

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
```

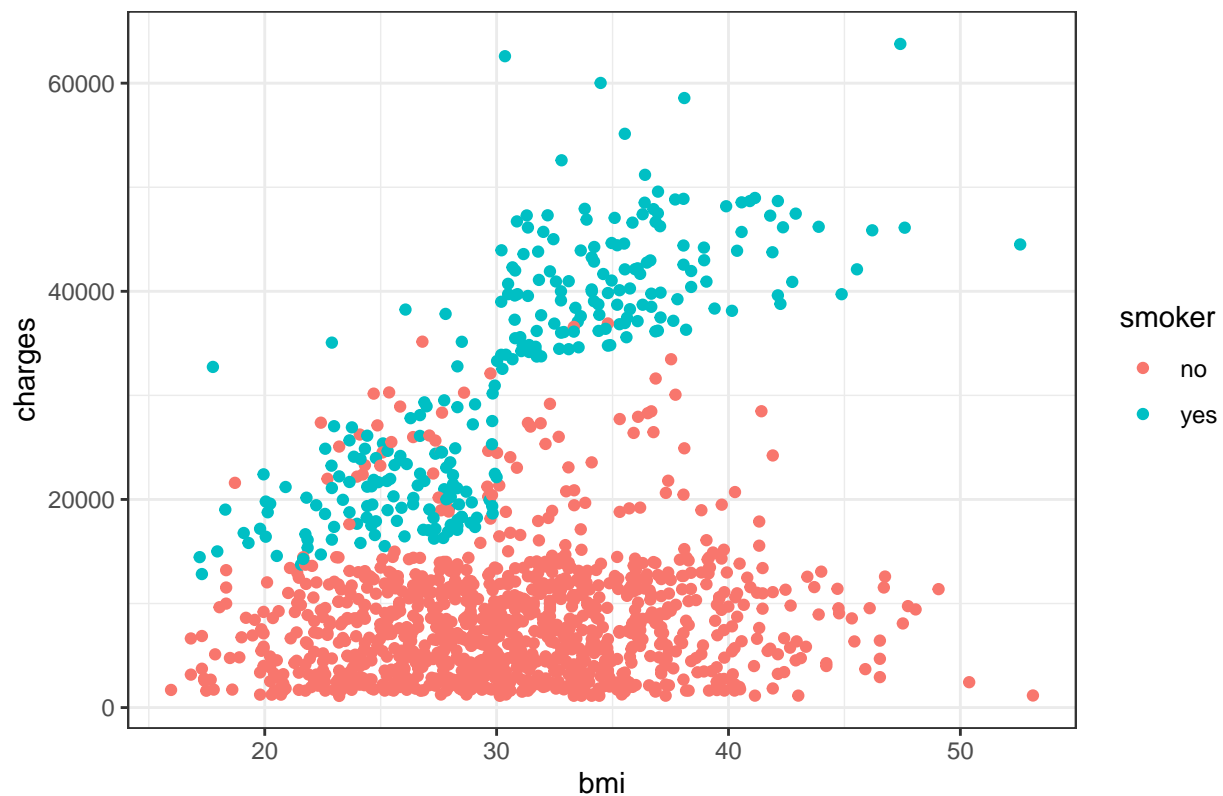
```
## 2  18  male 33.770      1    no southeast 1725.552
## 3  28  male 33.000      3    no southeast 4449.462
## 4  33  male 22.705      0    no northwest 21984.471
## 5  32  male 28.880      0    no northwest 3866.855
## 6  31 female 25.740      0    no southeast 3756.622
```

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

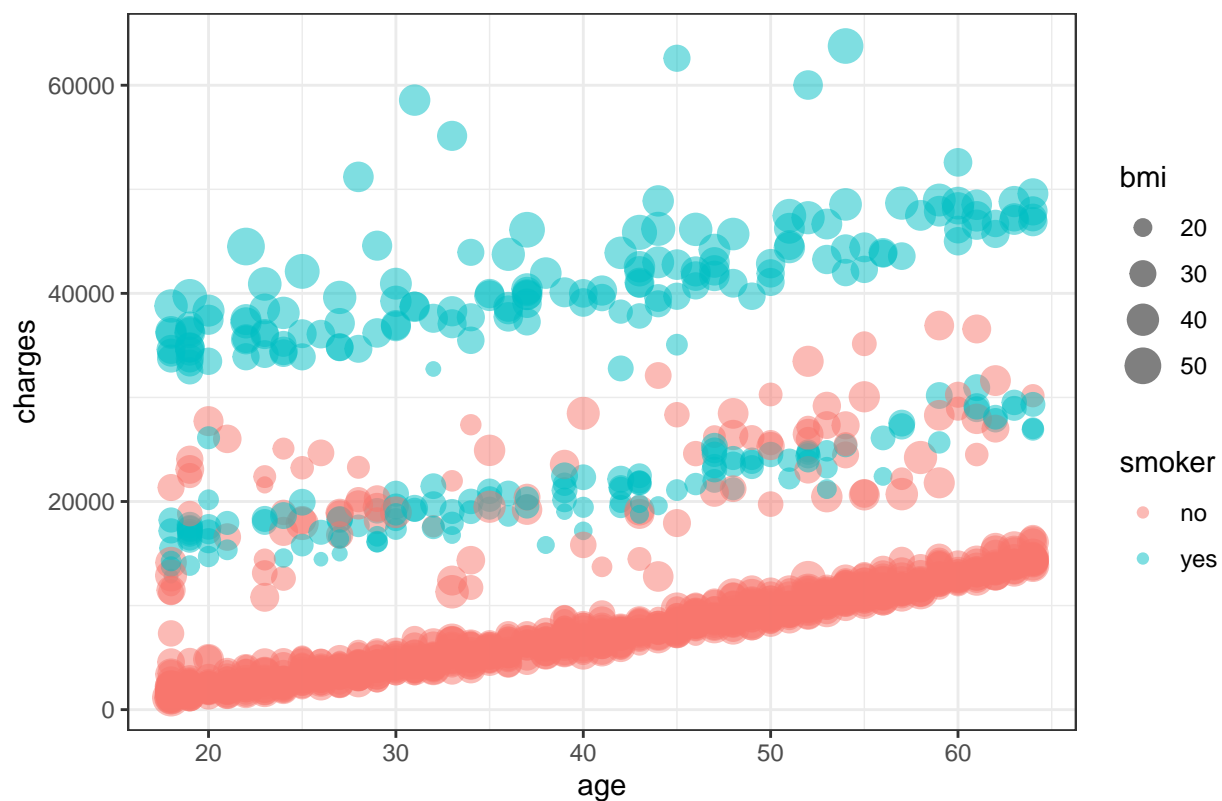
```
ggplot(aes(x=bmi, y=charges, color=smoker), data=data) +
  labs(title="Scatter Plot of Charges vs BMI by Smoker Status") +
  theme_bw() +
  geom_point()
```

Scatter Plot of Charges vs BMI by Smoker Status



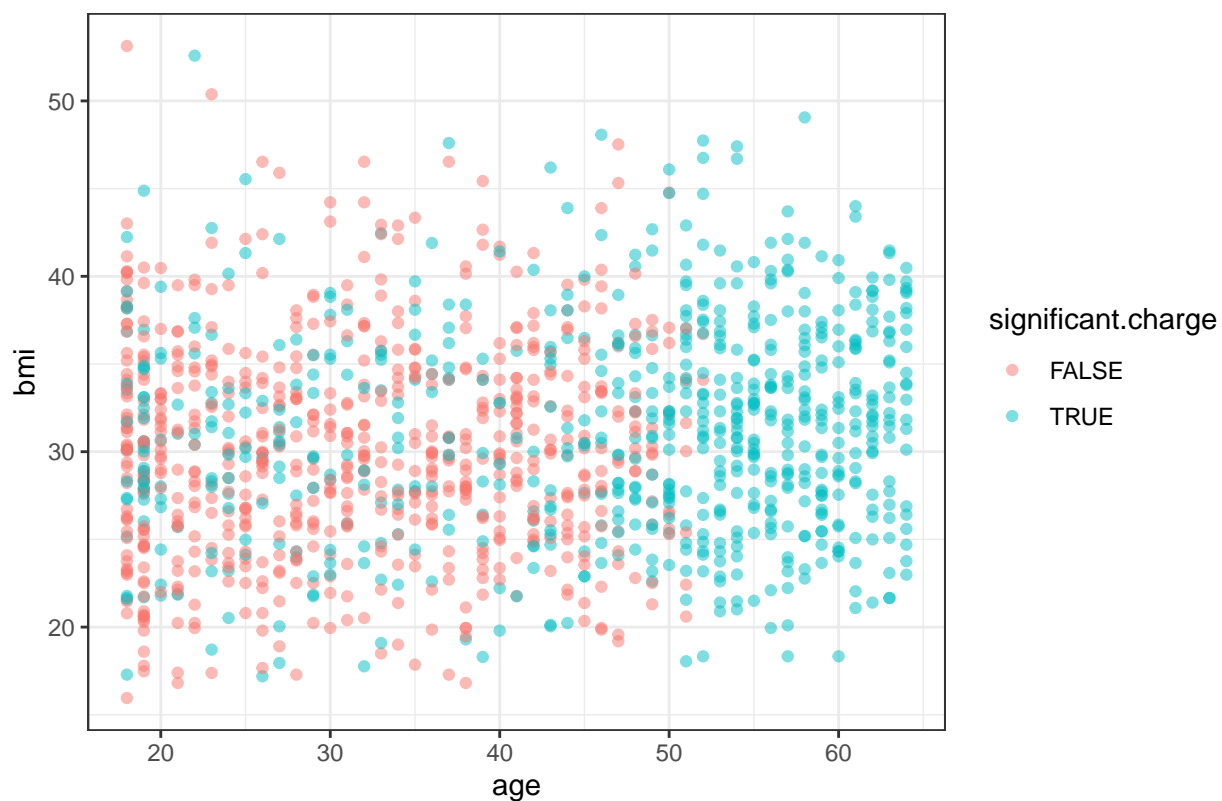
```
ggplot(aes(x=age,y=charges, color=smoker, size=bmi), data=data) +
  labs(title="Scatter plot of Charges vs Age by BMI and Smoker Status") +
  theme_bw() +
  geom_point(alpha=0.5)
```

Scatter plot of Charges vs Age by BMI and Smoker Status



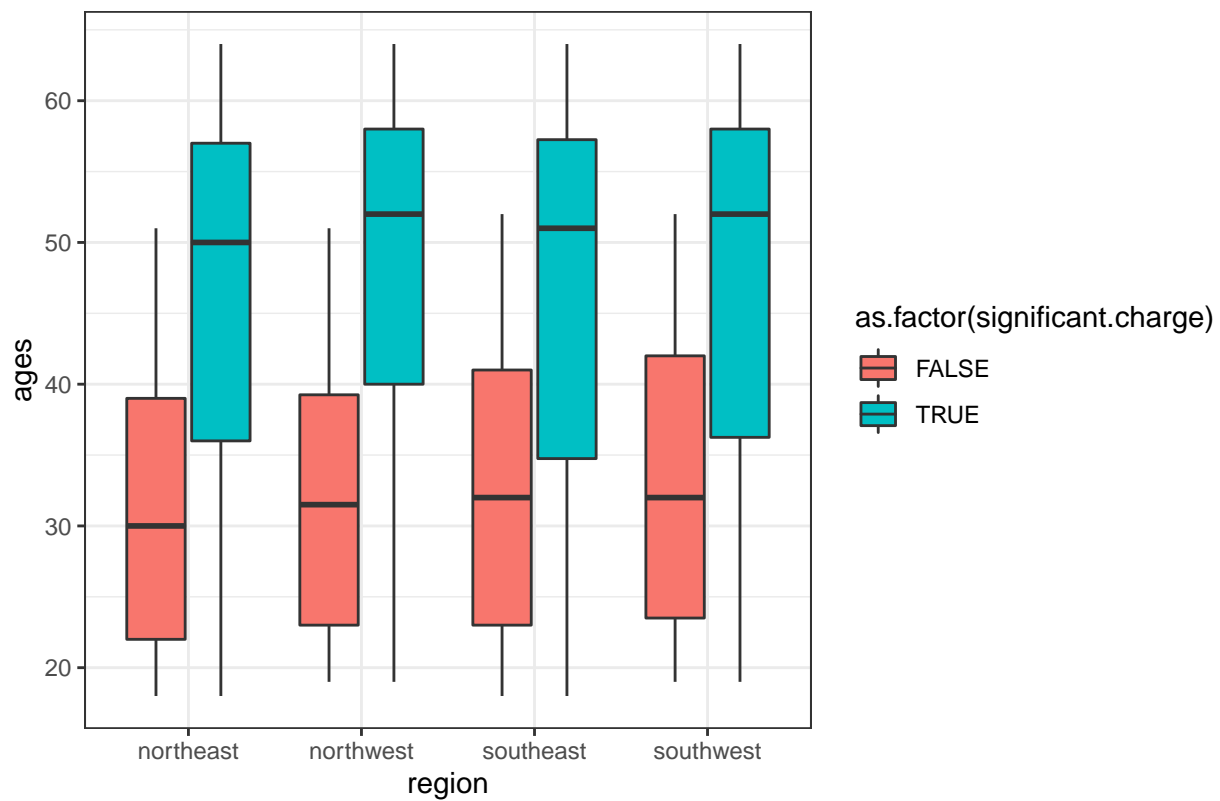
```
ggplot(aes(x=age,y=bmi, color=significant.charge), data=data) +  
  labs(title="Scatter plot of Charges vs Age by BMI and Smoker Status") +  
  theme_bw() +  
  geom_point(alpha=0.5)
```

Scatter plot of Charges vs Age by BMI and Smoker Status

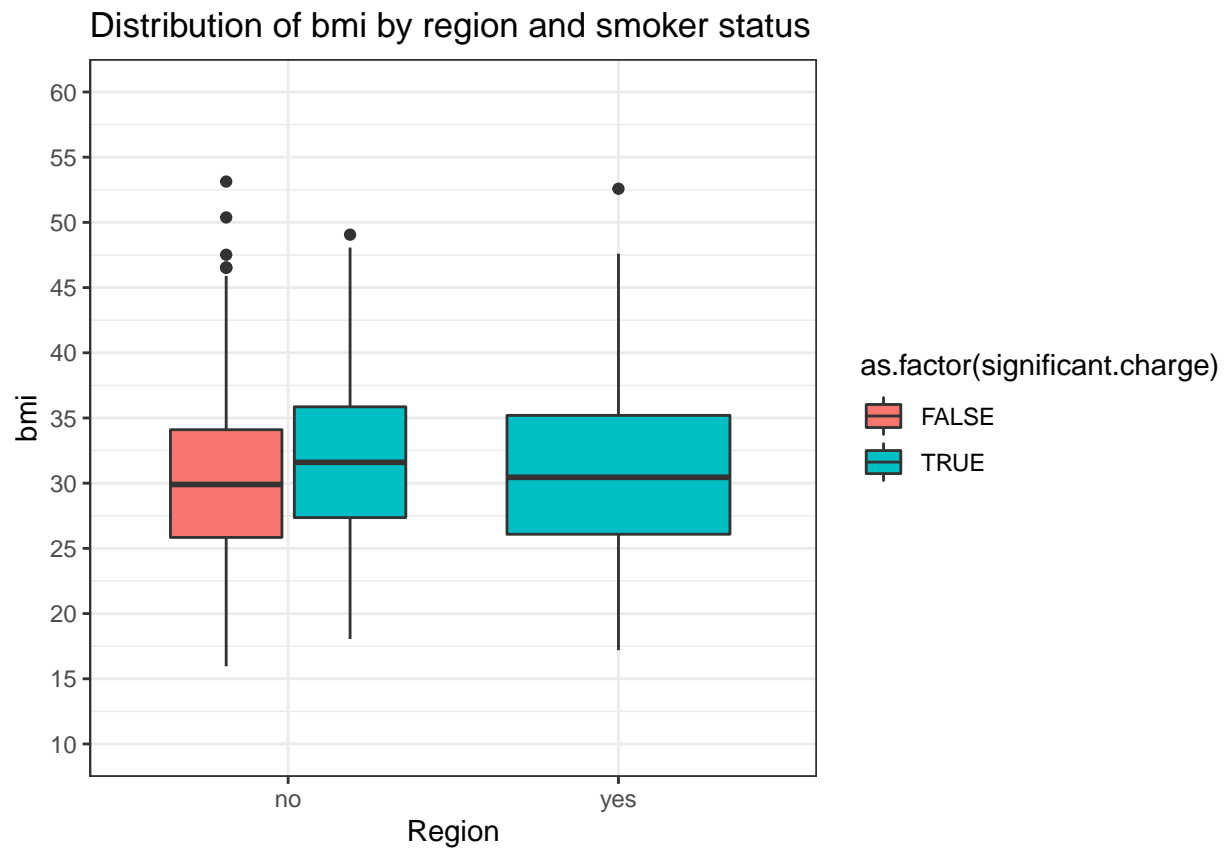


```
ggplot(data, aes(x=region, y=age, fill=as.factor(significant.charge)))+  
  geom_boxplot() +  
  theme_bw() +  
  labs(x="region", y="ages", title="Dist of bmi by region and smoker status")
```

Dist of bmi by region and smoker status

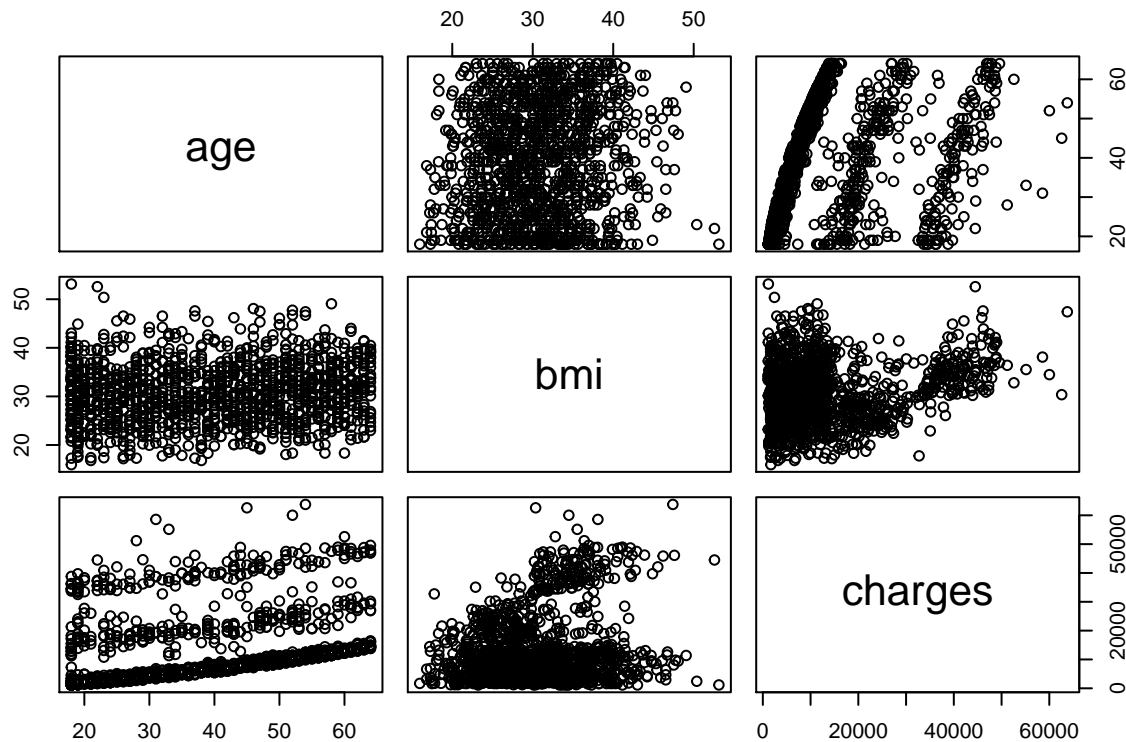


```
ggplot(data, aes(x=smoker, y=bmi, fill=as.factor(significant.charge)))+
  geom_boxplot() +
  theme_bw() +
  labs(x="Region", y="bmi", title="Distribution of bmi by region and smoker status") + scale_y_continuous
```



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
```

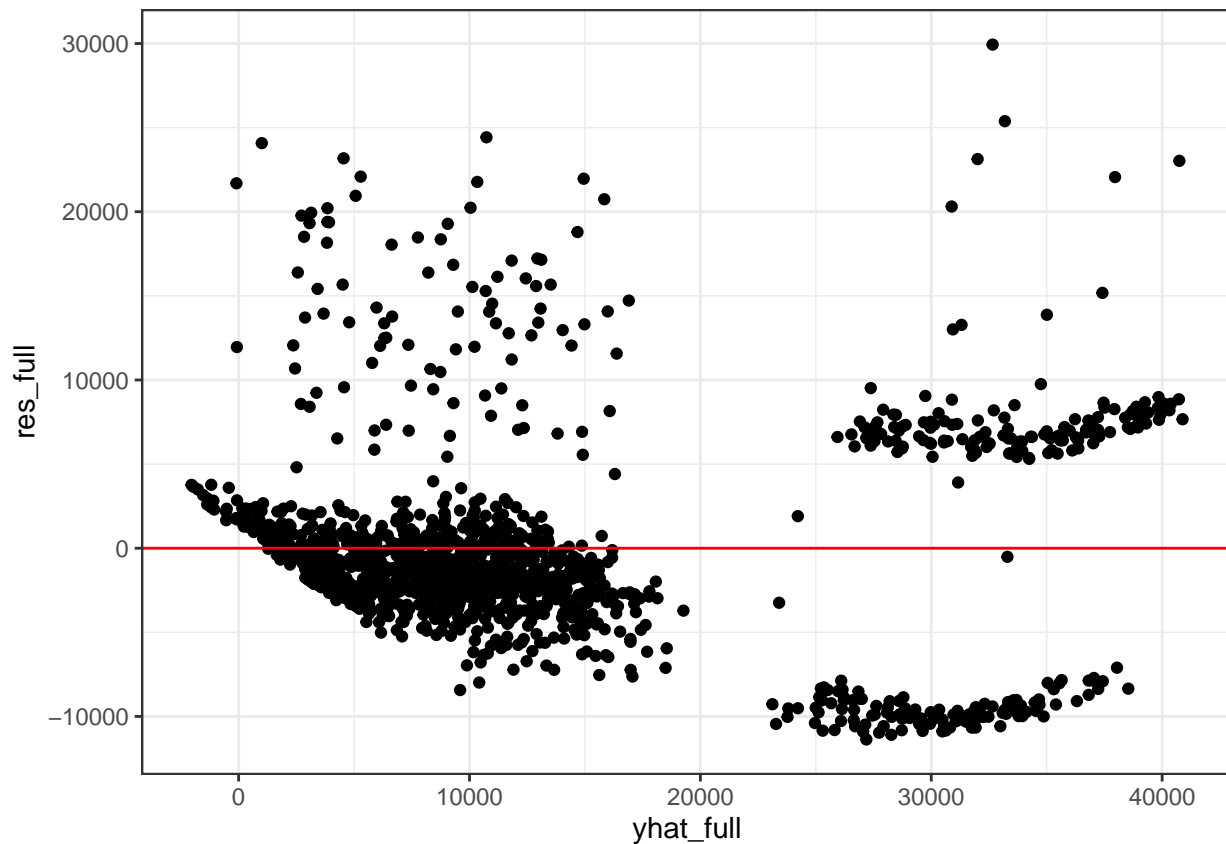
The full regression is as follows.

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

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 in the southwest. If the client is in the northeast I_3, I_4, I_5 will be zero, since this is the reference class.

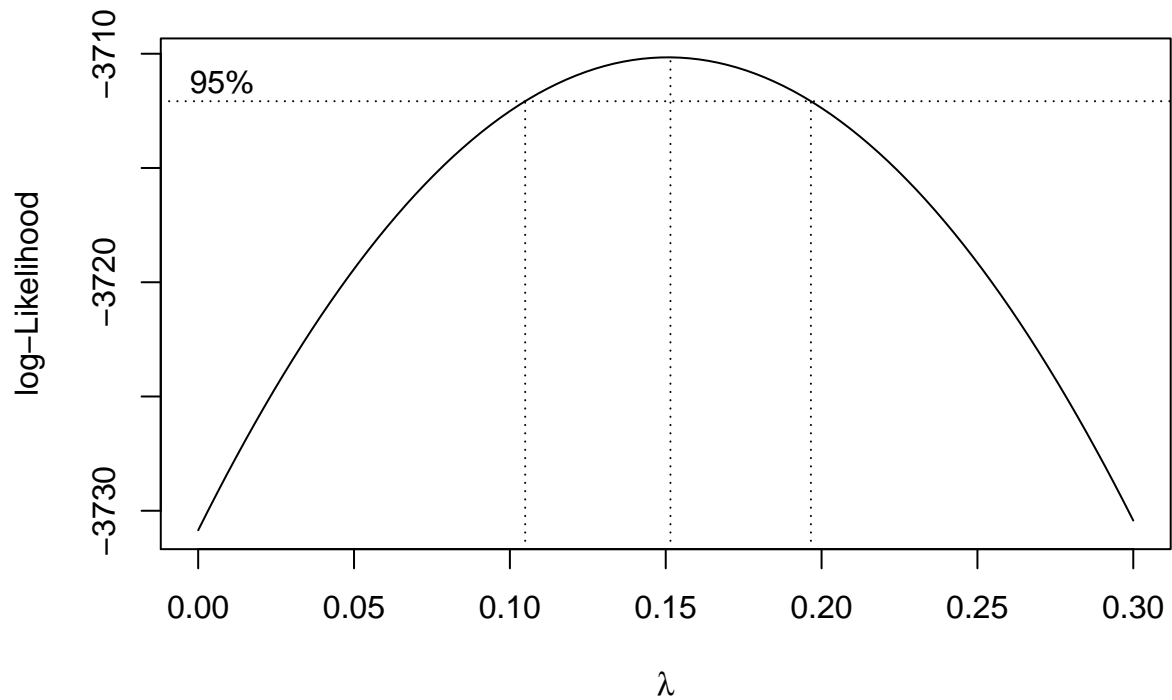
Assumption Check of Full Model

```
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))
```

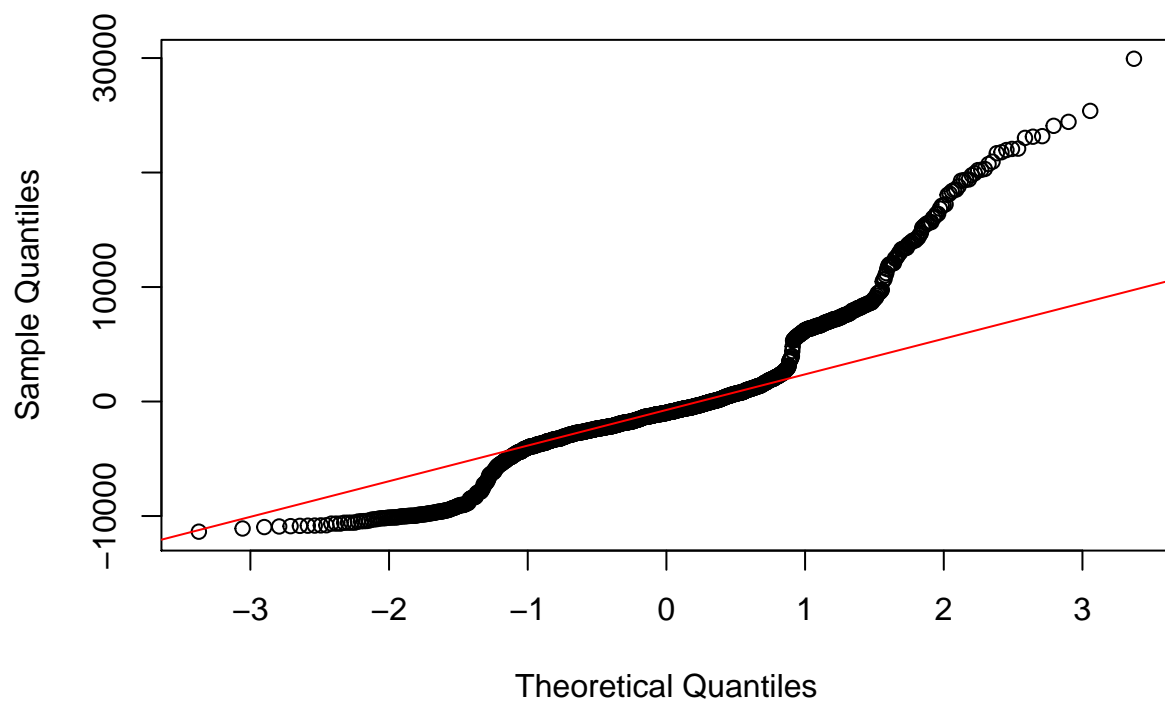


From the boxcox we can try a lambda value of 0.15 for transformation.

QQPlot

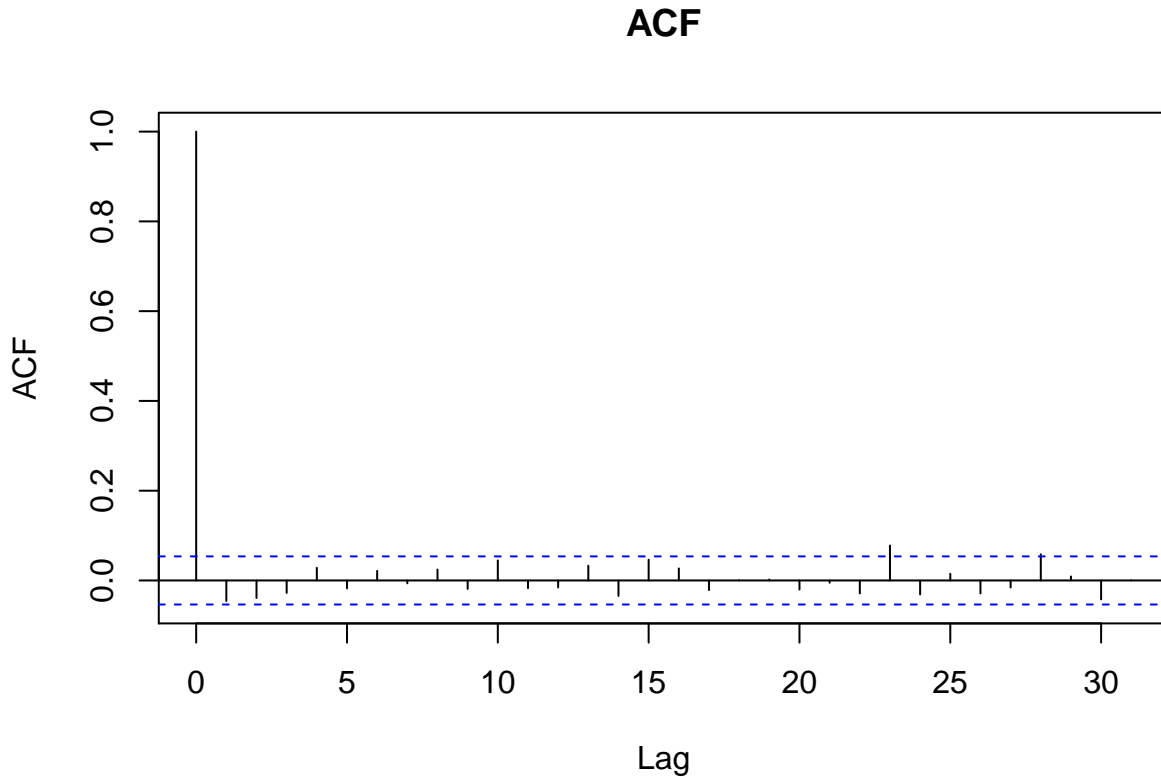
```
{
  qqnorm(mlr_full$residuals)
  qqline(mlr_full$residuals, col="red")
}
```

Normal Q-Q Plot



ACF

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



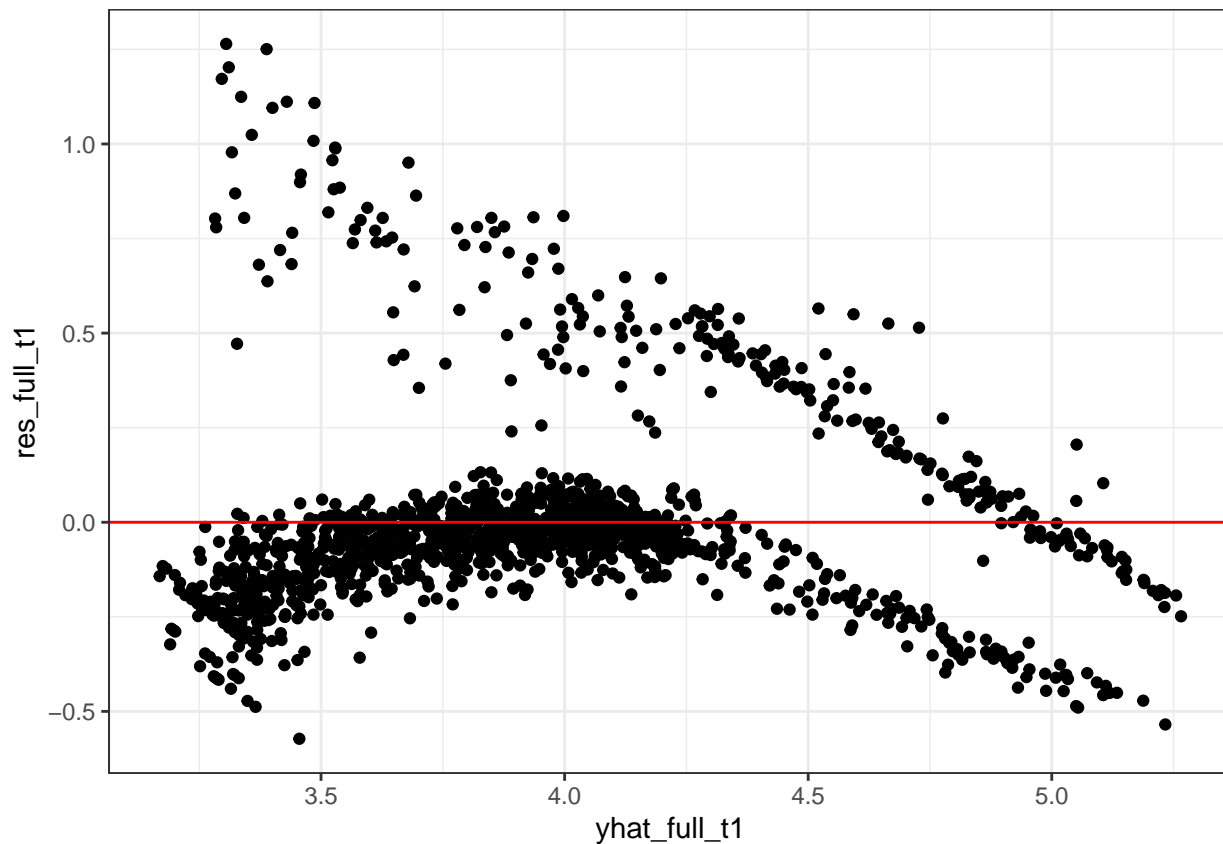
```
first_transformation_full <- data
first_transformation_full$charges <- first_transformation_full$charges^0.15
mlr_transform_first <- lm(charges ~ age + bmi + children + smoker + region, data=first_transformation_full)
summary(mlr_transform_first)
```

```
##
## 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
```

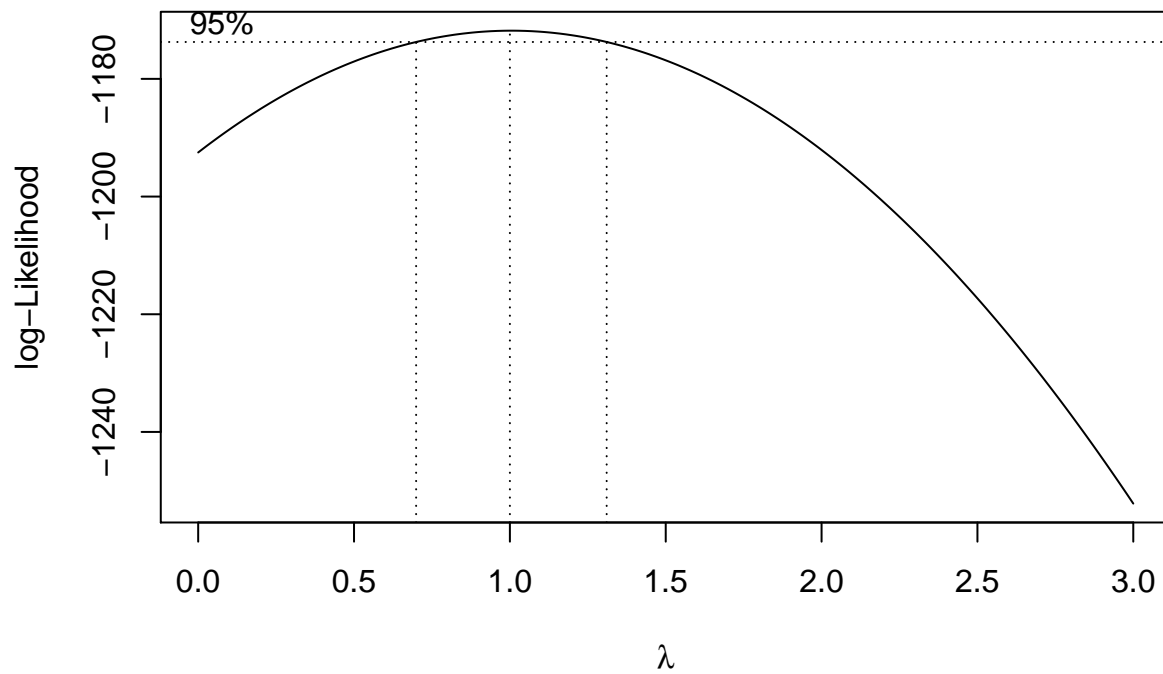
Residual Plot of the transformed model.

```
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

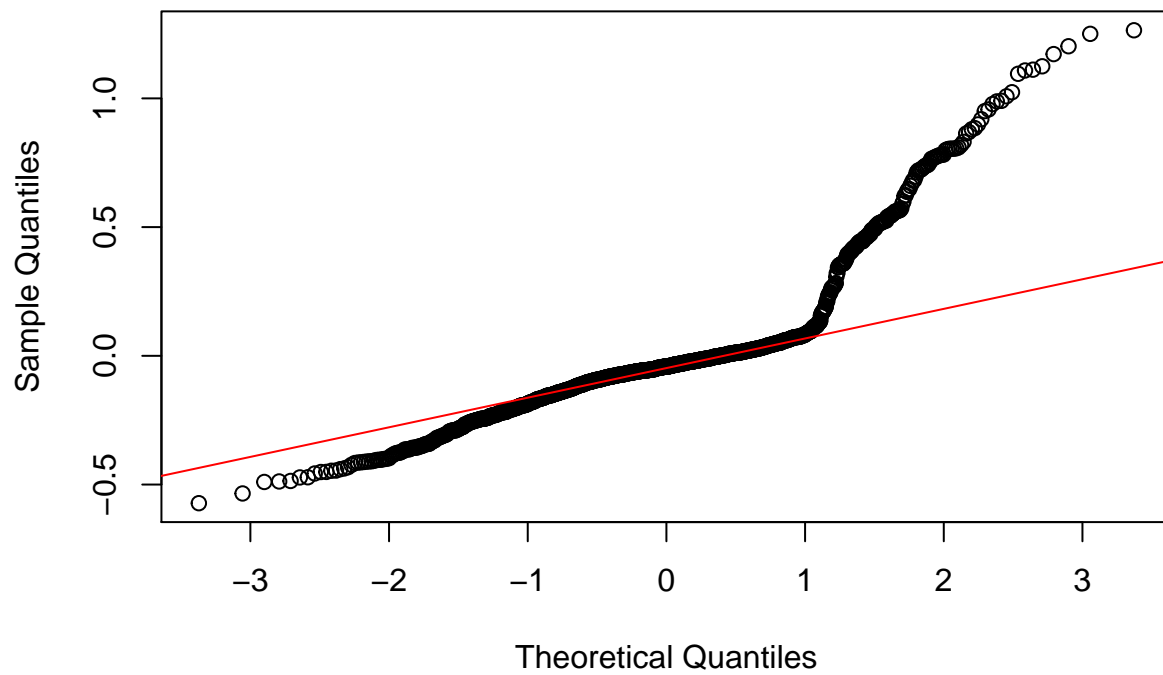
```
boxcox(mlr_transform_first, lambda=seq(0,3, 0.01))
```



QQPLOT

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

Normal Q-Q Plot



Possible Influential points?

```
n = dim(data)[1]
p = 8
COOKS<-cooks.distance(mlr_transform_first)
COOKS[COOKS>qf(0.5,p,n-p)]
```

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

Logistic

```
set.seed(6021) ##for reproducibility
sample<-sample.int(nrow(first_transformation_full), floor(.50*nrow(first_transformation_full)), replace=
train<- first_transformation_full[sample, ] ##training data frame
test<-first_transformation_full[-sample, ] ##test data frame
result<-glm(significant.charge ~ age + bmi + children + smoker + region, family="binomial", data=train)
summary(result)
```

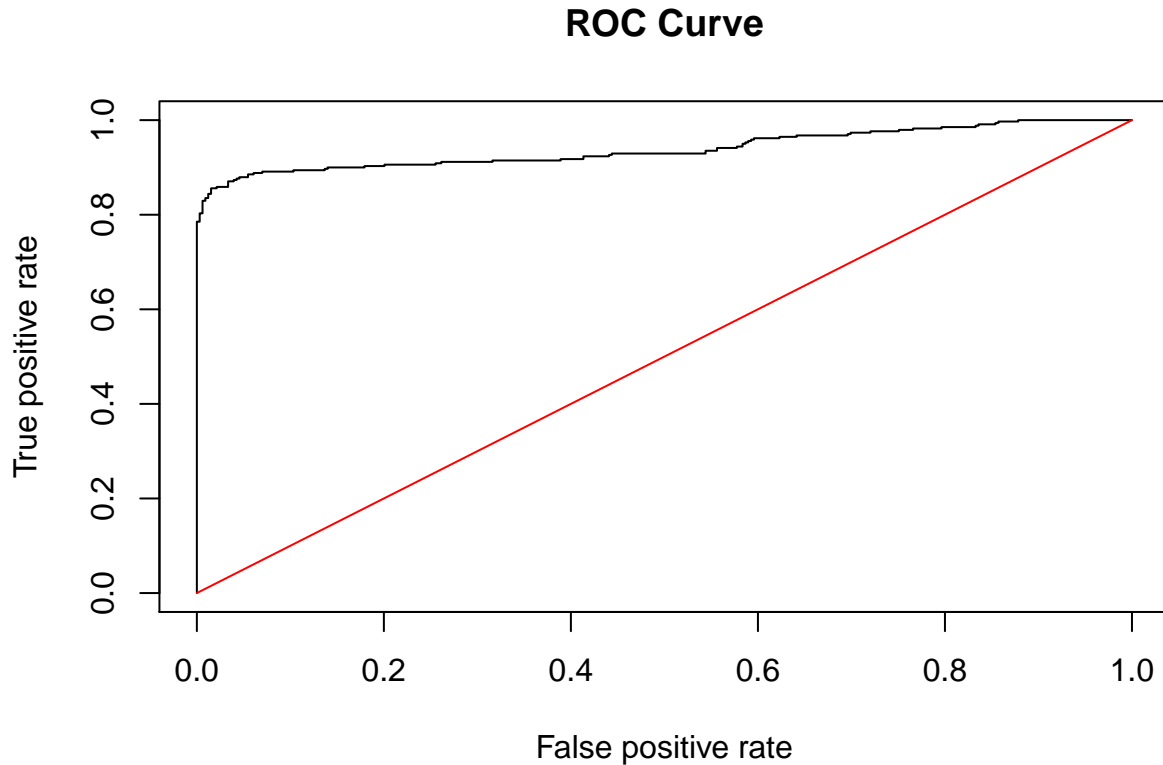
```
##
## Call:
## glm(formula = significant.charge ~ age + bmi + children + smoker +
##       region, family = "binomial", data = train)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.5477  -0.3328  -0.0757   0.3392   3.3986
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -9.80428    1.09743  -8.934  <2e-16 ***
## age             0.18932    0.01654  11.444  <2e-16 ***
## bmi             0.03258    0.02402   1.356    0.175
## children       0.19678    0.11015   1.786    0.074 .
## smokeryes     22.80340   693.10646   0.033    0.974
## regionnorthwest -0.37231    0.38719  -0.962    0.336
## regionsoutheast -0.47896    0.40576  -1.180    0.238
## regionsouthwest -0.17324    0.38930  -0.445    0.656
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 927.25  on 668  degrees of freedom
## Residual deviance: 355.27  on 661  degrees of freedom
## AIC: 371.27
##
## 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] 0.9392187
```

Matrix

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

```
##
##      FALSE TRUE
## FALSE   302   27
## TRUE    37   303
```

STOP

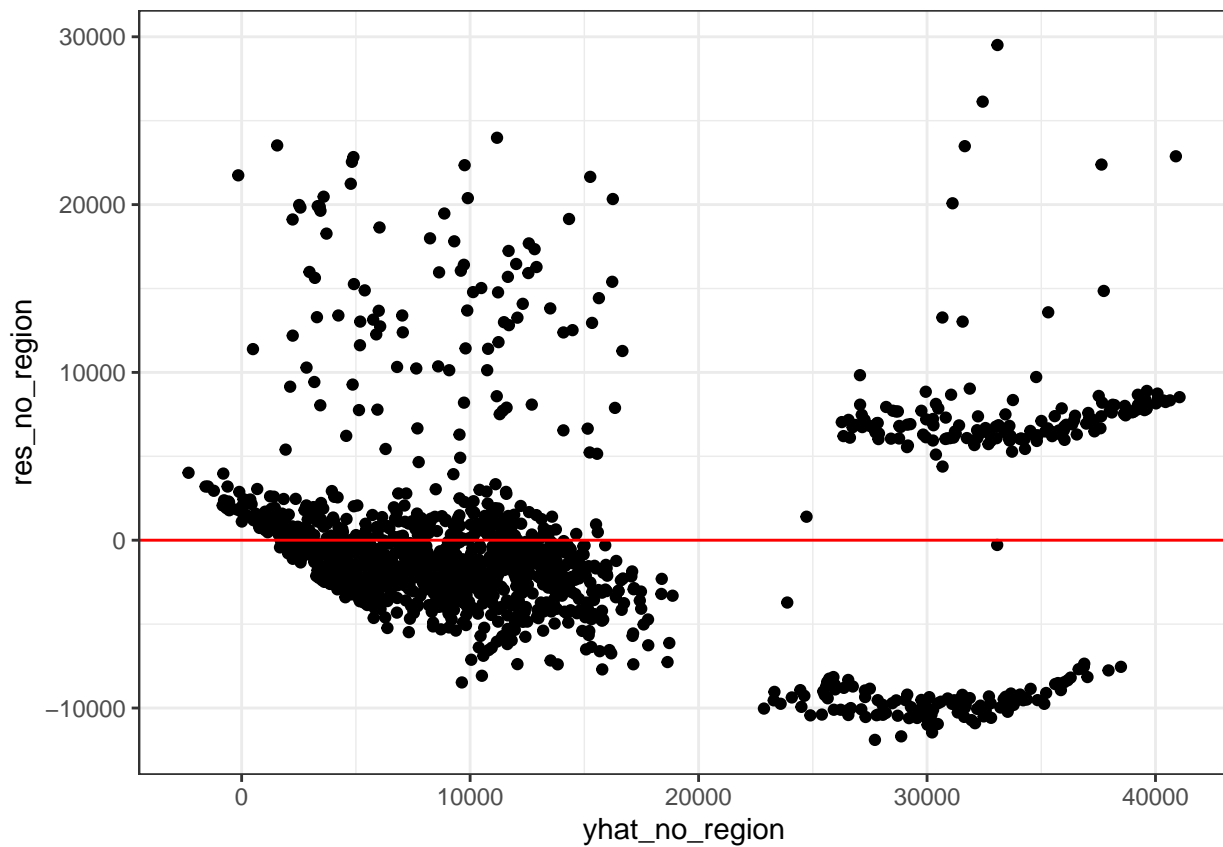
Without Region

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

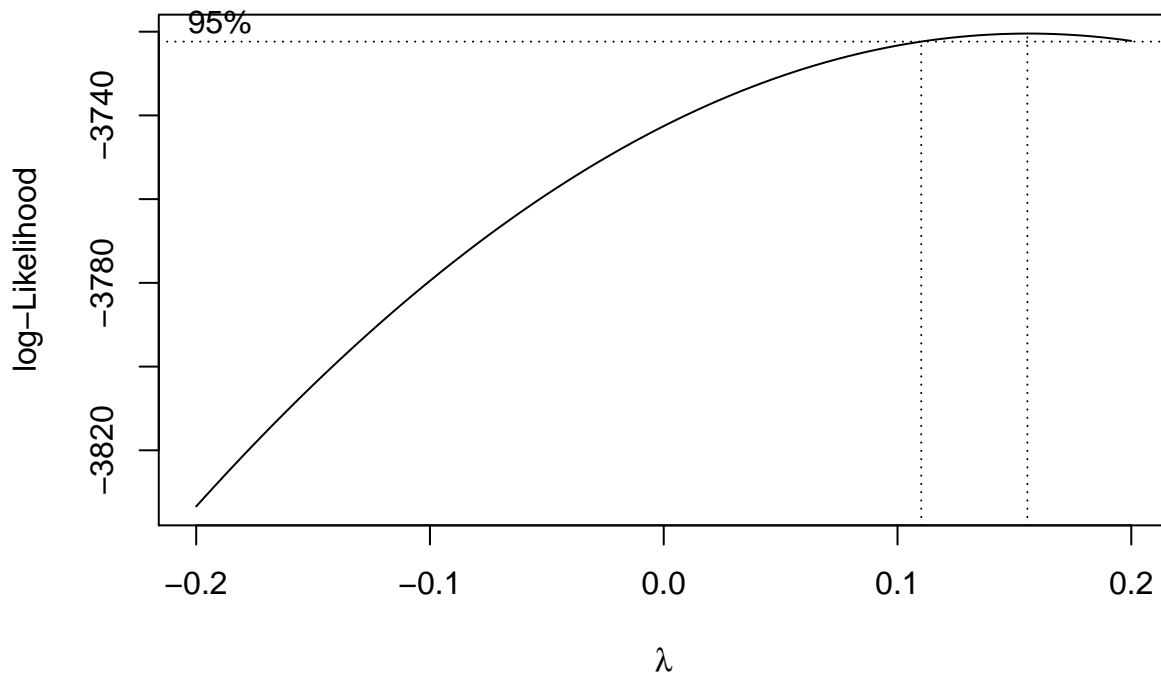
```
##
## Call:
## lm(formula = charges ~ age + bmi + children + smoker, data = no_class_predictor)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11897.9  -2920.8   -986.6   1392.2  29509.6
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -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_no_region <- mlr_no_region$fitted.values
res_no_region <- mlr_no_region$residuals
data %>%
  ggplot(aes(yhat_no_region, res_no_region)) +
  geom_point() +
  theme_bw() +
  geom_hline(yintercept = 0, color="red")
```



```
boxcox(mlr_no_region, lambda=seq(-0.2,0.2, 0.01))
```

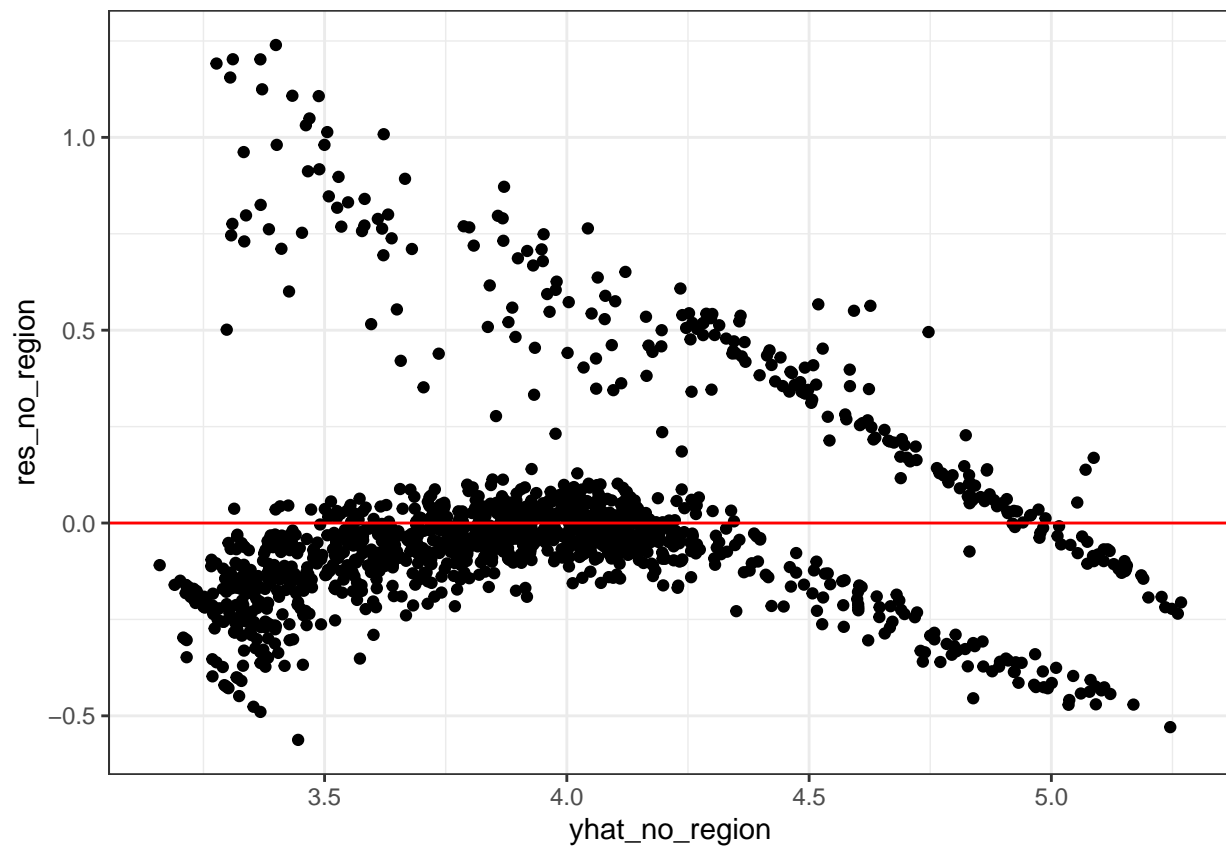


```
first_transformation_no_region <- no_class_predictor
first_transformation_no_region$charges <- first_transformation_no_region$charges^0.15
mlr_transform_first_no_region <- lm(charges ~ . - region, data=first_transformation_no_region)
summary(mlr_transform_first_no_region)
```

```
##
## Call:
## lm(formula = charges ~ . - region, data = first_transformation_no_region)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.56232 -0.12333 -0.04095  0.03491  1.23951
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.7280424  0.0405709  67.241  < 2e-16 ***
## age          0.0192130  0.0005077  37.844  < 2e-16 ***
## sexmale     -0.0368456  0.0142177  -2.592  0.00966 **
## bmi          0.0076871  0.0011694   6.574 7.03e-11 ***
## children     0.0526502  0.0058795   8.955  < 2e-16 ***
## smokeryes    0.9574101  0.0175940  54.417  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2589 on 1332 degrees of freedom
## Multiple R-squared:  0.773, Adjusted R-squared:  0.7722
## F-statistic: 907.4 on 5 and 1332 DF, p-value: < 2.2e-16
```

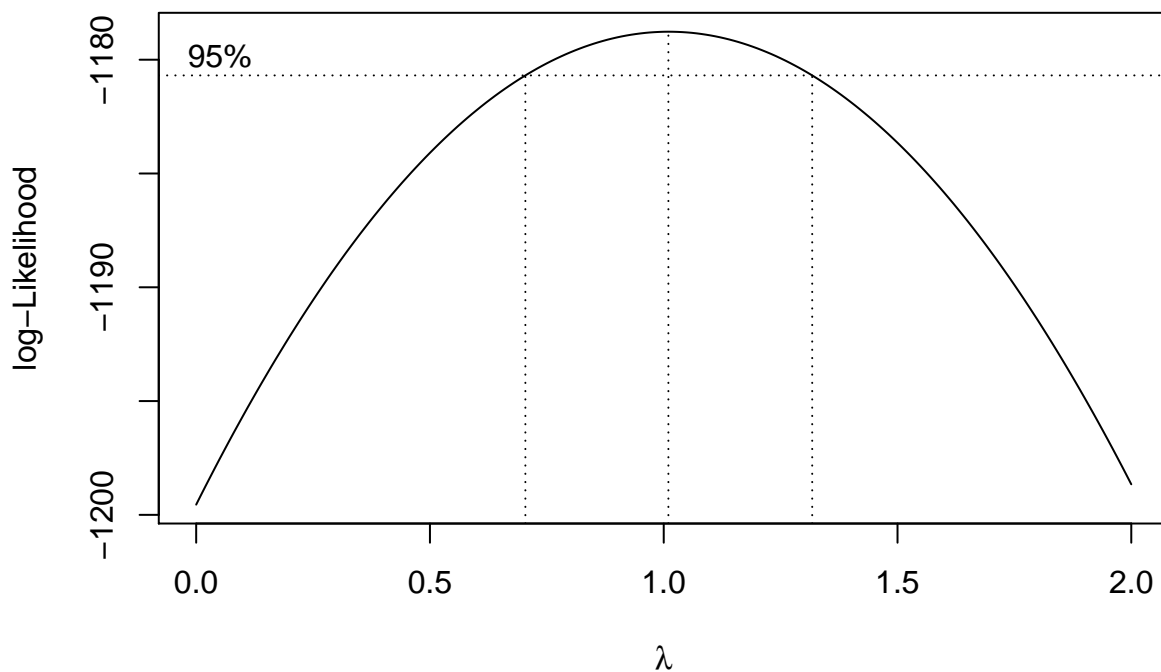
Residual Plot

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



Box Cox

```
boxcox(mlr_transform_first_no_region, lambda=seq(-0,2, 0.01))
```



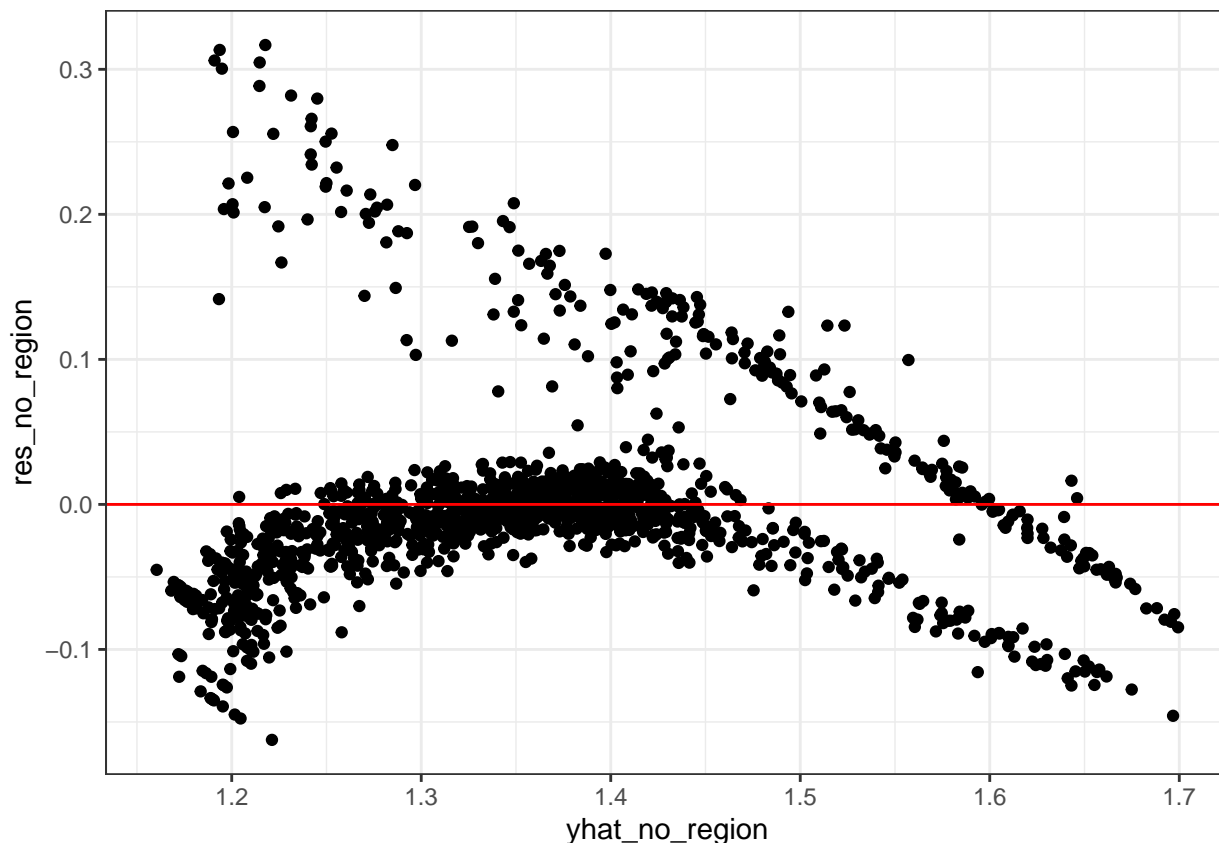
```
first_transformation_no_region_log <- first_transformation_no_region
first_transformation_no_region_log$charges <- log(first_transformation_no_region_log$charges)
```

```
mlr_transform_first_no_region_log <- lm(charges ~ . - region , data=first_transformation_no_region_log)
summary(mlr_transform_first_no_region_log)
```

```
##
## Call:
## lm(formula = charges ~ . - region, data = first_transformation_no_region_log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.16236 -0.03047 -0.00778  0.01058  0.31676
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.0518166  0.0105253  99.932  < 2e-16 ***
## age          0.0052074  0.0001317  39.536  < 2e-16 ***
## sexmale      -0.0112513  0.0036885  -3.050  0.00233 **
## bmi          0.0016363  0.0003034   5.394 8.16e-08 ***
## children     0.0152591  0.0015253  10.004  < 2e-16 ***
## smokeryes    0.2325355  0.0045644  50.946  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06716 on 1332 degrees of freedom
## Multiple R-squared:  0.7638, Adjusted R-squared:  0.7629
## F-statistic: 861.5 on 5 and 1332 DF,  p-value: < 2.2e-16
```

Residual Plot

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



Interaction Terms

```
dropped_sign <- data[1:7]
interaction1 <- lm( charges ~ (age + sex + bmi + children + region) * smoker, data = dropped_sign)
without_inter <- lm( charges ~ age + sex + bmi + children + region + smoker, data = dropped_sign)
anova(without_inter, interaction1)
```

Analysis of Variance Table

##

Model 1: charges ~ age + sex + bmi + children + region + smoker

Model 2: charges ~ (age + sex + bmi + children + region) * smoker

Res.Df RSS Df Sum of Sq F Pr(>F)

1 1329 4.8840e+10

2 1322 3.1055e+10 7 1.7784e+10 108.15 < 2.2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
interaction2 <- lm( charges ~ (bmi) * smoker + children + region + age, data = dropped_sign)
```

```
anova(without_inter, interaction2)
```

Analysis of Variance Table

##

Model 1: charges ~ age + sex + bmi + children + region + smoker

Model 2: charges ~ (bmi) * smoker + children + region + age

Res.Df RSS Df Sum of Sq F Pr(>F)

1 1329 4.8840e+10

2 1329 3.1275e+10 0 1.7565e+10

See which is interaction


```

transform <- data[1:7]
transform$charges <- dropped_sign$charges^0.15
reduced <- lm(charges ~ age + sex + bmi + children + region + smoker, data = transform)
bmi_smoker <- lm(charges ~ age + sex + bmi*smoker + children + region, data=transform)
anova(reduced, bmi_smoker)

## Analysis of Variance Table
##
## Model 1: charges ~ age + sex + bmi + children + region + smoker
## Model 2: charges ~ age + sex + bmi * smoker + children + region
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1     1329 87.884
## 2     1328 79.027   1    8.8566 148.83 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

bmi_smoker_age <- lm(charges ~ age*smoker + sex + bmi*smoker + children + region, data=transform)
anova(bmi_smoker, bmi_smoker_age)

## Analysis of Variance Table
##
## Model 1: charges ~ age + sex + bmi * smoker + children + region
## Model 2: charges ~ age * smoker + sex + bmi * smoker + children + region
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1     1328 79.027
## 2     1327 67.093   1    11.934 236.03 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

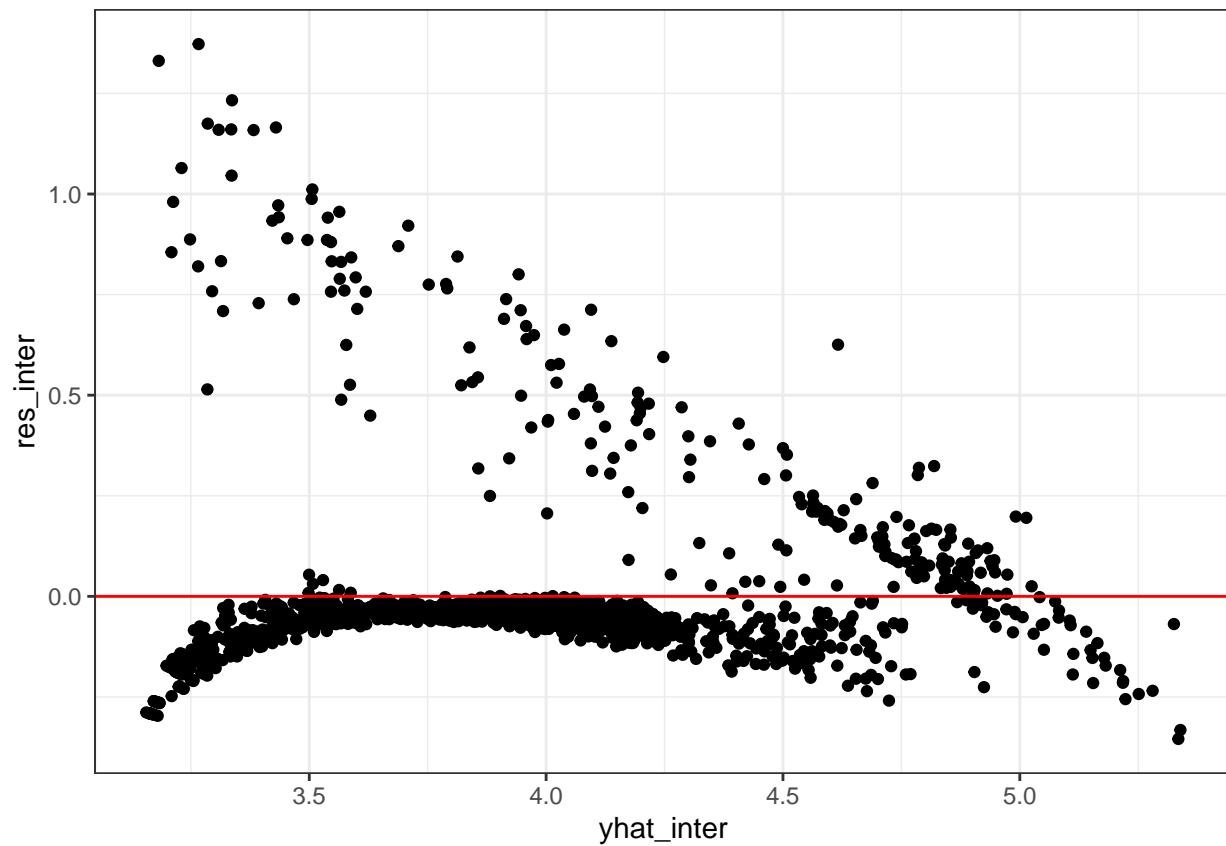
summary(bmi_smoker_age)

##
## Call:
## lm(formula = charges ~ age * smoker + sex + bmi * smoker + children +
##     region, data = transform)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35437 -0.08696 -0.04867 -0.02270  1.37264
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.8527139   0.0407222   70.053 < 2e-16 ***
## age           0.0227181   0.0004946   45.932 < 2e-16 ***
## smokeryes     0.5353670   0.0849260    6.304 3.94e-10 ***
## sexmale      -0.0445983   0.0123655   -3.607 0.000322 ***
## bmi           0.0008010   0.0011901    0.673 0.501034
## children      0.0552592   0.0051130   10.808 < 2e-16 ***
## regionnorthwest -0.0368187   0.0176716   -2.083 0.037397 *
## regionsoutheast -0.0796070   0.0177679   -4.480 8.09e-06 ***
## regionsouthwest -0.0839379   0.0177390   -4.732 2.46e-06 ***
## age:smokeryes  -0.0168791   0.0010987  -15.363 < 2e-16 ***
## smokeryes:bmi   0.0350887   0.0024492   14.326 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

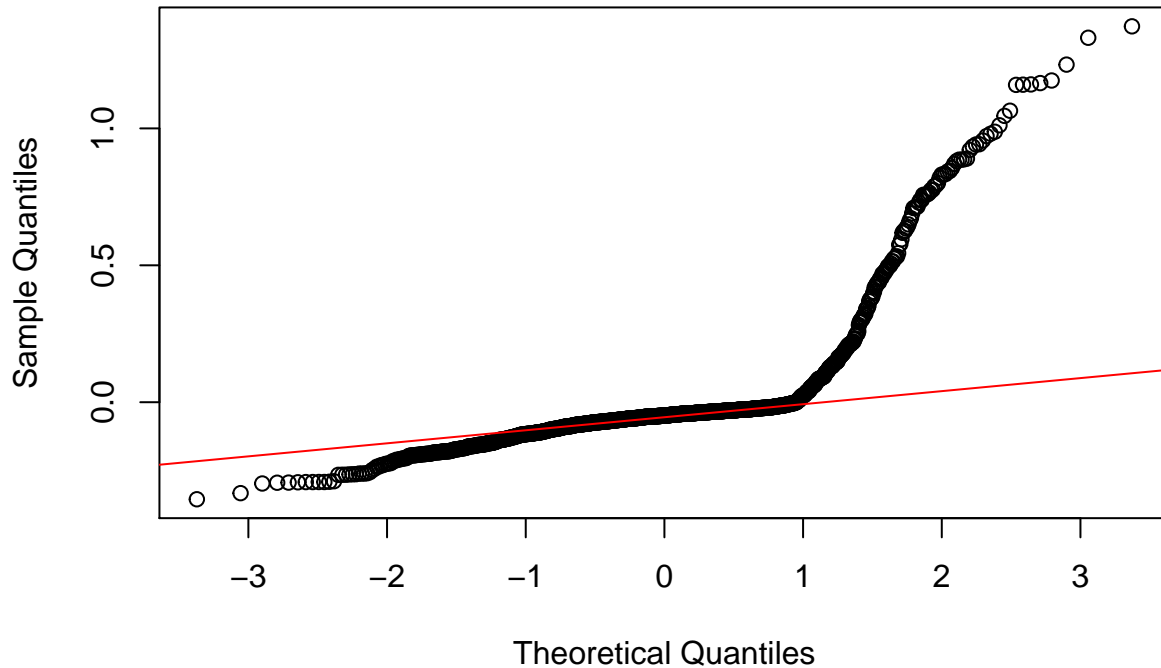
```
##
## Residual standard error: 0.2249 on 1327 degrees of freedom
## Multiple R-squared:  0.8294, Adjusted R-squared:  0.8281
## F-statistic: 645.2 on 10 and 1327 DF,  p-value: < 2.2e-16
```

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



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

Normal Q-Q Plot



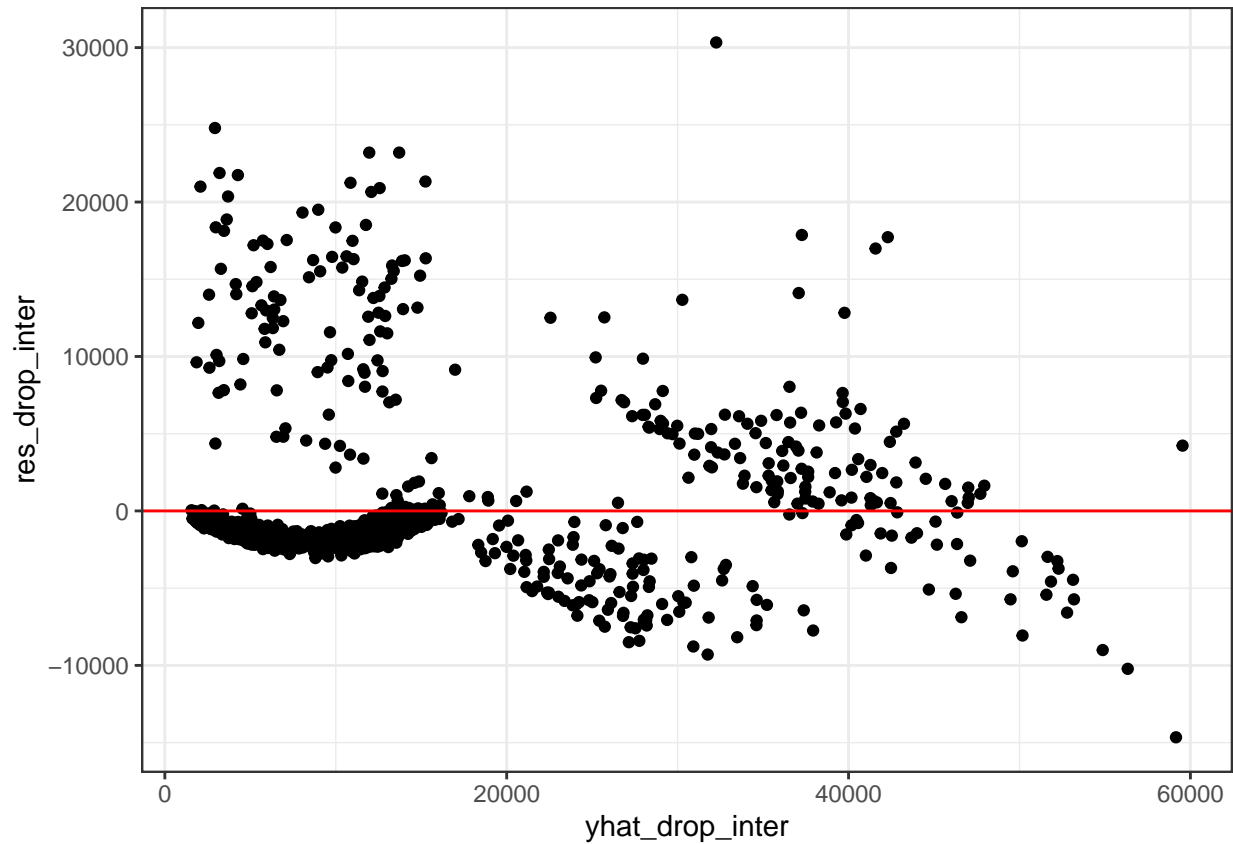
STOP

```
summary(interaction2)
```

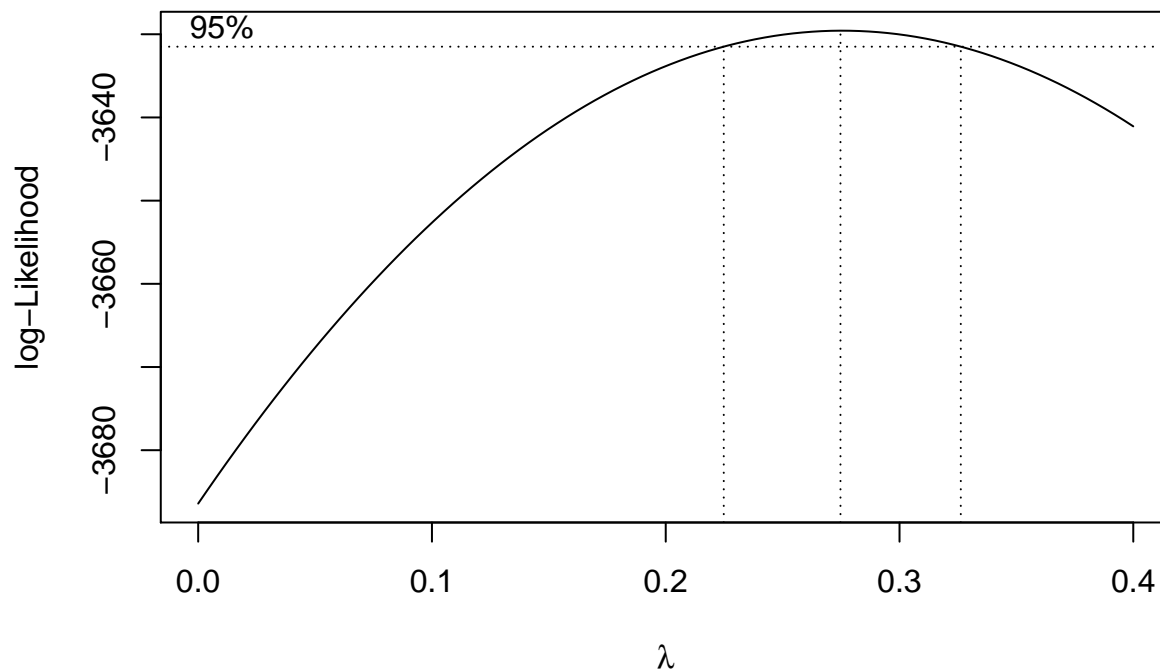
```
##
## Call:
## lm(formula = charges ~ (bmi) * smoker + children + region + age,
##     data = dropped_sign)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14655.4  -1918.9  -1313.4   -489.7   30333.1
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2453.564    857.695  -2.861  0.00429 **
## bmi           22.615     25.620   0.883  0.37756
## smokeryes    -20309.092   1648.861 -12.317 < 2e-16 ***
## children      512.713     110.266   4.650 3.65e-06 ***
## regionnorthwest -581.704     381.215  -1.526  0.12727
## regionsoutheast -1207.011     383.109  -3.151  0.00167 **
## regionsouthwest -1227.601     382.576  -3.209  0.00136 **
## age           264.042       9.522  27.729 < 2e-16 ***
## bmi:smokeryes  1438.108     52.630  27.325 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4851 on 1329 degrees of freedom
## Multiple R-squared:  0.8405, Adjusted R-squared:  0.8395
## F-statistic: 875.4 on 8 and 1329 DF, p-value: < 2.2e-16
```

Residual Plot

```
yhat_drop_inter <- interaction2$fitted.values  
res_drop_inter <- interaction2$residuals  
data %>%  
  ggplot(aes(yhat_drop_inter, res_drop_inter)) +  
  geom_point() +  
  theme_bw() +  
  geom_hline(yintercept = 0, color="red")
```



```
boxcox(interaction2, lambda=seq(-0,0.4, 0.01))
```



Lets try adding interaction term

```
interaction2
```

```
##
## Call:
## lm(formula = charges ~ (bmi) * smoker + children + region + age,
##     data = dropped_sign)
##
## Coefficients:
##      (Intercept)          bmi      smokeryes      children
##      -2453.56         22.61      -20309.09         512.71
## regionnorthwest regionsoutheast regionsouthwest         age
##      -581.70         -1207.01         -1227.60         264.04
##      bmi:smokeryes
##      1438.11

interaction3 <- lm( charges ~ (bmi + age) * smoker + children + region, data = dropped_sign)
summary(interaction3)
```

```
##
## Call:
## lm(formula = charges ~ (bmi + age) * smoker + children + region,
##     data = dropped_sign)
##
## 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 **
## bmi           22.444     25.679   0.874  0.38228
## age          264.558     10.672  24.791 < 2e-16 ***
```

```
## smokeryes      -20223.654   1831.889 -11.040 < 2e-16 ***
## 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 **
## bmi:smokeryes   1438.525      52.793  27.249 < 2e-16 ***
## age:smokeryes   -2.542      23.711  -0.107 0.91464
## ---
## 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
```

```
anova(interaction2, interaction3)
```

```
## Analysis of Variance Table
##
## Model 1: charges ~ (bmi) * smoker + children + region + age
## Model 2: charges ~ (bmi + age) * smoker + children + region
##   Res.Df      RSS Df Sum of Sq    F Pr(>F)
## 1    1329 3.1275e+10
## 2    1328 3.1274e+10  1    270702 0.0115 0.9146
```

Woops lets just go with the model

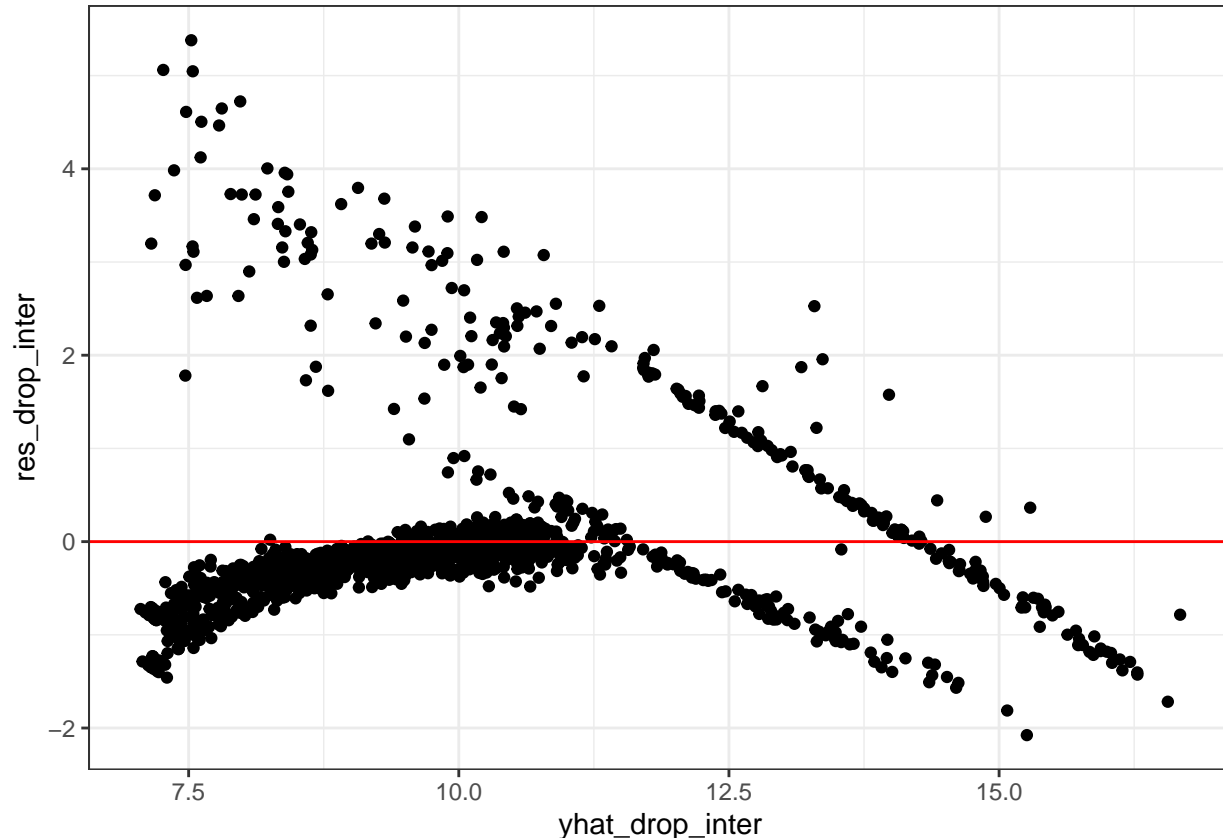
```
first_transform_interaction <- dropped_sign
first_transform_interaction$charges <- first_transform_interaction$charges^(0.25)
mlr_interaction_transform <- lm( charges ~ (bmi) * smoker + children + region + age, data = first_tran
summary(mlr_interaction_transform)
```

```
##
## Call:
## lm(formula = charges ~ (bmi) * smoker + children + region + age,
##     data = first_transform_interaction)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.0760 -0.4432 -0.1898  0.0246  5.3793
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.855000   0.177751  32.939 < 2e-16 ***
## bmi           0.007545   0.005310   1.421  0.1556
## smokeryes     -0.484587   0.341714  -1.418  0.1564
## children      0.204668   0.022852   8.956 < 2e-16 ***
## regionnorthwest -0.160264   0.079004  -2.029  0.0427 *
## regionsoutheast -0.351084   0.079397  -4.422 1.06e-05 ***
## regionsouthwest -0.311574   0.079286  -3.930 8.94e-05 ***
## age           0.077473   0.001973  39.258 < 2e-16 ***
## bmi:smokeryes   0.150073   0.010907  13.759 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.005 on 1329 degrees of freedom
## Multiple R-squared:  0.8065, Adjusted R-squared:  0.8053
```

```
## F-statistic: 692.3 on 8 and 1329 DF,  p-value: < 2.2e-16
```

Residual

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



BIC since nothing works lmao

```
model3_full <- lm( charges ~ bmi * smoker * region + children + age, data = dropped_sign)
model3_reduced <- lm( charges ~ bmi * smoker + children + age, data = dropped_sign)
summary(model3_full)
```

```
##
## Call:
## lm(formula = charges ~ bmi * smoker * region + children + age,
##     data = dropped_sign)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13441.4  -1905.8  -1240.7   -390.9   30472.5
##
## Coefficients:
```

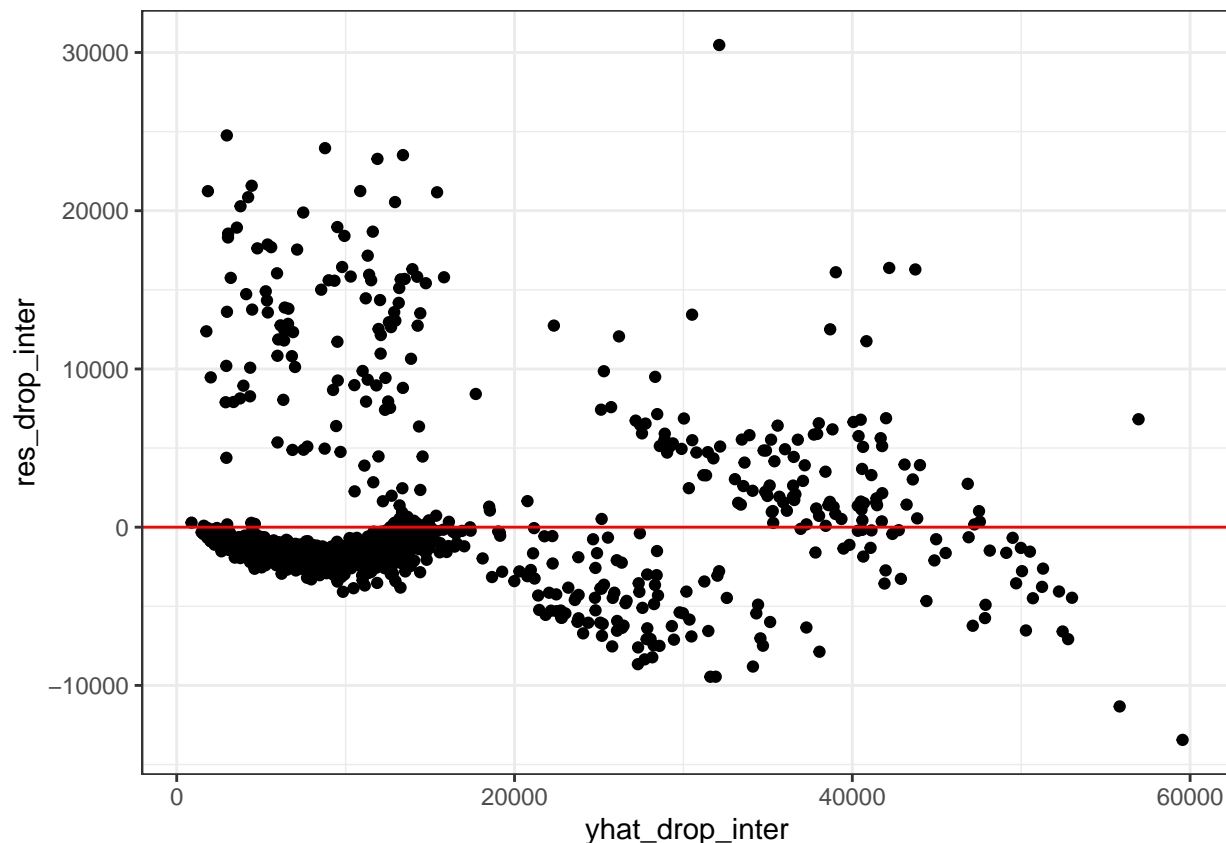
```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -4755.77   1569.63  -3.030  0.00249 **
## bmi             101.16     51.59   1.961  0.05011 .
## smokeryes      -19478.94  3195.05  -6.097 1.42e-09 ***
## regionnorthwest  519.13    2280.80   0.228  0.81999
## regionsoutheast 4029.91    2208.45   1.825  0.06826 .
## regionsouthwest 1807.67    2221.48   0.814  0.41595
## children        507.80     110.11   4.612 4.38e-06 ***
## age             263.61      9.53  27.662 < 2e-16 ***
## bmi:smokeryes   1414.58    108.75  13.007 < 2e-16 ***
## bmi:regionnorthwest -36.15    76.60  -0.472  0.63701
## bmi:regionsoutheast -160.31    69.41  -2.310  0.02105 *
## bmi:regionsouthwest -107.65    72.89  -1.477  0.13995
## smokeryes:regionnorthwest -7950.40  5385.74  -1.476  0.14013
## smokeryes:regionsoutheast  476.34  4329.57   0.110  0.91241
## smokeryes:regionsouthwest -5220.01  5158.76  -1.012  0.31178
## bmi:smokeryes:regionnorthwest 265.17   182.67   1.452  0.14684
## bmi:smokeryes:regionsoutheast -38.36   138.78  -0.276  0.78230
## bmi:smokeryes:regionsouthwest 197.64   168.73   1.171  0.24168
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4831 on 1320 degrees of freedom
## Multiple R-squared:  0.8429, Adjusted R-squared:  0.8409
## F-statistic: 416.5 on 17 and 1320 DF, p-value: < 2.2e-16
```

```
anova(model3_reduced, model3_full)
```

```
## Analysis of Variance Table
##
## Model 1: charges ~ bmi * smoker + children + age
## Model 2: charges ~ bmi * smoker * region + children + age
##   Res.Df      RSS Df Sum of Sq    F    Pr(>F)
## 1    1332 3.1600e+10
## 2    1320 3.0808e+10 12 791865991 2.8274 0.0007724 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual plots

```
yhat_drop_inter <- model3_full$fitted.values
res_drop_inter <- model3_full$residuals
data %>%
  ggplot(aes(yhat_drop_inter, res_drop_inter)) +
  geom_point() +
  theme_bw() +
  geom_hline(yintercept = 0, color="red")
```

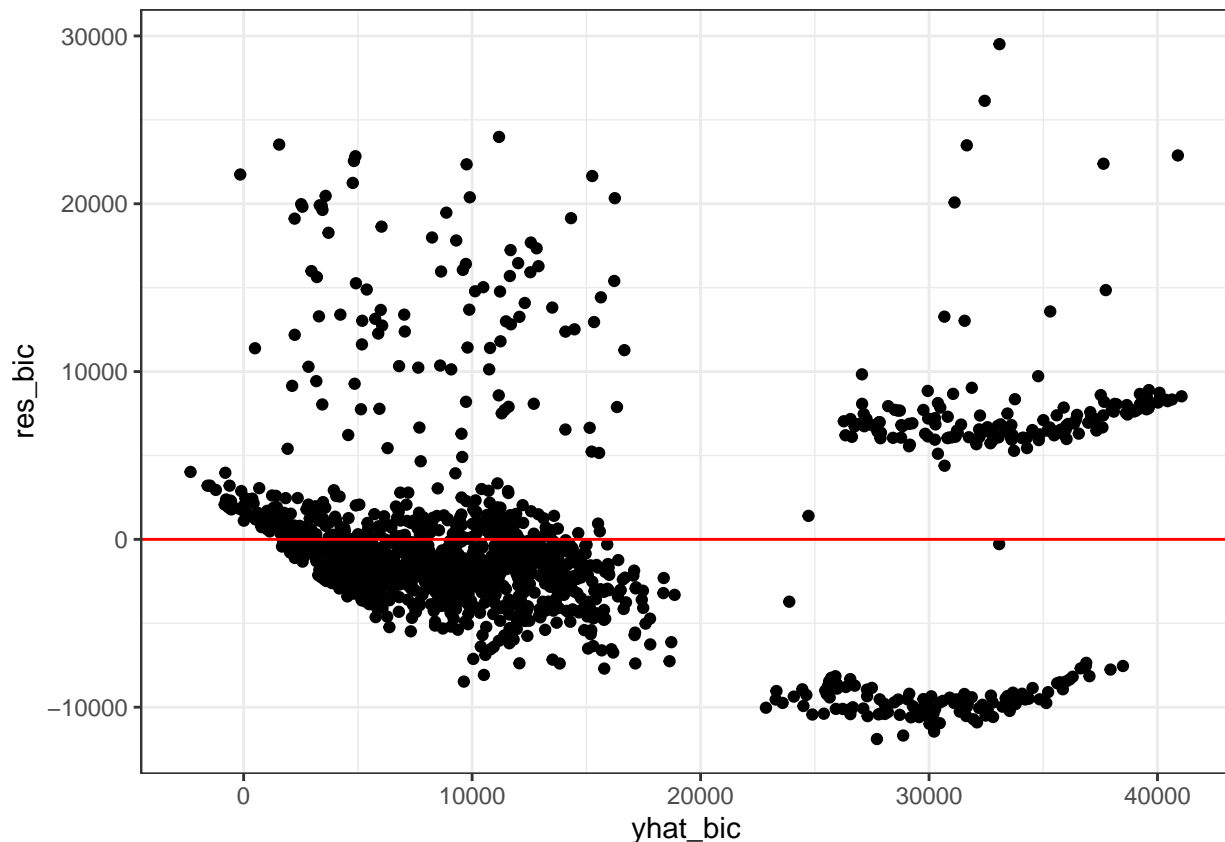



```
bic_model <- lm(charges ~ age + bmi + children + smoker, data = data)
summary(bic_model)
```

```
##
## 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

yhat_bic <- bic_model$fitted.values
res_bic <- bic_model$residuals
data %>%
```

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



```
bic_model_interaction <- lm(charges ~ (age+ bmi + children)*smoker, data = data)
anova(bic_model, bic_model_interaction)
```

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

```
##
```

```
## Model 1: charges ~ age + bmi + children + smoker
```

```
## Model 2: charges ~ (age + bmi + children) * smoker
```

```
##   Res.Df      RSS Df Sum of Sq    F    Pr(>F)
```

```
## 1   1333 4.9078e+10
```

```
## 2   1330 3.1556e+10  3 1.7522e+10 246.16 < 2.2e-16 ***
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
bic_model_interaction2 <- lm(charges ~ (bmi)*smoker + children, data = data)
```

```
anova(bic_model, bic_model_interaction2)
```

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

```
##
```

```
## Model 1: charges ~ age + bmi + children + smoker
```

```
## Model 2: charges ~ (bmi) * smoker + children
```

```
##   Res.Df      RSS Df Sum of Sq  F Pr(>F)
```

```
## 1   1333 4.9078e+10
```

```
## 2   1333 4.9848e+10  0 -769707304
```