Model Preparation

Manoj Kumar Nagabandi

16 07 2022

1. Fetch the data

This is the third part of the project. After cleaning data, exploring and engineering features, we prepare linear and generalized linear models.

```
file_input = 'rideshare_kaggle_modified.csv'
data = read.csv(file_input)
print(dim(data))
```

```
## [1] 293877    102
```

We randomly subset 3000 training points out of the initial data set. This is caused by R requirements for storing model parameters (algorithms refused to run on 15,000 instances).

```
subset_size = 293877  # Number of points that we use from the dataset

subsample_indices = sample.int(
  n = nrow(data),
  size = subset_size
)

data = data[subsample_indices, ]
```

```
train_percent = 0.6
val_percent = 0.2
train_val_percent = train_percent + val_percent
test_percent = 1 - train_val_percent
```

We randomly split our subset into train, validation and test sets.

```
train_val_sample = sample.int(
  n = nrow(data),
  size = floor(train_val_percent * nrow(data))
)

train_val = data[train_val_sample, ]
test = data[-train_val_sample, ]

train_sample = sample.int(
  n = nrow(train_val),
  size = floor(train_percent * nrow(data))
)

train = train_val[train_sample, ]
validation = train_val[-train_sample, ]
```

```r
ncat = function(...){
  cat(..., '\n')
}

ncat('Train size', nrow(train))
```

```
## Train size 176326
```

```r
ncat('Validation size', nrow(validation))
```

```
## Validation size 58775
```

```r
ncat('Test size', nrow(test))
```

```
## Test size 58776
```

Storing names of dependent and independent features.

```r
Y_colname = colnames(data)[length(colnames(data))]

X_colnames = colnames(data[1 : length(colnames(data)) - 1])
```

Linear regression assumptions

To use linear models, we need to answer a question - are they adequate for this task? In order to do this, we will verify compliance with linear model assumptions

1. There is a linear relationship between the predictors (x) and the outcome (y)
2. Residual Errors have a mean value of zero
3. Predictors (x) are independent and observed with negligible error
4. Residual Errors have constant variance
5. Residual Errors are independent from each other and predictors (x)

```r
assumption_test_model = lm(train[,Y_colname] ~ ., data = train[, X_colnames])
assumption_test_model_log = lm(log(train[,Y_colname]) ~ ., data = train[, X_colnames])

assumption_test_model_sqrt = lm(
  sqrt(train[,Y_colname]) ~ .,
  data = train[, X_colnames]
)
```

Assumption 1 and 2: Check linearity of the data and residual errors have zero mean value

To check, if there is a linear relationship between target and predictors (linearity of the data), we will use Residuals VS Fitted plot

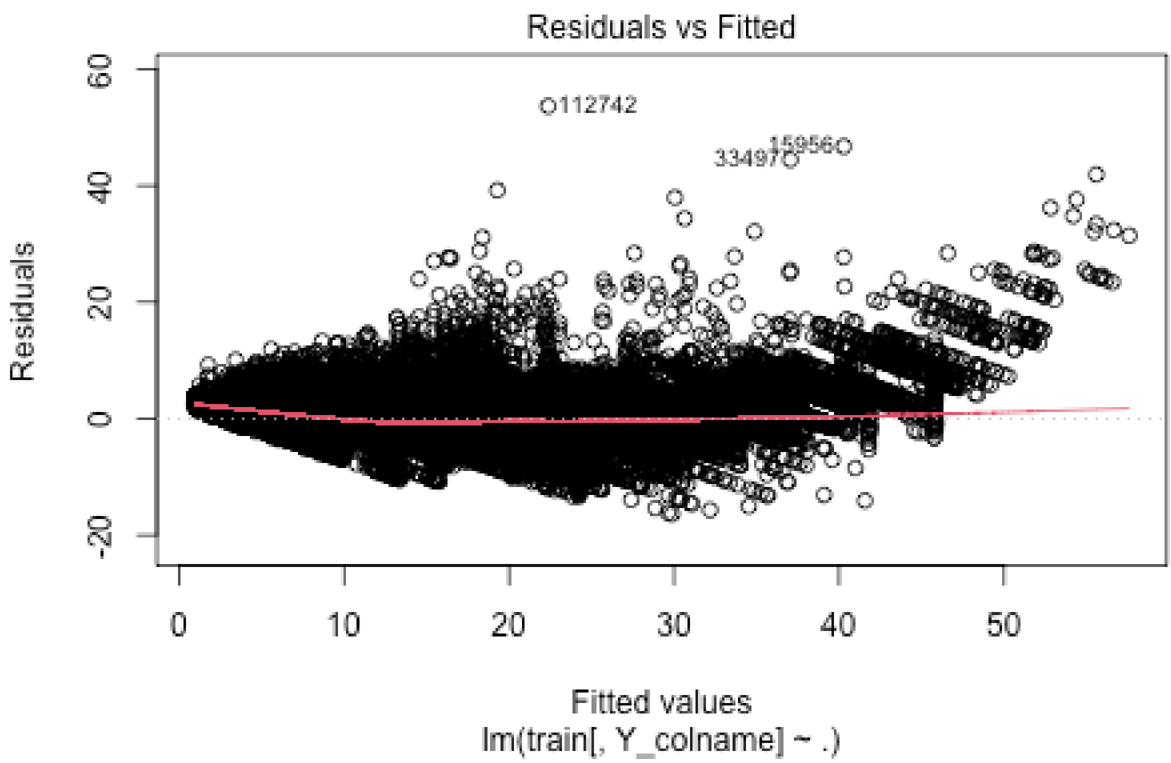
The graphs below show that the GLM model with log(Y) transformation has centric residuals around zero, without funnel-shaped, whereas the complete linear model shows the decrease in residues.

This suggests the suitability of log(Y) transformation using a generalized linear model.

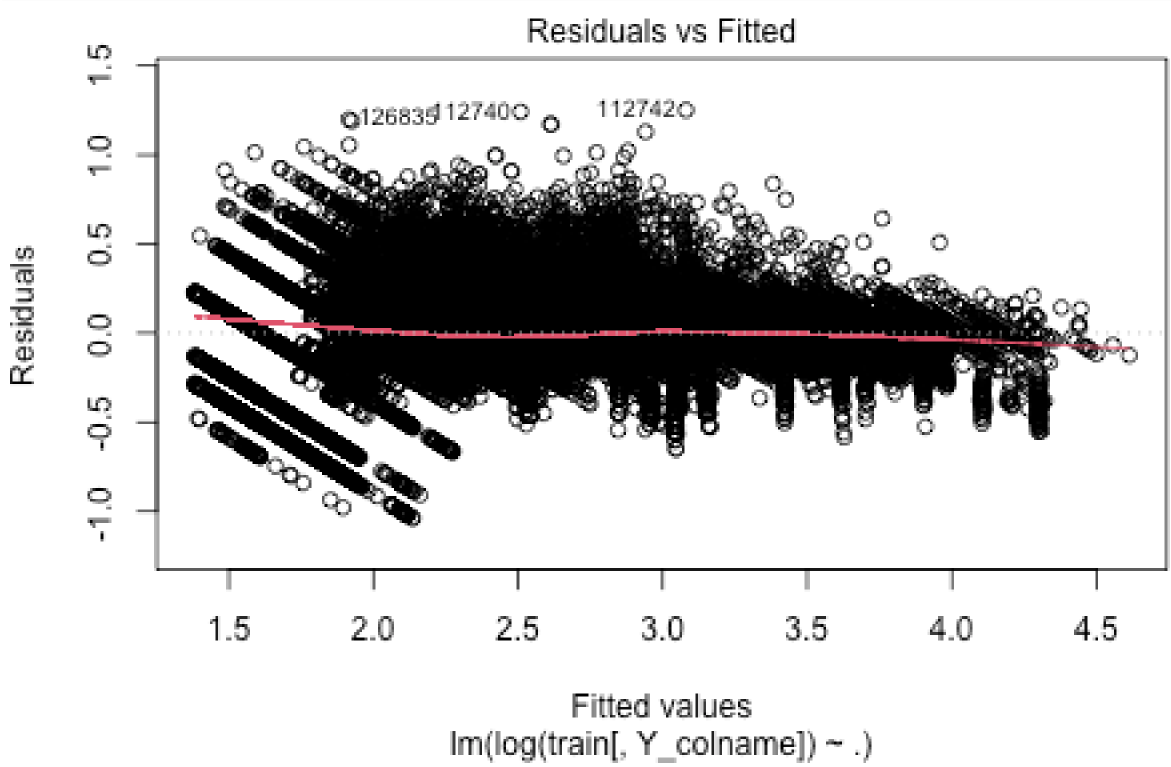
We have to pay attention to the Y-axis scale, which is where the difference is clear.

and from the residual plots below, we can see that GLM model has almost zero mean of residuals.

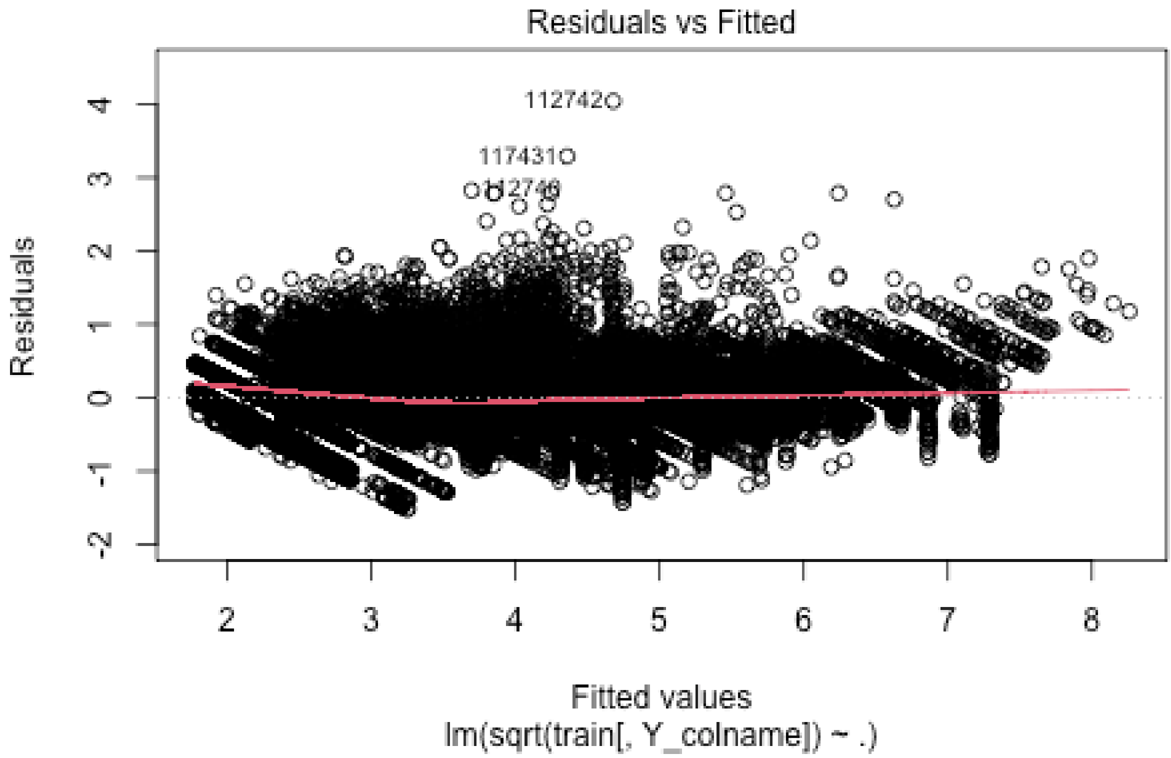
```r
par(mfrow = c(1, 1))

plot(assumption_test_model, 1)
```




```
plot(assumption_test_model_log, 1)
```




```
plot(assumption_test_model_sqrt, 1)
```




```r
par(mfrow = c(1, 1))
```

Assumption 3: Independence of predictors

Independence of predictors means predictors are independent and observed with negligible error.

For this, we use Durbin Watson test. The null hypothesis of the test states that there is no auto-correlation of residuals. Implicitly, our target is not enough evidence for rejecting H0 hypotheses.

We perform the test for full linear and GLM models. As seen below, both statistics give evidence in favor of correlated residuals, which is an argument against using GLM and linear models.

```r
library(car)
```

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

```r
durbinWatsonTest(assumption_test_model)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.00152302      1.996947   0.442
##  Alternative hypothesis: rho != 0
```

```r
durbinWatsonTest(assumption_test_model_log)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1    0.0002791436       1.99944   0.902
##  Alternative hypothesis: rho != 0
```

```r
durbinWatsonTest(assumption_test_model_sqrt)
```

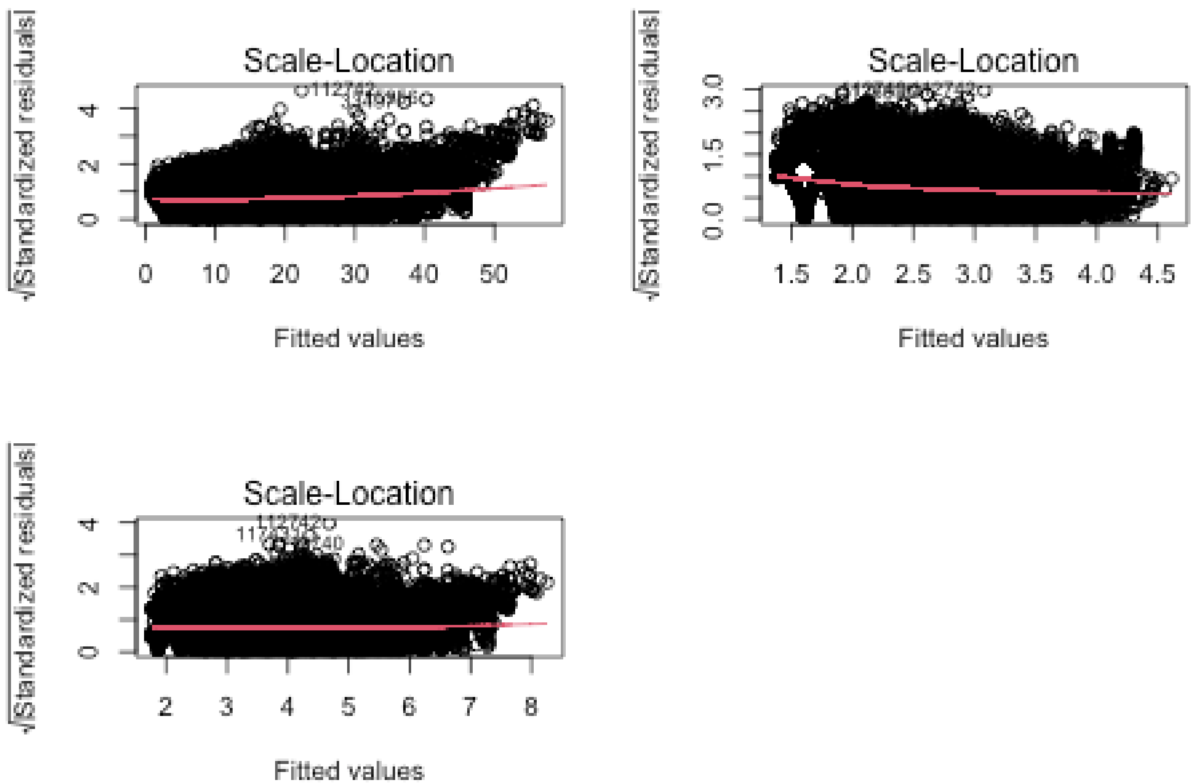
```
##  lag Autocorrelation D-W Statistic p-value
##    1     0.001359051      1.997277   0.584
##  Alternative hypothesis: rho != 0
```

Assumption 4: residual errors have constant variance

It seems this assumption is not met for GLM model and GLM model with sqrt transformation but is satified for GLM model with log transformations.

The red line is roughly horizontal across the plot. If it is, then the assumption of homoscedasticity is satisfied for a given regression model. That is, the spread of the residuals is roughly equal at all fitted values therefore variance is constant across the entire range.

```
par(mfrow = c(2, 2))

plot(assumption_test_model, 3)
plot(assumption_test_model_log, 3)
plot(assumption_test_model_sqrt, 3)

par(mfrow = c(1, 1))
```



We can use Breusch-Pagan Test to verify if homoscedasticity is met.

A Breusch-Pagan Test uses the following null and alternative hypotheses: Null Hypothesis (H0): The residuals are homoscedastic (i.e. evenly spread) Alternative Hypothesis (HA): The residuals are heteroscedastic (i.e. not evenly spread)

From the output we can see that the p-value of the test is less than 0.05, we fail to reject the null hypothesis. We have sufficient evidence to say that heteroscedasticity is present in the regression model.

```
#load lmtest package
library(lmtest)
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
#perform Breusch-Pagan Test
bptest(assumption_test_model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  assumption_test_model
## BP = 22000, df = 84, p-value < 2.2e-16
```

```r
bptest(assumption_test_model_log)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  assumption_test_model_log
## BP = 23398, df = 84, p-value < 2.2e-16
```

```r
bptest(assumption_test_model_sqrt)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  assumption_test_model_sqrt
## BP = 10641, df = 84, p-value < 2.2e-16
```

In conclusion, use of linear models and GLMs can be used for this task. However, for the log(Y) transformation, this model meets several assumptions. So, we will consider this model and compare it with the rest.

In the cells below, let us train several linear models. We leverage a range of linear models, with and without parameters selection techniques.

We experiment with following models: - Full linear model - Poisson GLM (log transform of target variable) - Backward and forward coefficients selection - Best model, based on Bayesian Information Criterion - Best model, based on Mallow's Cp coefficient

After training the models, we compare and select the best in terms of: - Adjuster R squared - Akaike Information Criterion - Bayesian Information Criterion - Number of paramters - Validation R squared

Furthermore, we share our considerations as to which criteria should be prioritized when choosing the model.

1. Full linear model

```r
full_model = lm(train[,Y_colname] ~ ., data = train[, X_colnames])
```

Model summary

The adjusted R squared for full linear model is 0.9247. With the following techniques, we will try to improve this metrics while decreasing the number of parameters considered.

```r
summary(full_model)
```

```
##
## Call:
```

```
## lm(formula = train[, Y_colname] ~ ., data = train[, X_colnames])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.380  -1.426  -0.152   1.262  53.628
##
## Coefficients: (17 not defined because of singularities)
##                                                          Estimate Std. Error
## (Intercept)                                            -5.561e+03  3.238e+04
## timestamp                                               3.443e-07  5.776e-06
## hour                                                   -2.151e-03  2.093e-02
## day                                                    -6.087e-02  5.006e-01
## month                                                  -1.673e+00  1.502e+01
## distance                                                2.893e+00  6.873e-03
## surge_multiplier                                        6.774e+01  2.331e-01
## latitude                                                1.569e+00  4.621e+00
## longitude                                              -1.059e+00  6.482e+00
## temperature                                            -5.472e-02  3.785e-02
## apparentTemperature                                     2.022e-02  1.250e-02
## precipIntensity                                        -5.222e-01  9.800e-01
## precipProbability                                       3.057e-01  5.182e-01
## humidity                                               -1.327e+00  1.194e+00
## windSpeed                                               7.426e-03  1.812e-02
## windGust                                               -1.207e-03  7.385e-03
## windGustTime                                           -1.662e-06  2.095e-06
## visibility                                              3.071e-02  1.728e-02
## temperatureHigh                                        -4.547e-03  1.207e-01
## temperatureHighTime                                     3.241e-07  3.447e-05
## temperatureLow                                         -1.081e-02  8.186e-02
## temperatureLowTime                                     -3.185e-06  3.562e-06
## apparentTemperatureHigh                                -2.199e-02  9.931e-02
## apparentTemperatureHighTime                            -1.339e-05  8.728e-06
## apparentTemperatureLow                                  1.080e-02  2.189e-02
## apparentTemperatureLowTime                             -7.927e-07  3.613e-06
## dewPoint                                                4.486e-02  3.398e-02
## pressure                                                2.529e-03  4.293e-03
## windBearing                                             4.769e-04  2.241e-04
## cloudCover                                             -2.008e-01  9.776e-02
## uvIndex                                                 4.036e-03  1.868e-02
## ozone                                                  -9.888e-05  1.256e-03
## sunriseTime                                            -5.376e-03  2.666e-02
## sunsetTime                                                     NA         NA
## moonPhase                                              -1.882e+00  9.681e+00
## precipIntensityMax                                     -2.092e+01  1.383e+01
## uvIndexTime                                             5.391e-03  2.668e-02
## temperatureMin                                         -3.139e-02  7.192e-02
## temperatureMinTime                                      3.761e-07  5.712e-06
## temperatureMax                                                 NA         NA
## temperatureMaxTime                                             NA         NA
## apparentTemperatureMin                                 -1.859e-02  1.434e-02
## apparentTemperatureMinTime                              6.361e-06  4.569e-06
## apparentTemperatureMax                                         NA         NA
## apparentTemperatureMaxTime                                     NA         NA
## V1_Lyft                                                -3.678e+00  2.920e-02
```

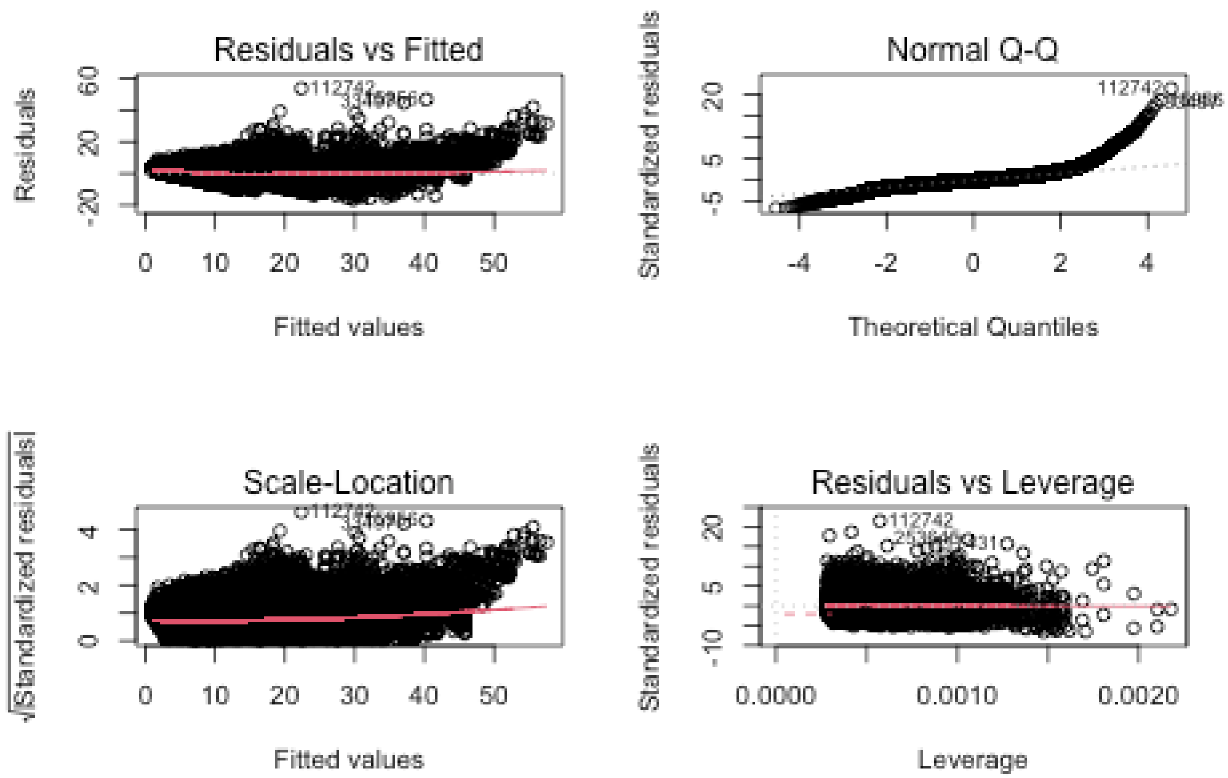
```
## V1_Uber                                                          NA         NA
## V1_Back.Bay                                                5.798e-02  2.922e-02
## V1_Beacon.Hill                                            -2.575e-01  2.916e-02
## V1_Boston.University                                      -8.258e-02  4.015e-02
## V1_Fenway                                                 -4.477e-01  3.993e-02
## V1_Financial.District                                      4.810e-01  2.928e-02
## V1_Haymarket.Square                                        9.408e-02  3.955e-02
## V1_North.End                                              -3.089e-02  3.948e-02
## V1_North.Station                                           2.705e-01  2.921e-02
## V1_Northeastern.University                                -9.175e-02  3.985e-02
## V1_South.Station                                          -1.849e-01  3.941e-02
## V1_Theatre.District                                        2.940e-01  2.908e-02
## V1_West.End                                                       NA         NA
## V1_.clear.day.                                            -2.328e-01  1.822e-01
## V1_.clear.night.                                          -1.435e-01  1.820e-01
## V1_.cloudy.                                               -1.284e-02  1.492e-01
## V1_.partly.cloudy.day.                                    -1.174e-01  1.687e-01
## V1_.partly.cloudy.night.                                  -1.002e-01  1.666e-01
## V1_.rain.                                                         NA         NA
## V1_.Light.rain.in.the.morning.and.overnight..             -1.967e+00  2.214e+00
## V1_.Light.rain.in.the.morning..                            1.109e-01  6.784e-01
## V1_.Light.rain.until.evening..                             7.464e-01  3.323e+00
## V1_.Mostly.cloudy.throughout.the.day..                     1.742e+01  9.588e+01
## V1_.Partly.cloudy.throughout.the.day..                    -2.376e+00  1.952e+00
## V1_.Rain.throughout.the.day..                              5.710e-01  3.312e+00
## V1_.Rain.until.morning..starting.again.in.the.evening..           NA         NA
## V1_Black                                                   1.080e+01  2.874e-02
## V1_Black.SUV                                               2.055e+01  2.868e-02
## V1_Lux                                                     1.104e+01  2.980e-02
## V1_Lux.Black                                               1.634e+01  2.980e-02
## V1_Lux.Black.XL                                            2.558e+01  2.976e-02
## V1_Lyft.1                                                  2.830e+00  2.976e-02
## V1_Lyft.XL                                                 8.555e+00  2.976e-02
## V1_Shared                                                         NA         NA
## V1_UberPool                                               -9.688e-01  2.861e-02
## V1_UberX                                                   2.863e-02  2.872e-02
## V1_UberXL                                                  5.948e+00  2.871e-02
## V1_WAV                                                            NA         NA
## V1_.Clear.                                                        NA         NA
## V1_.Drizzle.                                              -2.274e-03  2.085e-01
## V1_.Mostly.Cloudy.                                         6.047e-02  4.031e-02
## V1_.Overcast.                                                     NA         NA
## V1_.Partly.Cloudy.                                                NA         NA
## V1_.Possible.Drizzle.                                             NA         NA
## V1_Back.Bay.1                                             -8.406e-02  2.917e-02
## V1_Beacon.Hill.1                                          -3.460e-01  2.915e-02
## V1_Boston.University.1                                    -4.876e-01  3.043e-02
## V1_Fenway.1                                               -2.864e-01  2.990e-02
## V1_Financial.District.1                                    3.293e-01  2.931e-02
## V1_Haymarket.Square.1                                      1.985e-01  2.957e-02
## V1_North.End.1                                             3.985e-01  2.928e-02
## V1_North.Station.1                                        -8.502e-03  2.907e-02
## V1_Northeastern.University.1                              -4.819e-01  2.972e-02
## V1_South.Station.1                                                NA         NA
```

```
## V1_Theatre.District.1                                4.543e-01  2.918e-02
## V1_West.End.1                                              NA         NA
##                                                      t value Pr(>|t|)
## (Intercept)                                           -0.172  0.86363
## timestamp                                              0.060  0.95247
## hour                                                  -0.103  0.91814
## day                                                   -0.122  0.90322
## month                                                 -0.111  0.91130
## distance                                             420.874  < 2e-16 ***
## surge_multiplier                                     290.577  < 2e-16 ***
## latitude                                               0.339  0.73424
## longitude                                             -0.163  0.87016
## temperature                                           -1.446  0.14831
## apparentTemperature                                    1.618  0.10566
## precipIntensity                                       -0.533  0.59411
## precipProbability                                      0.590  0.55530
## humidity                                              -1.112  0.26613
## windSpeed                                              0.410  0.68187
## windGust                                              -0.163  0.87017
## windGustTime                                          -0.793  0.42767
## visibility                                             1.777  0.07553 .
## temperatureHigh                                       -0.038  0.96995
## temperatureHighTime                                    0.009  0.99250
## temperatureLow                                        -0.132  0.89498
## temperatureLowTime                                    -0.894  0.37111
## apparentTemperatureHigh                               -0.221  0.82472
## apparentTemperatureHighTime                           -1.535  0.12484
## apparentTemperatureLow                                 0.493  0.62172
## apparentTemperatureLowTime                            -0.219  0.82636
## dewPoint                                               1.320  0.18674
## pressure                                               0.589  0.55588
## windBearing                                            2.128  0.03335 *
## cloudCover                                            -2.054  0.04001 *
## uvIndex                                                0.216  0.82890
## ozone                                                 -0.079  0.93724
## sunriseTime                                           -0.202  0.84018
## sunsetTime                                                NA       NA
## moonPhase                                             -0.194  0.84585
## precipIntensityMax                                    -1.513  0.13033
## uvIndexTime                                            0.202  0.83986
## temperatureMin                                        -0.436  0.66248
## temperatureMinTime                                     0.066  0.94751
## temperatureMax                                            NA       NA
## temperatureMaxTime                                        NA       NA
## apparentTemperatureMin                                -1.297  0.19471
## apparentTemperatureMinTime                             1.392  0.16378
## apparentTemperatureMax                                    NA       NA
## apparentTemperatureMaxTime                                NA       NA
## V1_Lyft                                             -125.970  < 2e-16 ***
## V1_Uber                                                   NA       NA
## V1_Back.Bay                                            1.984  0.04723 *
## V1_Beacon.Hill                                        -8.831  < 2e-16 ***
## V1_Boston.University                                  -2.057  0.03969 *
## V1_Fenway                                            -11.211  < 2e-16 ***
```

```
## V1_Financial.District                                   16.431  < 2e-16 ***
## V1_Haymarket.Square                                      2.379  0.01737 *  
## V1_North.End                                            -0.782  0.43396    
## V1_North.Station                                         9.260  < 2e-16 ***
## V1_Northeastern.University                              -2.303  0.02130 *  
## V1_South.Station                                        -4.692 2.71e-06 ***
## V1_Theatre.District                                     10.111  < 2e-16 ***
## V1_West.End                                                 NA       NA    
## V1_.clear.day.                                          -1.278  0.20117    
## V1_.clear.night.                                        -0.789  0.43031    
## V1_.cloudy.                                             -0.086  0.93146    
## V1_.partly.cloudy.day.                                  -0.696  0.48654    
## V1_.partly.cloudy.night.                                -0.601  0.54752    
## V1_.rain.                                                   NA       NA    
## V1_.Light.rain.in.the.morning.and.overnight..           -0.889  0.37415    
## V1_.Light.rain.in.the.morning..                          0.163  0.87016    
## V1_.Light.rain.until.evening..                           0.225  0.82227    
## V1_.Mostly.cloudy.throughout.the.day..                   0.182  0.85585    
## V1_.Partly.cloudy.throughout.the.day..                  -1.217  0.22348    
## V1_.Rain.throughout.the.day..                            0.172  0.86312    
## V1_.Rain.until.morning..starting.again.in.the.evening..     NA       NA    
## V1_Black                                               375.831  < 2e-16 ***
## V1_Black.SUV                                           716.554  < 2e-16 ***
## V1_Lux                                                 370.440  < 2e-16 ***
## V1_Lux.Black                                           548.322  < 2e-16 ***
## V1_Lux.Black.XL                                        859.416  < 2e-16 ***
## V1_Lyft.1                                               95.091  < 2e-16 ***
## V1_Lyft.XL                                             287.479  < 2e-16 ***
## V1_Shared                                                   NA       NA    
## V1_UberPool                                            -33.862  < 2e-16 ***
## V1_UberX                                                 0.997  0.31884    
## V1_UberXL                                              207.166  < 2e-16 ***
## V1_WAV                                                      NA       NA    
## V1_.Clear.                                                  NA       NA    
## V1_.Drizzle.                                            -0.011  0.99130    
## V1_.Mostly.Cloudy.                                       1.500  0.13356    
## V1_.Overcast.                                               NA       NA    
## V1_.Partly.Cloudy.                                          NA       NA    
## V1_.Possible.Drizzle.                                       NA       NA    
## V1_Back.Bay.1                                           -2.882  0.00395 ** 
## V1_Beacon.Hill.1                                       -11.868  < 2e-16 ***
## V1_Boston.University.1                                 -16.022  < 2e-16 ***
## V1_Fenway.1                                             -9.580  < 2e-16 ***
## V1_Financial.District.1                                 11.233  < 2e-16 ***
## V1_Haymarket.Square.1                                    6.713 1.92e-11 ***
## V1_North.End.1                                          13.609  < 2e-16 ***
## V1_North.Station.1                                      -0.292  0.76995    
## V1_Northeastern.University.1                           -16.215  < 2e-16 ***
## V1_South.Station.1                                          NA       NA    
## V1_Theatre.District.1                                   15.569  < 2e-16 ***
## V1_West.End.1                                               NA       NA    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.499 on 176241 degrees of freedom
## Multiple R-squared:  0.9284, Adjusted R-squared:  0.9283
## F-statistic: 2.719e+04 on 84 and 176241 DF,  p-value: < 2.2e-16
```

```r
par(mfrow = c(2, 2))
plot(full_model)
```

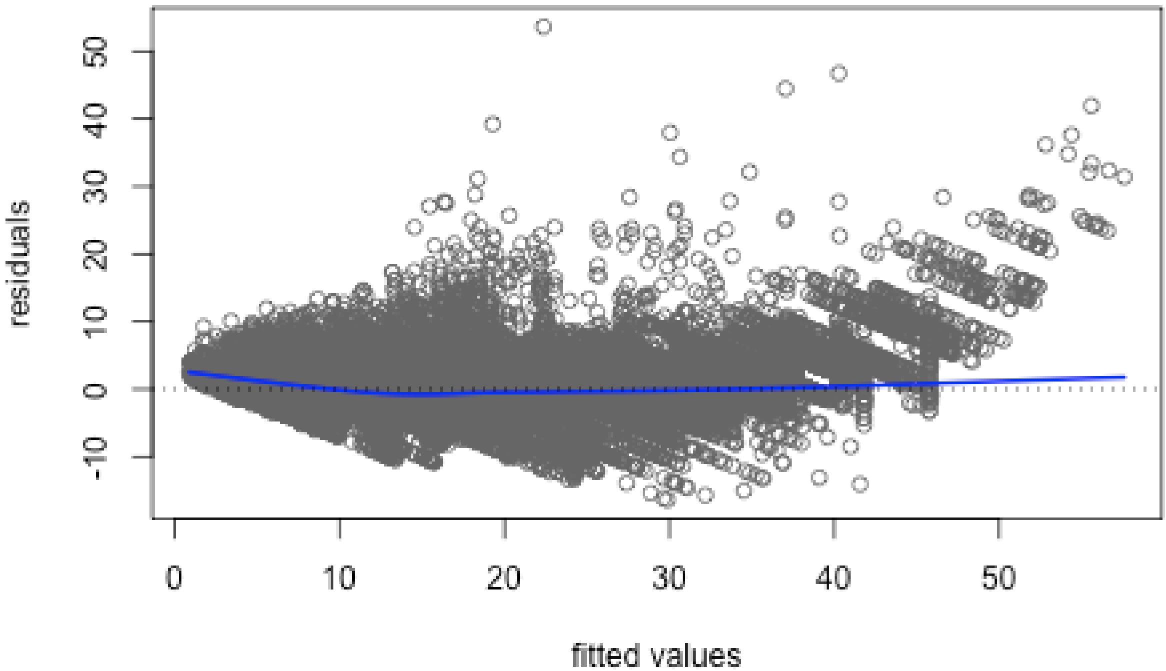



```r
par(mfrow = c(1, 1))
```

Model residuals plot

```r
plot_residuals <- function(model){
  plot(
    fitted(model),
    residuals(model),
    col = 'gray40',
    xlab = 'fitted values',
    ylab = 'residuals'
  )

  lines(
    loess.smooth(fitted(model), residuals(model)),
    col = "blue",
    lwd = 2
  )
  abline(h = 0, lty = 3)
}
```

```r
plot_residuals(full_model)
```

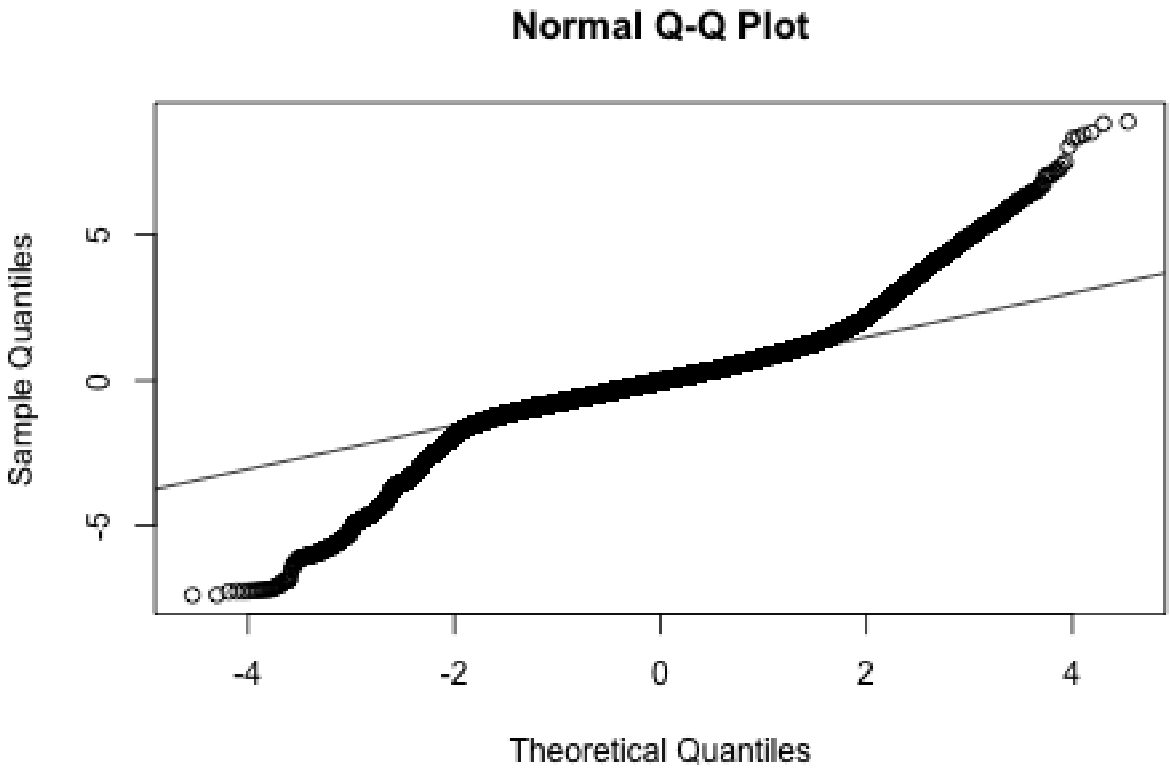


Full linear model (Y log transform) – Poisson GLM

```
full_model_log = lm(log(train[,Y_colname]) ~ ., data = train[, X_colnames])
```

Model summary

```
qqnorm(rstandard(full_model_log))
qqline(rstandard(full_model_log))
```



Log transformation helped to increase R squared up to 0.9386

```
summary(full_model_log)
```

```
##
## Call:
## lm(formula = log(train[, Y_colname]) ~ ., data = train[, X_colnames])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.03764 -0.07511 -0.00430  0.06871  1.24990
##
## Coefficients: (17 not defined because of singularities)
##                                                            Estimate Std. Error
## (Intercept)                                              -2.860e+03  1.824e+03
## timestamp                                                -1.864e-07  3.253e-07
## hour                                                      6.993e-04  1.179e-03
## day                                                       1.843e-02  2.819e-02
## month                                                     5.563e-01  8.458e-01
## distance                                                  1.754e-01  3.871e-04
## surge_multiplier                                          2.588e+00  1.313e-02
## latitude                                                  4.215e-01  2.603e-01
## longitude                                                -5.754e-01  3.651e-01
## temperature                                              -5.313e-04  2.132e-03
## apparentTemperature                                      -3.166e-04  7.039e-04
## precipIntensity                                          -1.170e-01  5.520e-02
## precipProbability                                         7.094e-02  2.919e-02
## humidity                                                 -3.743e-02  6.723e-02
## windSpeed                                                -5.345e-04  1.020e-03
## windGust                                                  1.572e-04  4.159e-04
## windGustTime                                             -1.263e-07  1.180e-07
```

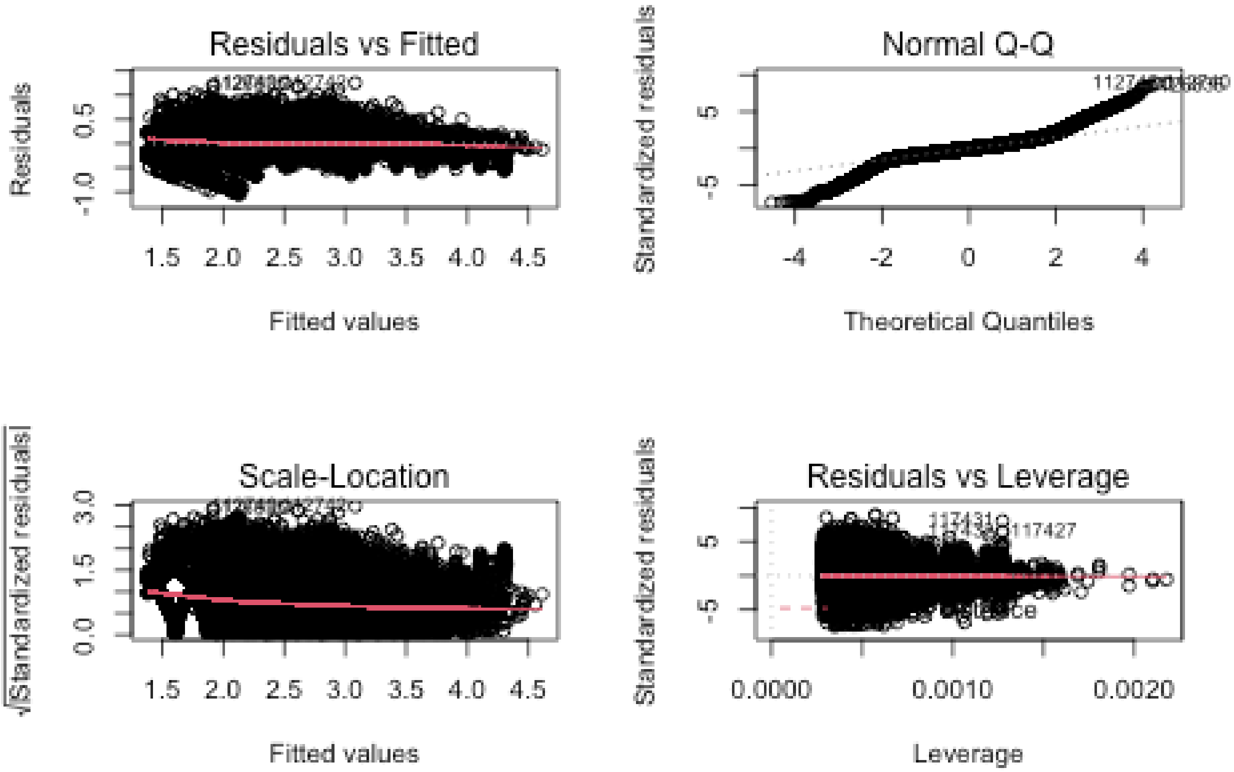
```
## visibility                                           1.749e-03  9.734e-04
## temperatureHigh                                      5.023e-04  6.798e-03
## temperatureHighTime                                  4.298e-06  1.942e-06
## temperatureLow                                      -9.469e-04  4.611e-03
## temperatureLowTime                                  -5.505e-08  2.006e-07
## apparentTemperatureHigh                             -4.708e-03  5.594e-03
## apparentTemperatureHighTime                         -1.217e-06  4.916e-07
## apparentTemperatureLow                               7.160e-04  1.233e-03
## apparentTemperatureLowTime                          -6.233e-08  2.035e-07
## dewPoint                                             9.736e-04  1.914e-03
## pressure                                             8.217e-05  2.418e-04
## windBearing                                          1.253e-05  1.262e-05
## cloudCover                                          -1.158e-02  5.506e-03
## uvIndex                                             -1.205e-03  1.052e-03
## ozone                                                1.030e-04  7.073e-05
## sunriseTime                                         -2.612e-03  1.501e-03
## sunsetTime                                                  NA         NA
## moonPhase                                            7.072e-01  5.452e-01
## precipIntensityMax                                   1.501e+00  7.790e-01
## uvIndexTime                                          2.611e-03  1.503e-03
## temperatureMin                                      -5.663e-04  4.051e-03
## temperatureMinTime                                   3.896e-07  3.217e-07
## temperatureMax                                              NA         NA
## temperatureMaxTime                                          NA         NA
## apparentTemperatureMin                              -9.076e-04  8.075e-04
## apparentTemperatureMinTime                           3.258e-07  2.573e-07
## apparentTemperatureMax                                      NA         NA
## apparentTemperatureMaxTime                                  NA         NA
## V1_Lyft                                             -5.125e-01  1.645e-03
## V1_Uber                                                     NA         NA
## V1_Back.Bay                                          9.367e-03  1.646e-03
## V1_Beacon.Hill                                       2.011e-03  1.642e-03
## V1_Boston.University                                -1.365e-02  2.261e-03
## V1_Fenway                                           -2.729e-02  2.249e-03
## V1_Financial.District                               -1.668e-02  1.649e-03
## V1_Haymarket.Square                                 -5.581e-03  2.228e-03
## V1_North.End                                        -5.179e-03  2.224e-03
## V1_North.Station                                     1.599e-03  1.645e-03
## V1_Northeastern.University                          -2.153e-03  2.244e-03
## V1_South.Station                                    -1.627e-02  2.220e-03
## V1_Theatre.District                                  2.729e-02  1.638e-03
## V1_West.End                                                 NA         NA
## V1_.clear.day.                                       6.019e-03  1.026e-02
## V1_.clear.night.                                     6.343e-03  1.025e-02
## V1_.cloudy.                                          1.571e-02  8.406e-03
## V1_.partly.cloudy.day.                               1.212e-02  9.502e-03
## V1_.partly.cloudy.night.                             1.027e-02  9.385e-03
## V1_.rain.                                                   NA         NA
## V1_.Light.rain.in.the.morning.and.overnight..        4.394e-02  1.247e-01
## V1_.Light.rain.in.the.morning..                     -2.497e-02  3.821e-02
## V1_.Light.rain.until.evening..                      -1.109e-01  1.872e-01
## V1_.Mostly.cloudy.throughout.the.day..               9.475e+00  5.400e+00
## V1_.Partly.cloudy.throughout.the.day..               6.270e-02  1.099e-01
## V1_.Rain.throughout.the.day..                       -3.368e-01  1.865e-01
```

```
## V1_.Rain.until.morning..starting.again.in.the.evening..         NA         NA
## V1_Black                                                 7.455e-01  1.618e-03
## V1_Black.SUV                                             1.149e+00  1.615e-03
## V1_Lux                                                   1.074e+00  1.679e-03
## V1_Lux.Black                                             1.341e+00  1.678e-03
## V1_Lux.Black.XL                                          1.691e+00  1.676e-03
## V1_Lyft.1                                                4.677e-01  1.676e-03
## V1_Lyft.XL                                               9.245e-01  1.676e-03
## V1_Shared                                                       NA         NA
## V1_UberPool                                             -1.065e-01  1.612e-03
## V1_UberX                                                 2.010e-03  1.618e-03
## V1_UberXL                                                4.653e-01  1.617e-03
## V1_WAV                                                          NA         NA
## V1_.Clear.                                                      NA         NA
## V1_.Drizzle.                                            -2.034e-02  1.175e-02
## V1_.Mostly.Cloudy.                                       2.365e-03  2.270e-03
## V1_.Overcast.                                                   NA         NA
## V1_.Partly.Cloudy.                                              NA         NA
## V1_.Possible.Drizzle.                                           NA         NA
## V1_Back.Bay.1                                           -3.009e-03  1.643e-03
## V1_Beacon.Hill.1                                        -3.867e-03  1.642e-03
## V1_Boston.University.1                                  -4.100e-02  1.714e-03
## V1_Fenway.1                                             -2.047e-02  1.684e-03
## V1_Financial.District.1                                 -3.368e-02  1.651e-03
## V1_Haymarket.Square.1                                   -1.550e-02  1.666e-03
## V1_North.End.1                                           1.989e-02  1.649e-03
## V1_North.Station.1                                      -7.880e-03  1.637e-03
## V1_Northeastern.University.1                            -2.724e-02  1.674e-03
## V1_South.Station.1                                              NA         NA
## V1_Theatre.District.1                                    2.793e-02  1.644e-03
## V1_West.End.1                                                   NA         NA
##                                                          t value Pr(>|t|)
## (Intercept)                                               -1.568   0.1169
## timestamp                                                 -0.573   0.5666
## hour                                                       0.593   0.5530
## day                                                        0.654   0.5133
## month                                                      0.658   0.5107
## distance                                                 453.165  < 2e-16 ***
## surge_multiplier                                         197.107  < 2e-16 ***
## latitude                                                   1.619   0.1054
## longitude                                                 -1.576   0.1150
## temperature                                               -0.249   0.8032
## apparentTemperature                                       -0.450   0.6528
## precipIntensity                                           -2.121   0.0340 *
## precipProbability                                          2.430   0.0151 *
## humidity                                                  -0.557   0.5777
## windSpeed                                                 -0.524   0.6004
## windGust                                                   0.378   0.7055
## windGustTime                                              -1.070   0.2846
## visibility                                                 1.797   0.0723 .
## temperatureHigh                                            0.074   0.9411
## temperatureHighTime                                        2.214   0.0269 *
## temperatureLow                                            -0.205   0.8373
## temperatureLowTime                                        -0.274   0.7838
```

```
## apparentTemperatureHigh                                  -0.842   0.4000
## apparentTemperatureHighTime                              -2.477   0.0133 *
## apparentTemperatureLow                                    0.581   0.5614
## apparentTemperatureLowTime                               -0.306   0.7594
## dewPoint                                                  0.509   0.6109
## pressure                                                  0.340   0.7340
## windBearing                                               0.993   0.3208
## cloudCover                                               -2.103   0.0355 *
## uvIndex                                                  -1.145   0.2521
## ozone                                                     1.457   0.1452
## sunriseTime                                              -1.740   0.0819 .
## sunsetTime                                                   NA       NA
## moonPhase                                                 1.297   0.1946
## precipIntensityMax                                        1.926   0.0541 .
## uvIndexTime                                               1.737   0.0823 .
## temperatureMin                                           -0.140   0.8888
## temperatureMinTime                                        1.211   0.2259
## temperatureMax                                               NA       NA
## temperatureMaxTime                                           NA       NA
## apparentTemperatureMin                                   -1.124   0.2611
## apparentTemperatureMinTime                                1.266   0.2054
## apparentTemperatureMax                                       NA       NA
## apparentTemperatureMaxTime                                   NA       NA
## V1_Lyft                                                -311.644  < 2e-16 ***
## V1_Uber                                                      NA       NA
## V1_Back.Bay                                               5.691 1.26e-08 ***
## V1_Beacon.Hill                                            1.224   0.2208
## V1_Boston.University                                     -6.037 1.57e-09 ***
## V1_Fenway                                               -12.133  < 2e-16 ***
## V1_Financial.District                                   -10.118  < 2e-16 ***
## V1_Haymarket.Square                                      -2.505   0.0122 *
## V1_North.End                                             -2.329   0.0199 *
## V1_North.Station                                          0.972   0.3310
## V1_Northeastern.University                               -0.959   0.3374
## V1_South.Station                                         -7.331 2.29e-13 ***
## V1_Theatre.District                                      16.662  < 2e-16 ***
## V1_West.End                                                  NA       NA
## V1_.clear.day.                                            0.587   0.5574
## V1_.clear.night.                                          0.619   0.5361
## V1_.cloudy.                                               1.869   0.0617 .
## V1_.partly.cloudy.day.                                    1.276   0.2021
## V1_.partly.cloudy.night.                                  1.094   0.2738
## V1_.rain.                                                    NA       NA
## V1_.Light.rain.in.the.morning.and.overnight..             0.352   0.7245
## V1_.Light.rain.in.the.morning..                          -0.654   0.5134
## V1_.Light.rain.until.evening..                           -0.593   0.5534
## V1_.Mostly.cloudy.throughout.the.day..                    1.755   0.0793 .
## V1_.Partly.cloudy.throughout.the.day..                    0.570   0.5684
## V1_.Rain.throughout.the.day..                            -1.806   0.0710 .
## V1_.Rain.until.morning..starting.again.in.the.evening..      NA       NA
## V1_Black                                                460.616  < 2e-16 ***
## V1_Black.SUV                                            711.541  < 2e-16 ***
## V1_Lux                                                  640.015  < 2e-16 ***
## V1_Lux.Black                                            798.926  < 2e-16 ***
```

```
## V1_Lux.Black.XL                                    1008.830  < 2e-16 ***
## V1_Lyft.1                                           278.988  < 2e-16 ***
## V1_Lyft.XL                                          551.570  < 2e-16 ***
## V1_Shared                                                NA       NA
## V1_UberPool                                         -66.098  < 2e-16 ***
## V1_UberX                                              1.243   0.2140
## V1_UberXL                                           287.739  < 2e-16 ***
## V1_WAV                                                   NA       NA
## V1_.Clear.                                               NA       NA
## V1_.Drizzle.                                         -1.732   0.0833 .
## V1_.Mostly.Cloudy.                                    1.042   0.2975
## V1_.Overcast.                                            NA       NA
## V1_.Partly.Cloudy.                                       NA       NA
## V1_.Possible.Drizzle.                                    NA       NA
## V1_Back.Bay.1                                        -1.831   0.0670 .
## V1_Beacon.Hill.1                                     -2.356   0.0185 *
## V1_Boston.University.1                              -23.917  < 2e-16 ***
## V1_Fenway.1                                         -12.154  < 2e-16 ***
## V1_Financial.District.1                             -20.397  < 2e-16 ***
## V1_Haymarket.Square.1                                -9.308  < 2e-16 ***
## V1_North.End.1                                       12.061  < 2e-16 ***
## V1_North.Station.1                                   -4.812 1.49e-06 ***
## V1_Northeastern.University.1                        -16.270  < 2e-16 ***
## V1_South.Station.1                                       NA       NA
## V1_Theatre.District.1                                16.993  < 2e-16 ***
## V1_West.End.1                                            NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1408 on 176241 degrees of freedom
## Multiple R-squared:  0.9387, Adjusted R-squared:  0.9387
## F-statistic: 3.213e+04 on 84 and 176241 DF,  p-value: < 2.2e-16
```

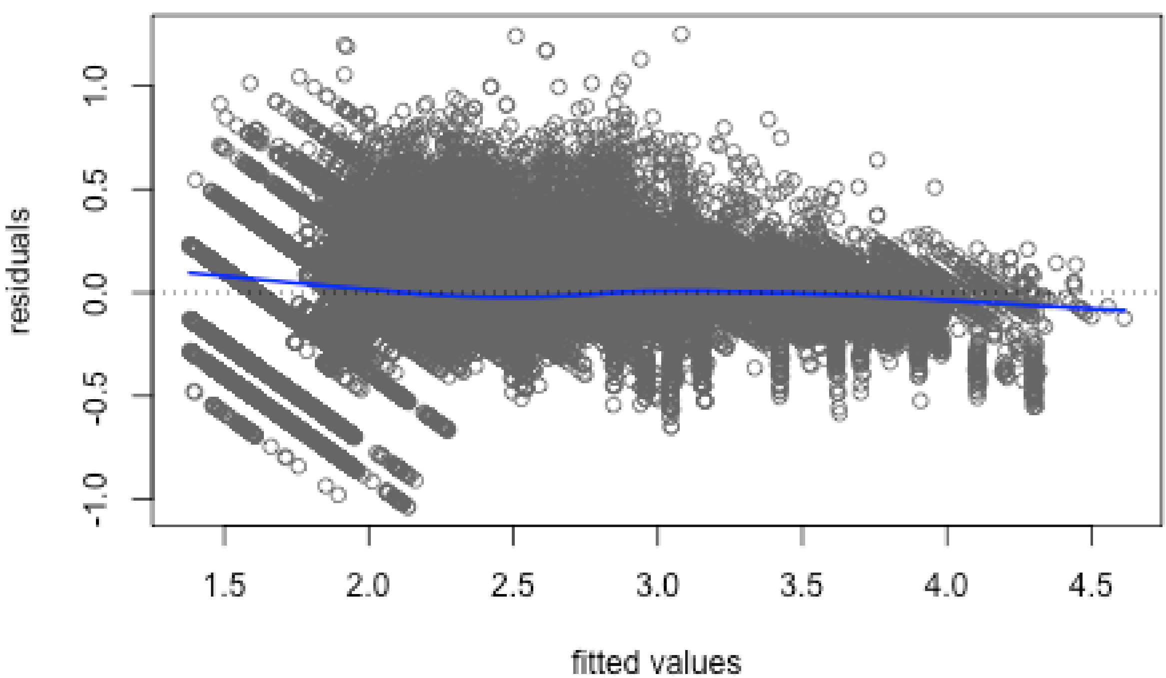
```r
par(mfrow = c(2, 2))
plot(full_model_log)
```




```
par(mfrow = c(1, 1))
```

Also, we note that the residuals are centered closer around zero, compared to full linear model.

```
plot_residuals(full_model_log)
```



Linear model (forward / backward model selection)

However, we can reduce variance of the model by selecting both backward and forward model. Generally, backward selection works better but we also keep the forward selection for comparison.

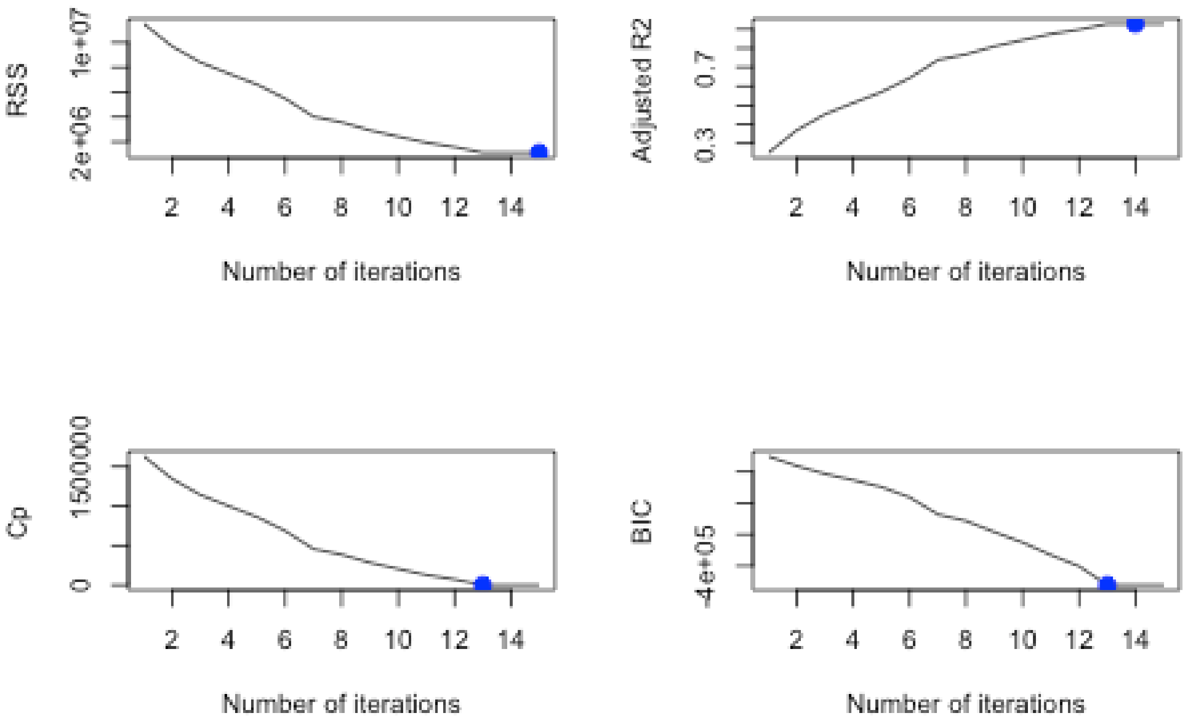
```r
library(leaps)

correlated_indices = match(c("timestamp", "hour", "day", "month", "apparentTemperature", "precipIntensi
stopifnot(sum(is.na(correlated_indices)) == 0)
nvmax = 15

forward_subsets = regsubsets(
  price~.,
  data = train[, -correlated_indices],
  method = 'forward',
  nvmax = nvmax
)

backward_subsets = regsubsets(
  price~.,
  data = train[, -correlated_indices],
  method = 'backward',
  nvmax = nvmax
)
```

Model summary

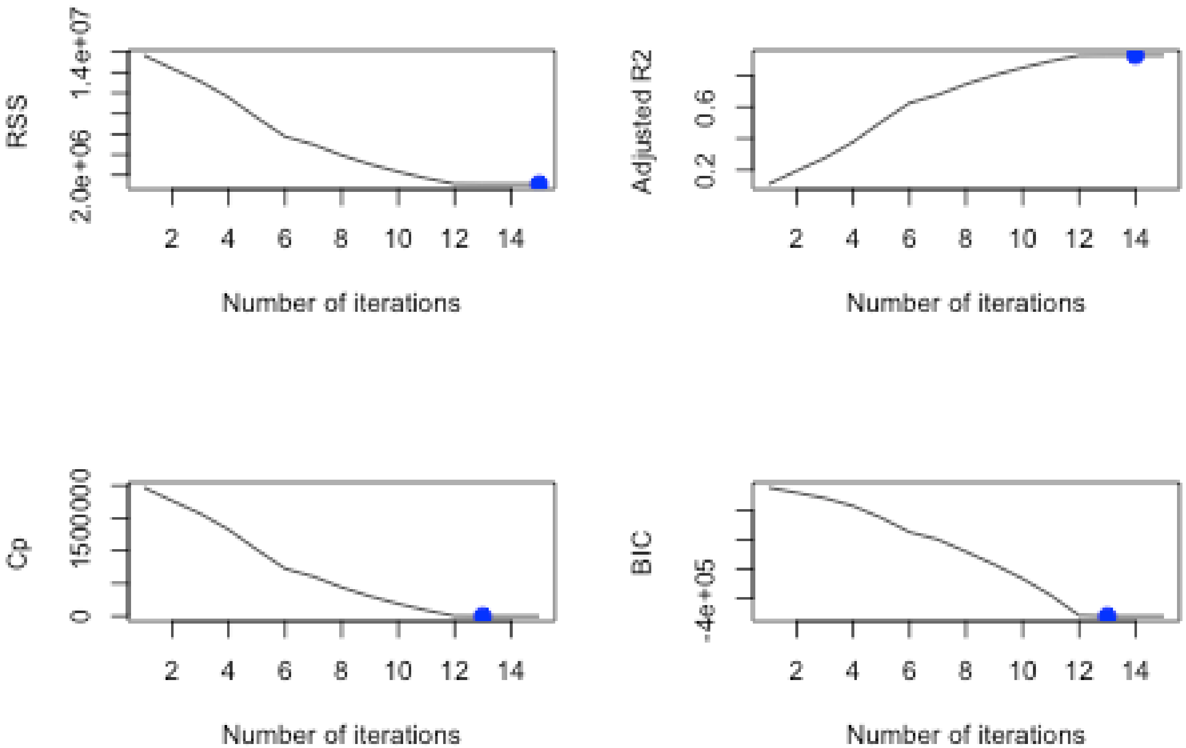
```r
plot_max_point <- function(values){
  max_idx = which.max(values)
  points(max_idx, values[max_idx], col = 'blue', cex = 2, pch = 20)
}

plot_min_point <- function(values){
  min_idx = which.min(values)
  points(min_idx, values[min_idx], col = 'blue', cex = 2, pch = 20)
}

plot_subsets_summary <- function(subsets_model){
  subset_summary = summary(subsets_model)

  xlabel = 'Number of iterations'
  linetype = 'l'

  par(mfrow = c(2, 2))

  plot(subset_summary$rss, xlab = xlabel, ylab = 'RSS', type = linetype)
  plot_min_point(values = subset_summary$rss)

  plot(subset_summary$adjr2, xlab = xlabel, ylab = 'Adjusted R2', type = linetype)
  plot_max_point(values = subset_summary$adjr2)

  plot(subset_summary$cp, xlab = xlabel, ylab = 'Cp', type = linetype)
  plot_min_point(values = subset_summary$cp)

  plot(subset_summary$bic, xlab = xlabel, ylab = 'BIC', type = 'l')
  plot_min_point(values = subset_summary$bic)
```

```
  # max_idx = which.max(subset_summary$adjr2)
  # points(max_idx, reg.summary$adjr2[max_idx], col="red",cex=2,pch=20)

  par(mfrow = c(1, 1))
}
```

```
plot_subsets_summary(forward_subsets)
```




```
plot_subsets_summary(backward_subsets)
```



In the following code cells„ we aim to select best models after forward and backward elimination - in terms of BIC and Mallow's Cp coefficients.

```
get_best_n_params <- function(subsets_model){
  model_summary = summary(subsets_model)
  best_bic = which.min(model_summary$bic)
  best_cp = which.min(model_summary$cp)
  return (list('best_bic' = best_bic, 'best_cp' = best_cp))
}

get_best_coefficients <- function(subsets_model){
  best_params = get_best_n_params(subsets_model)

  n_best_bic = best_params$best_bic
  n_best_cp = best_params$best_cp

  best_bic_coefs = names(coefficients(subsets_model, n_best_bic))[-1]
  best_cp_coefs = names(coefficients(subsets_model, n_best_cp))[-1]

  return(list("best_bic_coefs" = best_bic_coefs, "best_cp_coefs" = best_cp_coefs))
}

best_params_back = get_best_n_params(backward_subsets)
best_params_forward = get_best_n_params(forward_subsets)

n_best_bic_back = best_params_back$best_bic
n_best_cp_back = best_params_back$best_cp

n_best_bic_fwd = best_params_forward$best_bic
n_best_cp_fwd = best_params_forward$best_cp
```

```r
cat('Best # of parameters (backward)', n_best_bic_back, n_best_cp_back, '\n')
```

```
## Best # of parameters (backward) 13 13
```

```r
cat('Best # of parameters (forward)', n_best_bic_fwd, n_best_cp_fwd, '\n')
```

```
## Best # of parameters (forward) 13 13
```

```r
best_forward_coefs = get_best_coefficients(forward_subsets)
best_backward_coefs = get_best_coefficients(backward_subsets)

bic_forward_coefs = best_forward_coefs$best_bic_coefs
bic_backward_coefs = best_backward_coefs$best_bic_coefs

cp_forward_coefs = best_forward_coefs$best_cp_coefs
cp_backward_coefs = best_backward_coefs$best_cp_coefs
```

Here are the chosen coefficients after model selection:

```r
print(cp_backward_coefs)
```

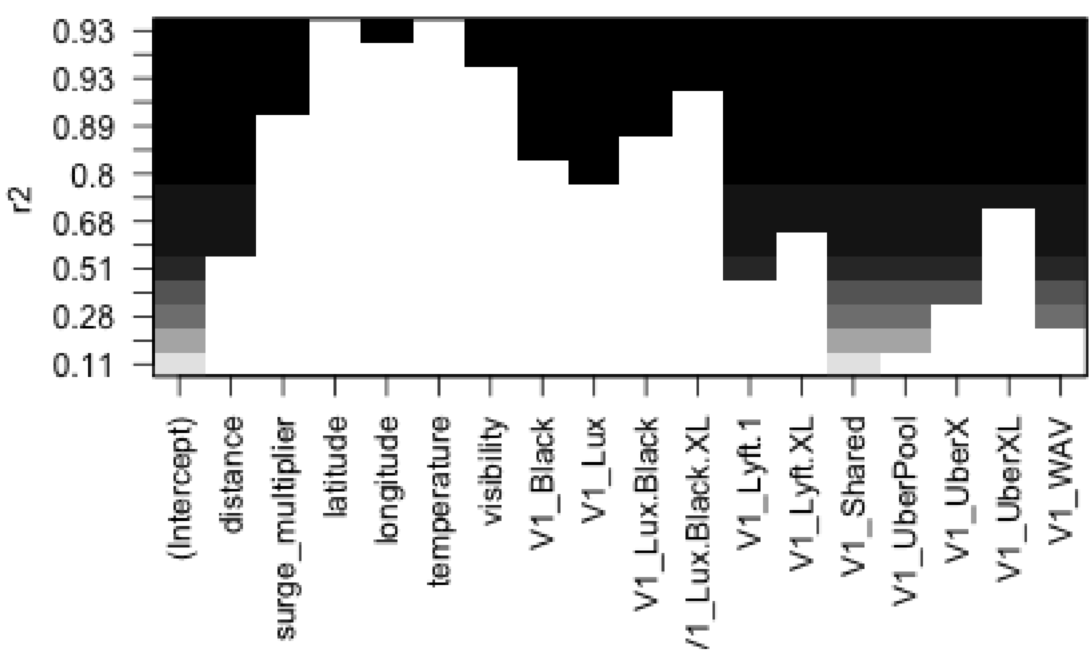
```
##  [1] "distance"         "surge_multiplier" "V1_Black"         "V1_Lux"
##  [5] "V1_Lux.Black"     "V1_Lux.Black.XL"  "V1_Lyft.1"        "V1_Lyft.XL"
##  [9] "V1_Shared"        "V1_UberPool"      "V1_UberX"         "V1_UberXL"
## [13] "V1_WAV"
```

```r
print(bic_backward_coefs)
```

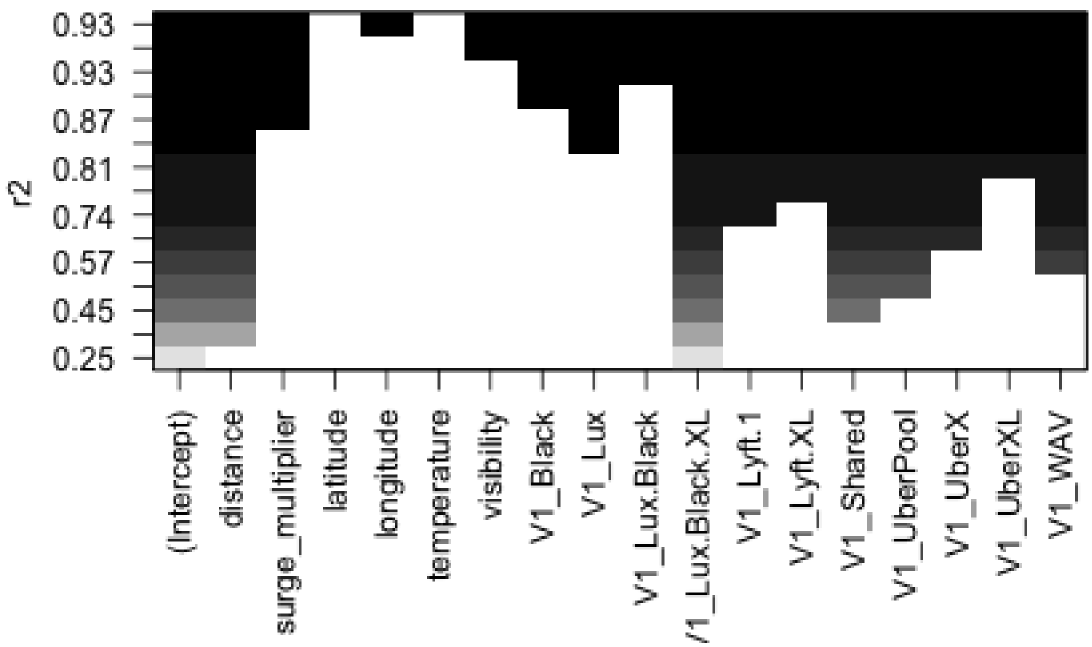
```
##  [1] "distance"         "surge_multiplier" "V1_Black"         "V1_Lux"
##  [5] "V1_Lux.Black"     "V1_Lux.Black.XL"  "V1_Lyft.1"        "V1_Lyft.XL"
##  [9] "V1_Shared"        "V1_UberPool"      "V1_UberX"         "V1_UberXL"
## [13] "V1_WAV"
```

To represent the most significant variables, here we report the variable elimination plots:

```r
plot(backward_subsets, scale = 'r2')
```



```r
plot(forward_subsets, scale = 'r2')
```



After elimination of variables, let us retrain best BIC and Cp model:

```r
bic_best_model = lm(train[, Y_colname] ~ ., data = train[, bic_backward_coefs])
```

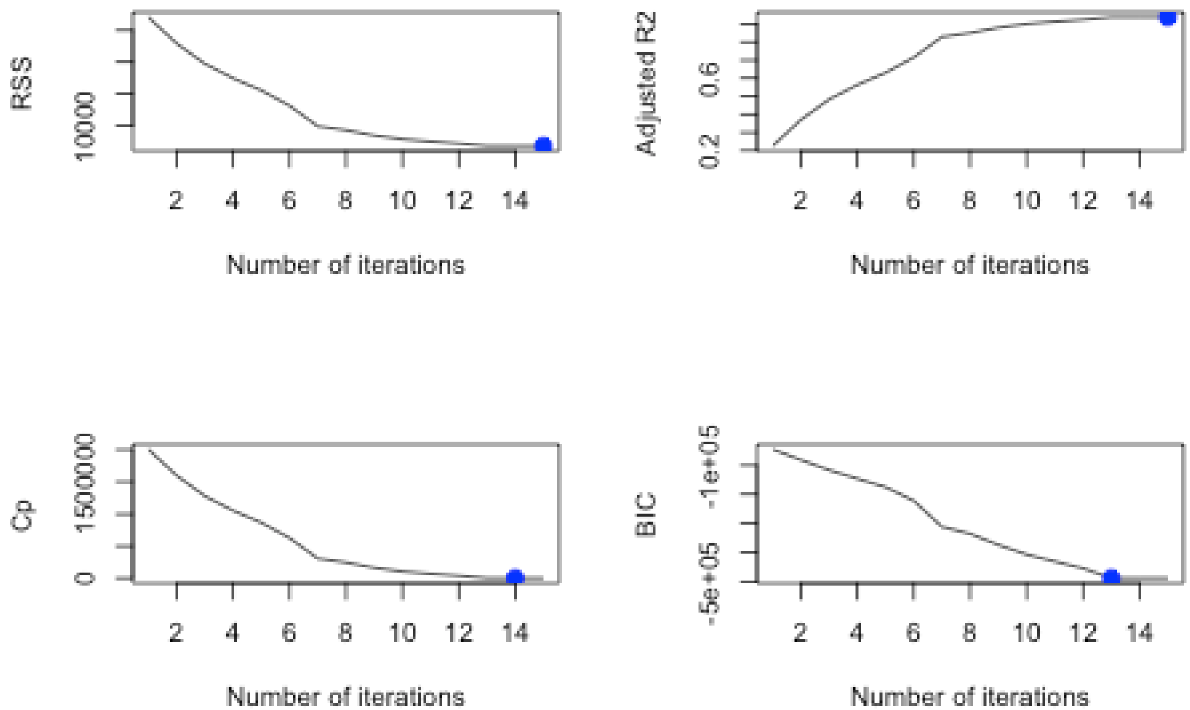
```r
cp_best_model = lm(train[, Y_colname] ~ ., data = train[, cp_backward_coefs])
```

Forward / backward selection for log(Y) transform

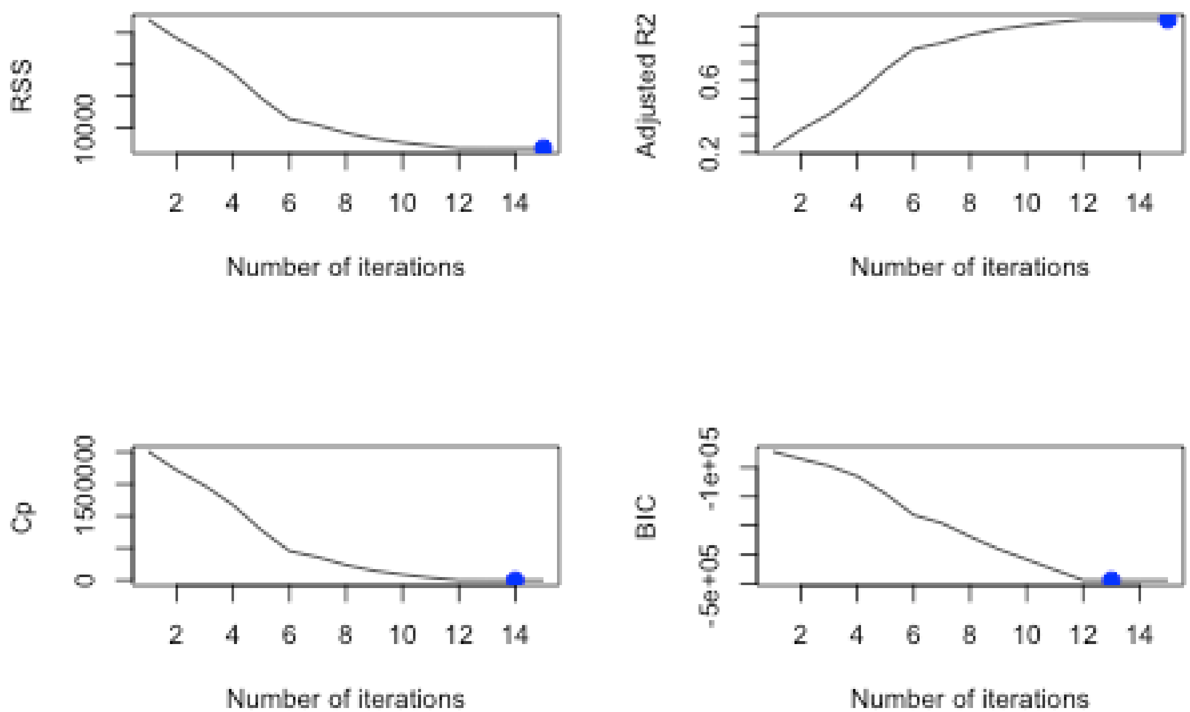
Here we reiterate the same procedure, taking Log() transformation of target

```r
nvmax = 15

log_forward_subsets = regsubsets(
  log(price) ~ .,
  data = train[, -correlated_indices],
  method = 'forward',
  nvmax = nvmax
)

log_backward_subsets = regsubsets(
  log(price) ~ .,
  data = train[, -correlated_indices],
  method = 'backward',
  nvmax = nvmax
)
```

```r
plot_subsets_summary(log_forward_subsets)
```




```r
plot_subsets_summary(log_backward_subsets)
```




```r
log_best_coefficients = get_best_coefficients(log_backward_subsets)
log_bic_backward_coefs = log_best_coefficients$best_bic_coefs
log_cp_backward_coefs = log_best_coefficients$best_cp_coefs
```

```r
cat(
  'Number of best coefficients',
  length(log_bic_backward_coefs),
```

```
  length(log_cp_backward_coefs),
  '\n'
)
```

```
## Number of best coefficients 13 14
```

```
log_bic_backward_coefs
```

```
##  [1] "distance"         "surge_multiplier" "V1_Black"         "V1_Lux"
##  [5] "V1_Lux.Black"     "V1_Lux.Black.XL"  "V1_Lyft.1"        "V1_Lyft.XL"
##  [9] "V1_Shared"        "V1_UberPool"      "V1_UberX"         "V1_UberXL"
## [13] "V1_WAV"
```

```
log_cp_backward_coefs
```

```
##  [1] "distance"         "surge_multiplier" "visibility"       "V1_Black"
##  [5] "V1_Lux"           "V1_Lux.Black"     "V1_Lux.Black.XL"  "V1_Lyft.1"
##  [9] "V1_Lyft.XL"       "V1_Shared"        "V1_UberPool"      "V1_UberX"
## [13] "V1_UberXL"        "V1_WAV"
```

```
log_cp_best_model = lm(log(train[, Y_colname]) ~ ., data = train[, log_cp_backward_coefs])
```

```
log_bic_best_model = lm(log(train[, Y_colname]) ~ ., data = train[, log_bic_backward_coefs])
```

Comparison of models

How do we choose proper criteria for comparing models? We have several options: 1. R^2 (adjusted) on validation set 2. R^2 (adjusted) on train set 3. Bayesian Information Criteria 4. Akaike information criteria 5. Mallow's Cp criteria

We select the proper characteristics, according to following considerations: 1. Cp, AIC, BIC and adjusted R^2 - all these account for both good fit and simplicity of the model 2. Validation R^2 parameter is preferred, as it shows the performance on the unseen data. Ideally, a cross-validation adjusted R^2 metrics will give better understanding 3. Mallow's Cp and AIC are equivalent and result in the selection of the same model 4. We are looking to minimize Cp, AIC and BIC, but to maximize R^2-related metrics 5. BIC statistics penalizer big models heavier that Cp and AIC –> BIC criterion results in selection of more 'lightweight' model

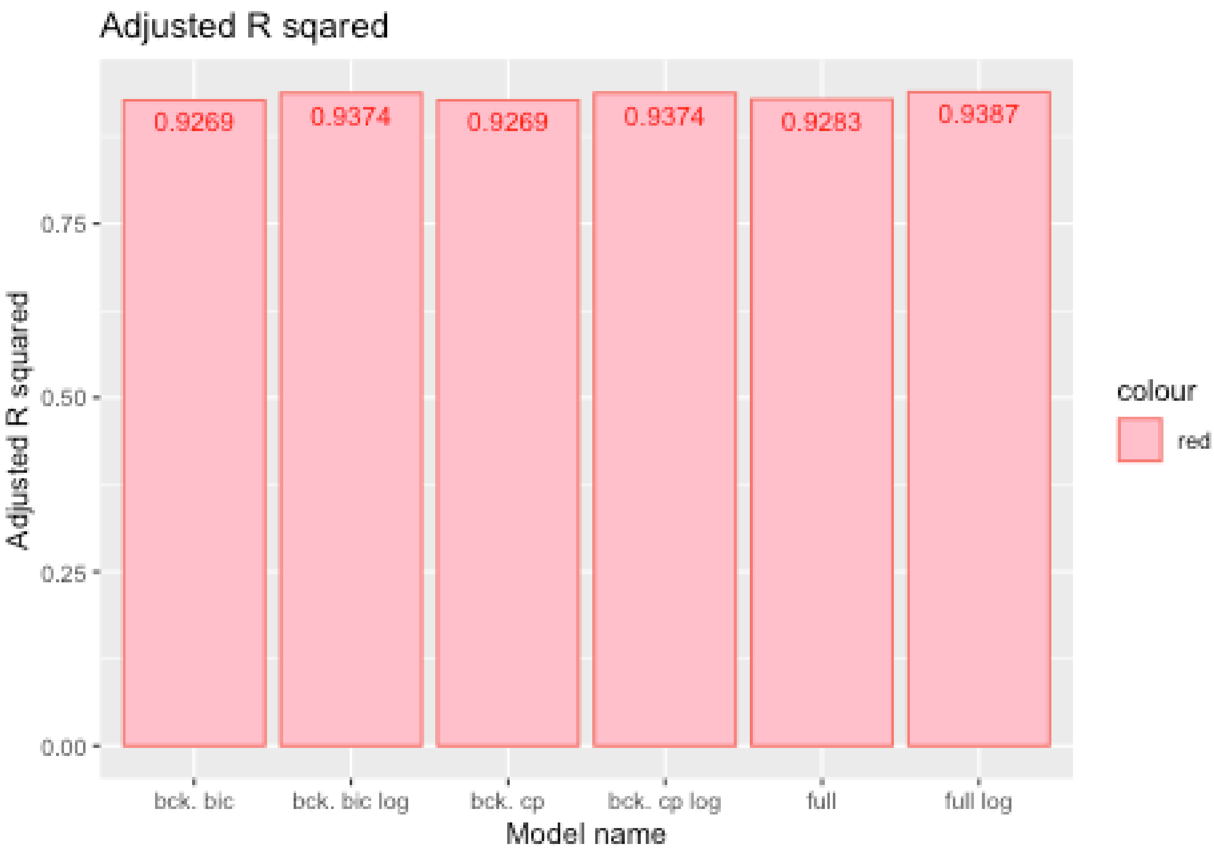
According to this, we will compare the models based on following characteristics: 1. (Cross) validation adjusted R^2 2. AIC (or, equivalently, Mallow's Cp) 3. BIC (BIC criterion will be of the highest priority, if we target at the simplest model possible)

```
models = list(
  full_model = list(name = 'full', model = full_model),
  full_model_log = list(name = 'full log', model = full_model_log),

  backward_bic_best_model = list(name = 'bck. bic', model = bic_best_model),
  backward_cp_best_model = list(name = 'bck. cp', model = cp_best_model),

  backward_cp_best_model_log = list(name = 'bck. cp log',
                                    model = log_cp_best_model),
  backward_bic_best_model_log = list(name = 'bck. bic log',
                                     model = log_bic_best_model)
)
```

Comparison of adjusted R^2

Main conclusion that can be made here - is that generalized linear model was capable of achieving better R squared results, independently of parameter selection technique.

```
library(ggplot2)
names = c()
r.squareds = c()
for(m in models){
  names = append(names, m$name)
  r.squareds = append(r.squareds, summary(m$model)$adj.r.squared)
}

ggplot(data.frame(names,r.squareds),aes(x=names, y=r.squareds, color= "red")) +
  geom_bar(stat="identity", fill="pink")+ xlab("Model name") +
  ylab("Adjusted R squared") +
  ggtitle("Adjusted R sqared")+geom_text(aes(label=round(r.squareds,4)), vjust=1.6, color="red", size=3
```



Comparison of AIC

A remarkable difference is also observed in terms of Akaike and Bayesian information criteria

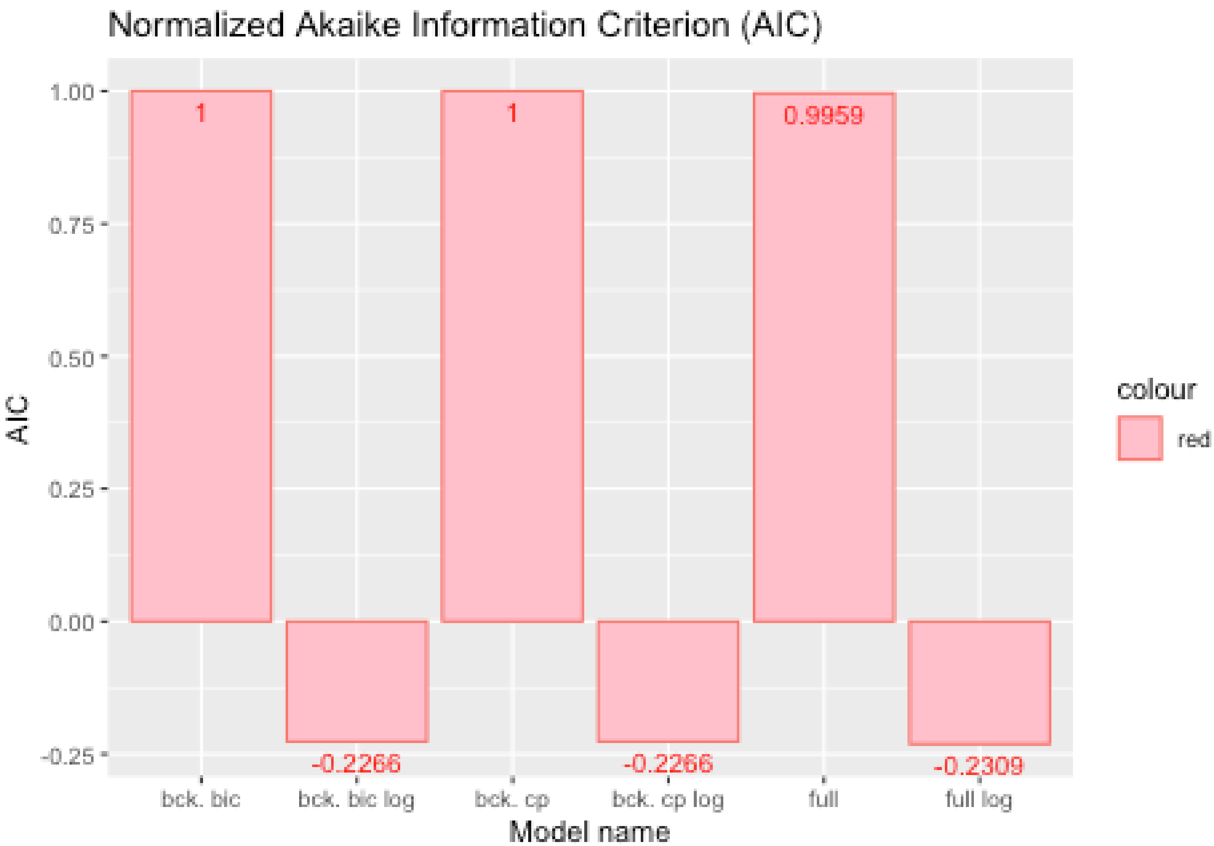
The lower the AIC the better, and a negative AIC indicates a lower degree of information loss than a positive AIC

```
library(olsrr)
```

```
##
## Attaching package: 'olsrr'

## The following object is masked from 'package:datasets':
```

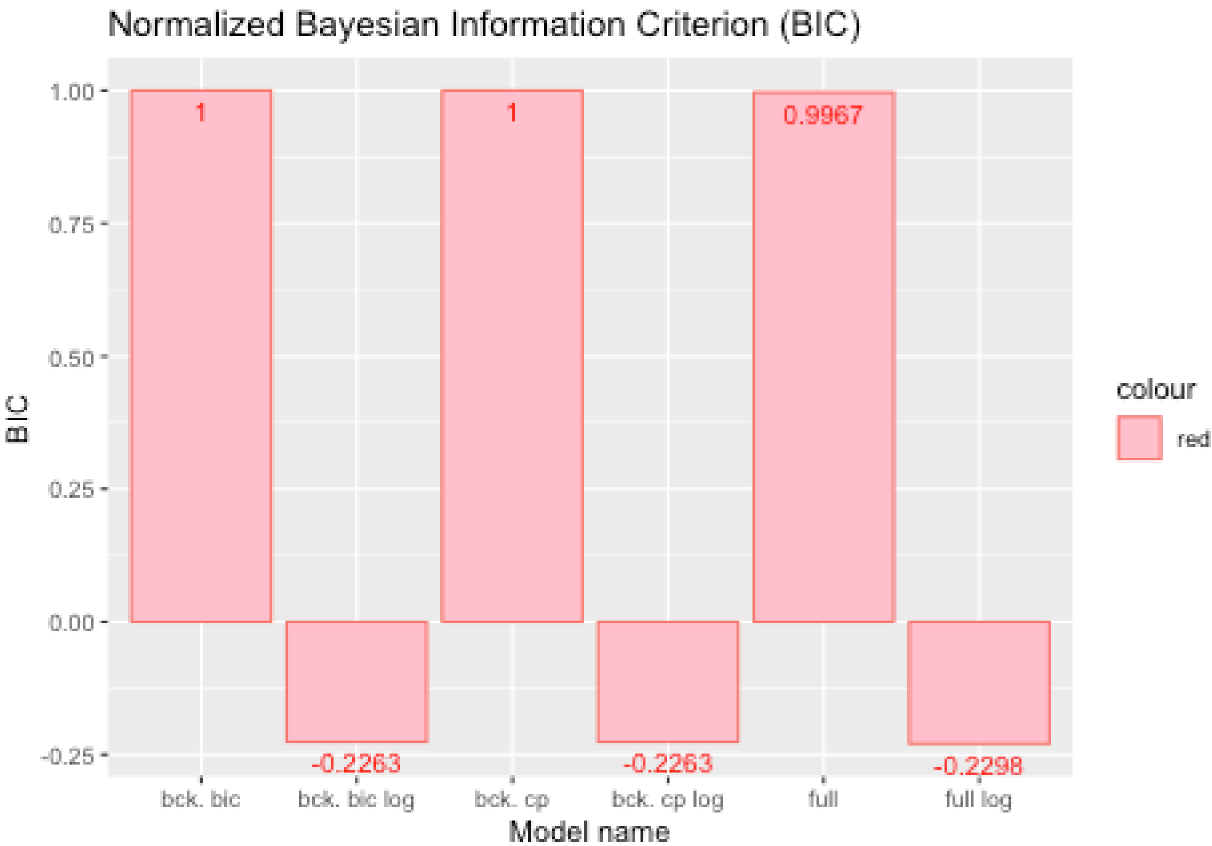
```
##
##     rivers
```

```r
AICs = c()

for(m in models){
  AICs = append(AICs, AIC(m$model))
}

AICs_norm = AICs / max(AICs)

ggplot(data.frame(names,AICs_norm),aes(x=names, y=AICs_norm, color= "red")) +
  geom_bar(stat="identity", fill="pink")+ xlab("Model name") +
  ylab("AIC") +
  ggtitle("Normalized Akaike Information Criterion (AIC)")+geom_text(aes(label=round(AICs_norm,4)), vju
```



Comparison of BIC

```r
BICs = c()

for(m in models){
  BICs = append(BICs, BIC(m$model))
}

BICs_norm = BICs / max(BICs)
ggplot(data.frame(names,BICs_norm),aes(x=names, y=BICs_norm, color= "red")) +
  geom_bar(stat="identity", fill="pink")+ xlab("Model name") +
  ylab("BIC") +
  ggtitle("Normalized Bayesian Information Criterion (BIC)")+geom_text(aes(label=round(BICs_norm,4)), v
```

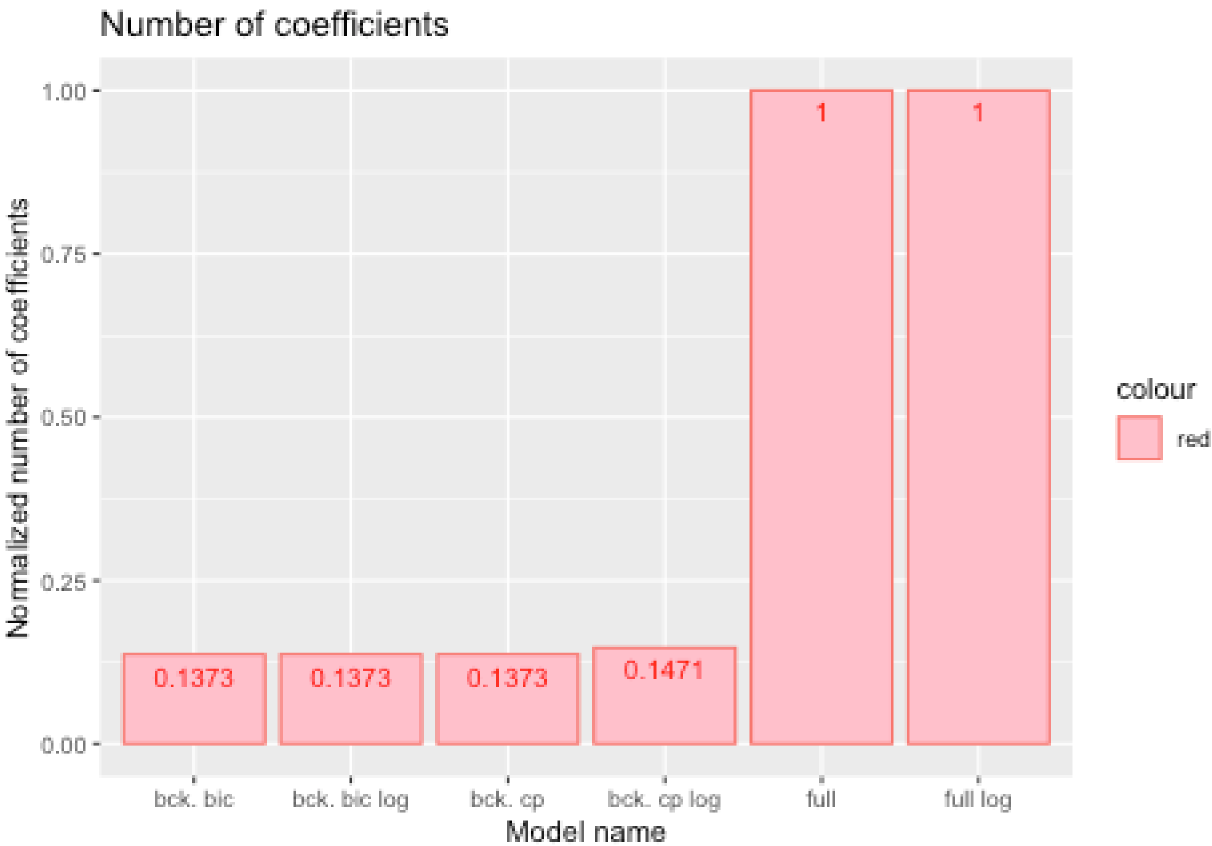


Comparison of parameters number

A powerful result can be made here - having up to 87 percent less parameters, the lightweight models still were able to achieve better R squared metrics.

WE can see GLM with backward selection and GLM log model with backward selection has less parameters.

```
model_sizes = c()
for(m in models){
  model_sizes = append(model_sizes, length(coefficients(m$model)))
}
model_sizes_norm = model_sizes / max(model_sizes)

ggplot(data.frame(names,model_sizes_norm),aes(x=names, y=model_sizes_norm, color= "red")) +
  geom_bar(stat="identity", fill="pink")+ xlab("Model name") +
  ylab("Normalized number of coefficients") +
  ggtitle("Number of coefficients")+ geom_text(aes(label=round(model_sizes_norm,4)), vjust=1.6, color="
```



Comparison on the validation set

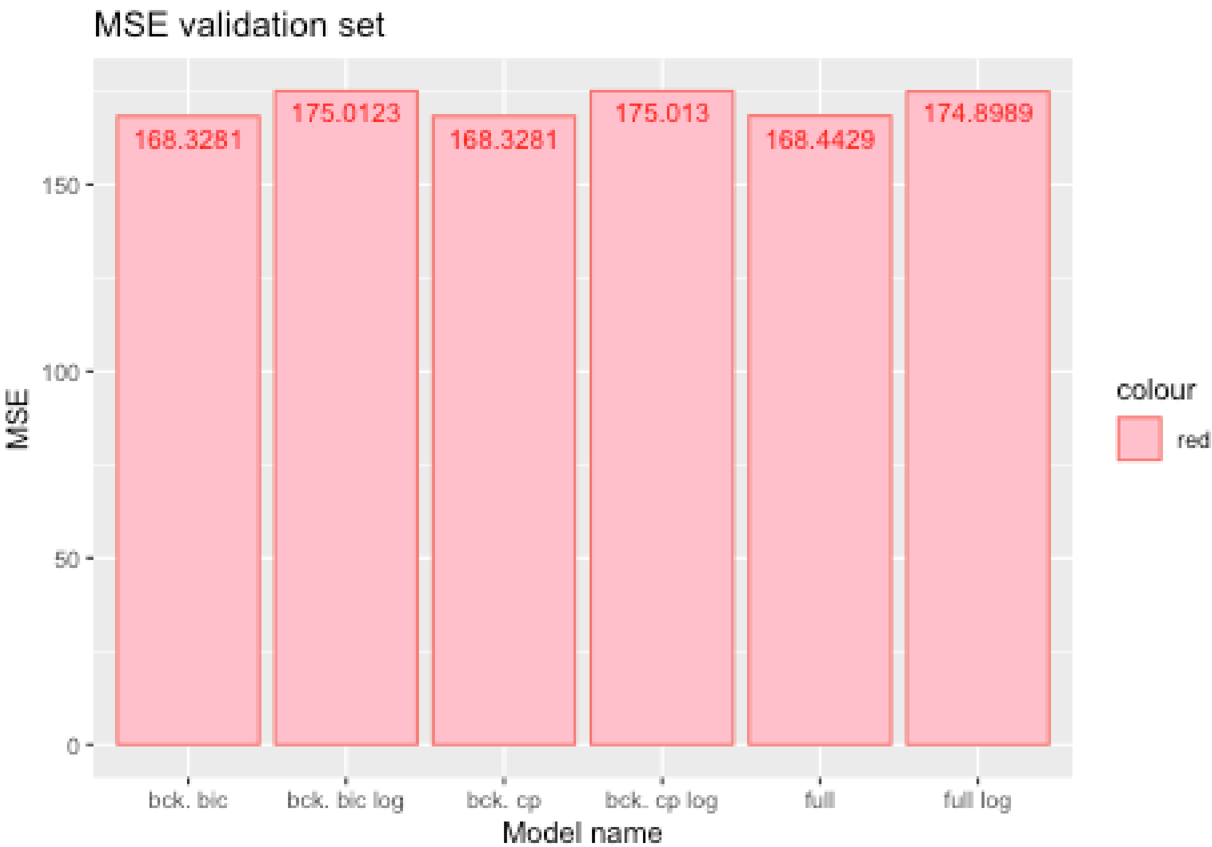
On the validation set, all models achieved comparable results.

```
library(stringr)

validation_r2 = c()

for(m in models){
  model = m$model##
  name = m$name

  preds = predict(
    model,
    data1 = validation[, coefficients(model)]
  )

  if(str_detect(name, 'log')){
    preds = exp(preds)
  }

  r2 = mean((preds - test$price)^2)
  cat('Model:', name, 'MSE validation:', r2, '\n')
  validation_r2 = append(validation_r2, r2)
}
```

```
## Model: full MSE validation: 168.4429
## Model: full log MSE validation: 174.8989
## Model: bck. bic MSE validation: 168.3281
```

```
## Model: bck. cp MSE validation: 168.3281
## Model: bck. cp log MSE validation: 175.013
## Model: bck. bic log MSE validation: 175.0123
```

```
ggplot(data.frame(names,model_sizes_norm),aes(x=names, y = validation_r2, color= "red")) +
  geom_bar(stat="identity", fill="pink")+ xlab("Model name") +
  ylab("MSE") +
  ggtitle("MSE validation set")+ geom_text(aes(label=round(validation_r2,4)), vjust=1.6, color="red", s
```



Our final conclusion regarding the linear models: The best model is a Poisson GLM, with backward coefficients selection. It has 87% less parameters than full model, allowing for around 0.9386 adjusted R squared.