

final-test-analysis.R

Shawn

2021-03-28

```
pollutants <- read.csv("C:/Users/Shawn/Downloads/pollutants.csv")
library("car")

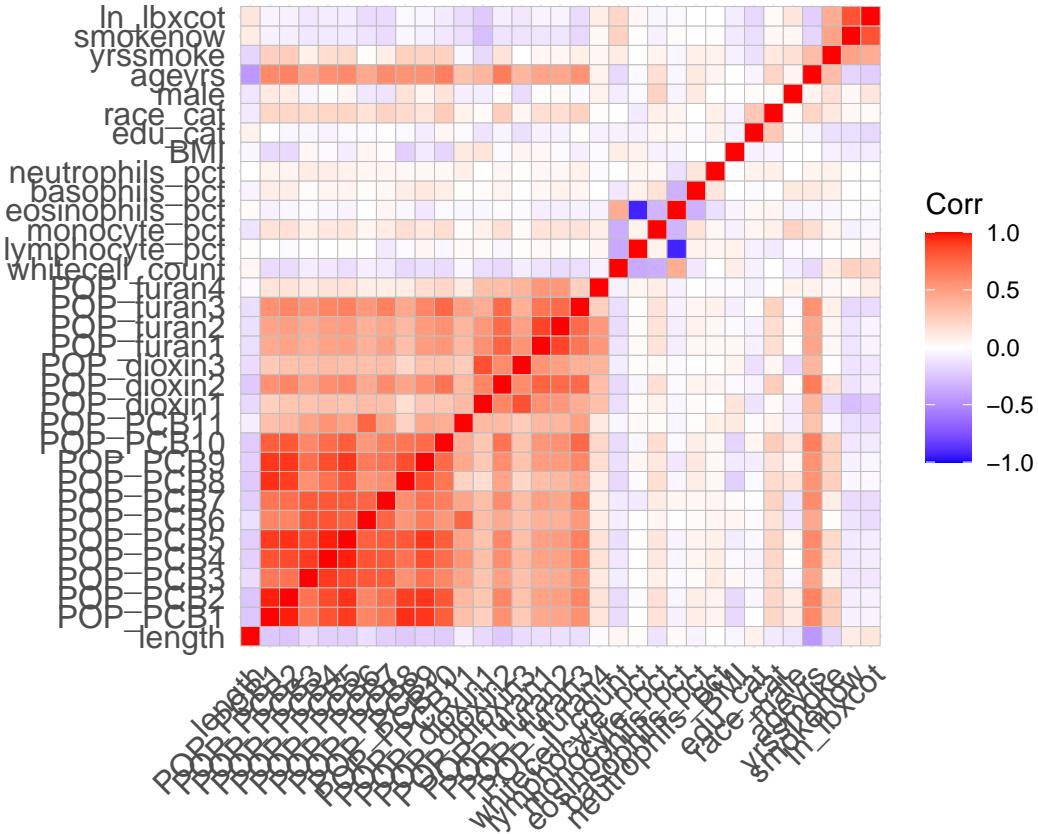
## Loading required package: carData
library(ggplot2)
library(ggcrrplot)
library(caret)

## Loading required package: lattice
library(glmnet)

## Loading required package: Matrix

## Loaded glmnet 4.1-1
library(MASS)
#remove the x
pollutants[ "X" ] = NULL

#calculate correlation matrix before removing the multicolinearity covariates
corr_mat = cor(pollutants)
#graph colored corr matrix
ggcorrplot(corr_mat)
```



```
#set the factor
pollutants$edu_cat = as.factor(pollutants$edu_cat)
pollutants$race_cat = as.factor(pollutants$race_cat)
pollutants$male = as.factor(pollutants$male)
pollutants$smokenow = as.factor(pollutants$smokenow)
#fit model
model = lm(length ~ ., data = pollutants)
#summary
summary(model)
```

```
##
## Call:
## lm(formula = length ~ ., data = pollutants)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.5023 -0.1540 -0.0290  0.1224  1.1904 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -5.516e-02  9.700e+00 -0.006  0.9955    
## POP_PCB1    -1.604e-06  1.075e-06 -1.492  0.1361    
## POP_PCB2     7.240e-07  3.023e-06  0.240  0.8108    
## POP_PCB3     1.189e-06  2.157e-06  0.551  0.5816    
## POP_PCB4    -1.800e-07  1.026e-06 -0.175  0.8608    
## POP_PCB5     1.496e-07  1.070e-06  0.140  0.8889
```

```

## POP_PCB6      2.754e-07  1.059e-06  0.260  0.7949
## POP_PCB7     -5.768e-07  1.207e-06 -0.478  0.6328
## POP_PCB8      1.644e-06  2.447e-06  0.672  0.5021
## POP_PCB9      6.043e-07  2.115e-06  0.286  0.7751
## POP_PCB10     1.181e-03  8.919e-04  1.324  0.1858
## POP_PCB11     3.405e-05  3.079e-04  0.111  0.9120
## POP_dioxin1    2.773e-05  3.056e-04  0.091  0.9277
## POP_dioxin2   -1.732e-04  4.398e-04 -0.394  0.6939
## POP_dioxin3   -1.876e-05  3.027e-05 -0.620  0.5356
## POP_furan1     2.522e-03  3.846e-03  0.656  0.5122
## POP_furan2    -2.915e-04  4.504e-03 -0.065  0.9484
## POP_furan3     4.498e-03  2.762e-03  1.629  0.1038
## POP_furan4    -6.489e-04  9.201e-04 -0.705  0.4808
## whitecell_count -5.233e-03  4.410e-03 -1.186  0.2358
## lymphocyte_pct   1.420e-02  9.698e-02  0.146  0.8836
## monocyte_pct    9.448e-03  9.697e-02  0.097  0.9224
## eosinophils_pct  1.545e-02  9.697e-02  0.159  0.8734
## basophils_pct    1.651e-02  9.706e-02  0.170  0.8650
## neutrophils_pct  2.754e-02  9.816e-02  0.281  0.7791
## BMI            -1.367e-03  1.411e-03 -0.969  0.3329
## edu_cat2       2.439e-02  2.218e-02  1.100  0.2718
## edu_cat3       4.781e-02  2.166e-02  2.207  0.0276 *
## edu_cat4       3.259e-02  2.557e-02  1.275  0.2028
## race_cat2      -2.198e-02  3.267e-02 -0.673  0.5013
## race_cat3      2.464e-02  3.372e-02  0.730  0.4653
## race_cat4      -3.479e-02  2.993e-02 -1.162  0.2455
## male1          -3.947e-02  1.772e-02 -2.227  0.0262 *
## ageyrs          -6.234e-03  7.447e-04 -8.372  2.41e-16 ***
## yrssmoke       -5.325e-04  7.276e-04 -0.732  0.4645
## smokenow1      1.914e-03  3.587e-02  0.053  0.9575
## ln_lbxcot      5.371e-03  3.928e-03  1.367  0.1719
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2221 on 827 degrees of freedom
## Multiple R-squared:  0.2448, Adjusted R-squared:  0.2119
## F-statistic: 7.447 on 36 and 827 DF,  p-value: < 2.2e-16
#show the VIF
vif(model)

```

```

##                      GVIF Df GVIF^(1/(2*Df))
## POP_PCB1        33.044120  1      5.748401
## POP_PCB2        34.281125  1      5.855009
## POP_PCB3        9.351143  1      3.057964
## POP_PCB4       31.742239  1      5.634025
## POP_PCB5       59.896895  1      7.739308
## POP_PCB6       11.386658  1      3.374412
## POP_PCB7        4.870075  1      2.206825
## POP_PCB8       12.982575  1      3.603134
## POP_PCB9       12.441595  1      3.527264
## POP_PCB10      6.020678  1      2.453707
## POP_PCB11      4.725769  1      2.173883
## POP_dioxin1     5.276251  1      2.297009
## POP_dioxin2     5.413132  1      2.326614

```

```

## POP_dioxin3      4.398509  1     2.097262
## POP_furan1      6.154213  1     2.480769
## POP_furan2      6.195336  1     2.489043
## POP_furan3      4.464346  1     2.112900
## POP_furan4      1.821809  1     1.349744
## whitecell_count 1.548380  1     1.244339
## lymphocyte_pct   12250.336528 1    110.681238
## monocyte_pct     726.843372  1    26.960033
## eosinophils_pct 15071.561945 1    122.766290
## basophils_pct   867.412798  1    29.451873
## neutrophils_pct 37.984114  1    6.163125
## BMI             1.263662  1    1.124127
## edu_cat         1.543109  3    1.074978
## race_cat        2.052848  3    1.127352
## male            1.350324  1    1.162034
## ageyrs          3.238631  1    1.799620
## yrssmoke        2.204139  1    1.484634
## smokenow        4.006708  1    2.001676
## ln_lbxcot       3.963407  1    1.990831

#get set a dataset with no categorical covariates
no_cat = pollutants
no_cat$edu_cat = NULL
no_cat$race_cat = NULL
no_cat$male = NULL
no_cat$smokenow = NULL
#summary of the dataset
summary(no_cat)

```

```

##      length      POP_PCB1      POP_PCB2      POP_PCB3
##  Min.   :0.5266  Min.   : 2000  Min.   : 2000  Min.   : 2000
##  1st Qu.:0.8754  1st Qu.: 9975  1st Qu.: 4800  1st Qu.: 3700
##  Median :1.0286  Median : 27600  Median : 11500  Median : 6200
##  Mean   :1.0543  Mean   : 38082  Mean   : 15637  Mean   : 10158
##  3rd Qu.:1.2095  3rd Qu.: 53325  3rd Qu.: 21825  3rd Qu.: 12000
##  Max.   :2.3512  Max.   :572000  Max.   :165000  Max.   :123000
##      POP_PCB4      POP_PCB5      POP_PCB6      POP_PCB7
##  Min.   : 2100  Min.   : 2100  Min.   : 2000  Min.   : 1100
##  1st Qu.:11475  1st Qu.: 15600  1st Qu.: 4400  1st Qu.: 4000
##  Median :25550  Median : 36300  Median : 9400  Median : 7450
##  Mean   :38456  Mean   : 52650  Mean   : 16820  Mean   : 12682
##  3rd Qu.:50650  3rd Qu.: 68625  3rd Qu.: 19500 3rd Qu.: 15625
##  Max.   :487000  Max.   :708000  Max.   :319000  Max.   :144000
##      POP_PCB8      POP_PCB9      POP_PCB10     POP_PCB11
##  Min.   : 1100  Min.   : 1100  Min.   : 1.70  Min.   : 1.30
##  1st Qu.: 3800  1st Qu.: 3900  1st Qu.: 9.10  1st Qu.: 14.80
##  Median : 6950  Median : 8050  Median : 18.35  Median : 24.50
##  Mean   :10530  Mean   : 12220  Mean   : 24.49  Mean   : 38.15
##  3rd Qu.:14425  3rd Qu.: 16025 3rd Qu.: 34.90  3rd Qu.: 42.95
##  Max.   :187000  Max.   :144000  Max.   :172.00  Max.   :845.00
##      POP_dioxin1    POP_dioxin2    POP_dioxin3    POP_furan1
##  Min.   : 1.90  Min.   : 1.40  Min.   : 36.8  Min.   : 1.000
##  1st Qu.: 23.90 1st Qu.: 21.27 1st Qu.: 197.0 1st Qu.: 3.200
##  Median : 41.35  Median : 37.80  Median : 342.5  Median : 5.200
##  Mean   : 57.65  Mean   : 47.81  Mean   : 494.4  Mean   : 6.371

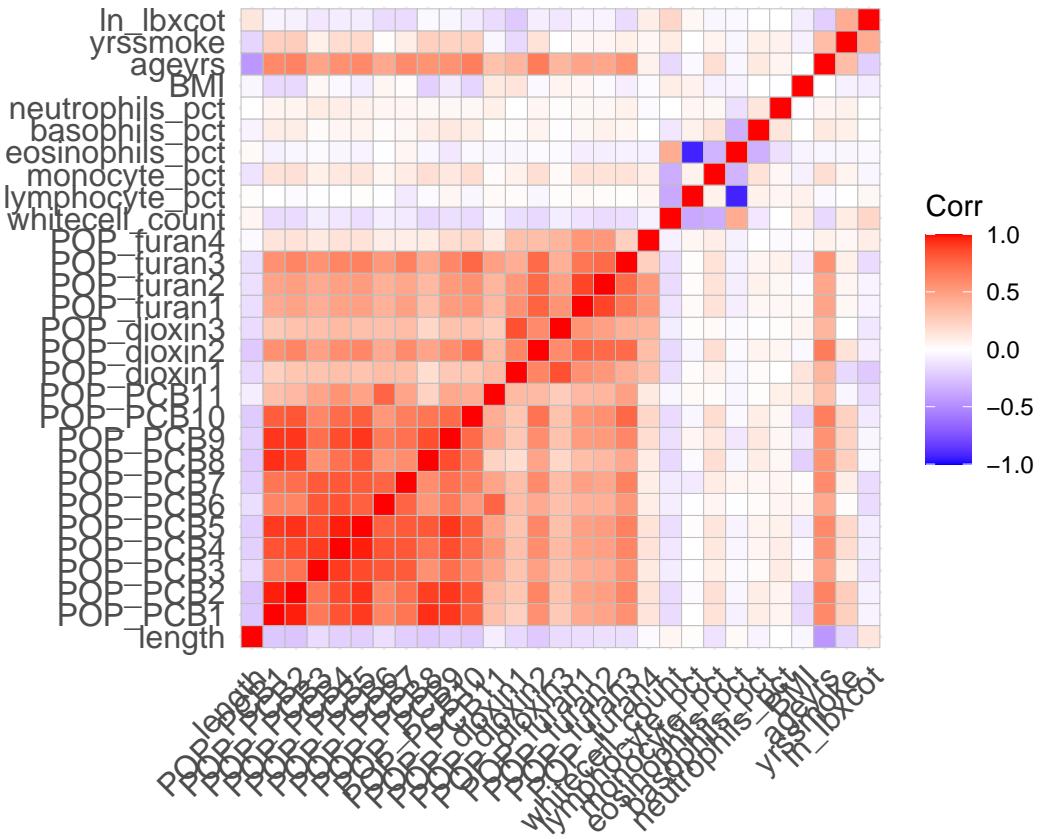
```

```

## 3rd Qu.: 71.62   3rd Qu.: 62.42   3rd Qu.: 603.0   3rd Qu.: 7.700
## Max.    :760.00   Max.    :281.00   Max.    :8190.0   Max.    :44.400
## POP_furan2      POP_furan3      POP_furan4      whitecell_count
## Min.    : 0.800   Min.    : 0.700   Min.    : 0.90   Min.    : 2.300
## 1st Qu.: 2.600   1st Qu.: 2.200   1st Qu.: 6.40   1st Qu.: 5.600
## Median  : 4.200   Median  : 5.050   Median  : 9.65   Median  : 6.900
## Mean    : 5.390   Mean    : 6.669   Mean    :11.54   Mean    : 7.191
## 3rd Qu.: 6.825   3rd Qu.: 9.300   3rd Qu.:14.00   3rd Qu.: 8.300
## Max.    :33.500   Max.    :38.300   Max.    :234.00   Max.    :20.100
## lymphocyte_pct   monocyte_pct   eosinophils_pct basophils_pct
## Min.    : 5.80   Min.    :1.600   Min.    :21.60   Min.    : 0.000
## 1st Qu.:24.00   1st Qu.: 6.600   1st Qu.:52.35   1st Qu.: 1.500
## Median  :28.95   Median  : 7.700   Median  :59.30   Median  : 2.300
## Mean    :29.92   Mean    : 7.936   Mean    :58.62   Mean    : 2.903
## 3rd Qu.:35.42   3rd Qu.: 9.100   3rd Qu.:65.22   3rd Qu.: 3.700
## Max.    :73.40   Max.    :23.800   Max.    :88.10   Max.    :28.200
## neutrophils_pct   BMI       ageyrs      yrssmoke
## Min.    :0.0000   Min.    :16.16   Min.    :20.00   Min.    : 0.0
## 1st Qu.:0.4000   1st Qu.:23.88   1st Qu.:34.00   1st Qu.: 0.0
## Median  :0.6000   Median :27.38   Median :46.00   Median : 0.0
## Mean    :0.6669   Mean    :28.09   Mean    :48.36   Mean    :10.6
## 3rd Qu.:0.8000   3rd Qu.:31.17   3rd Qu.:63.00   3rd Qu.:20.0
## Max.    :5.5000   Max.    :62.99   Max.    :85.00   Max.    :69.0
## ln_lbxcot
## Min.    :-4.5099
## 1st Qu.:-4.0745
## Median :-2.7334
## Mean   :-0.9804
## 3rd Qu.: 2.8000
## Max.   : 6.5848

#calculate correlation matrix
corr_matrix = cor(no_cat)
#graph colored corr matrix
ggcorrplot(corr_matrix)

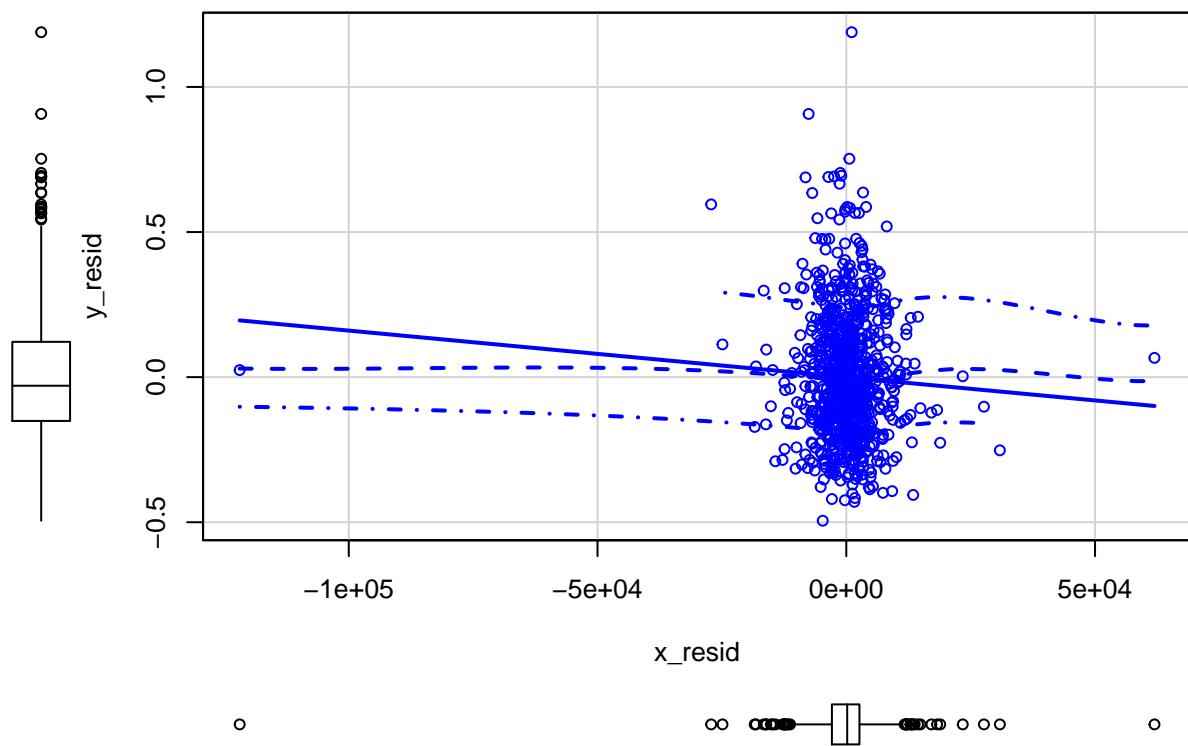
```

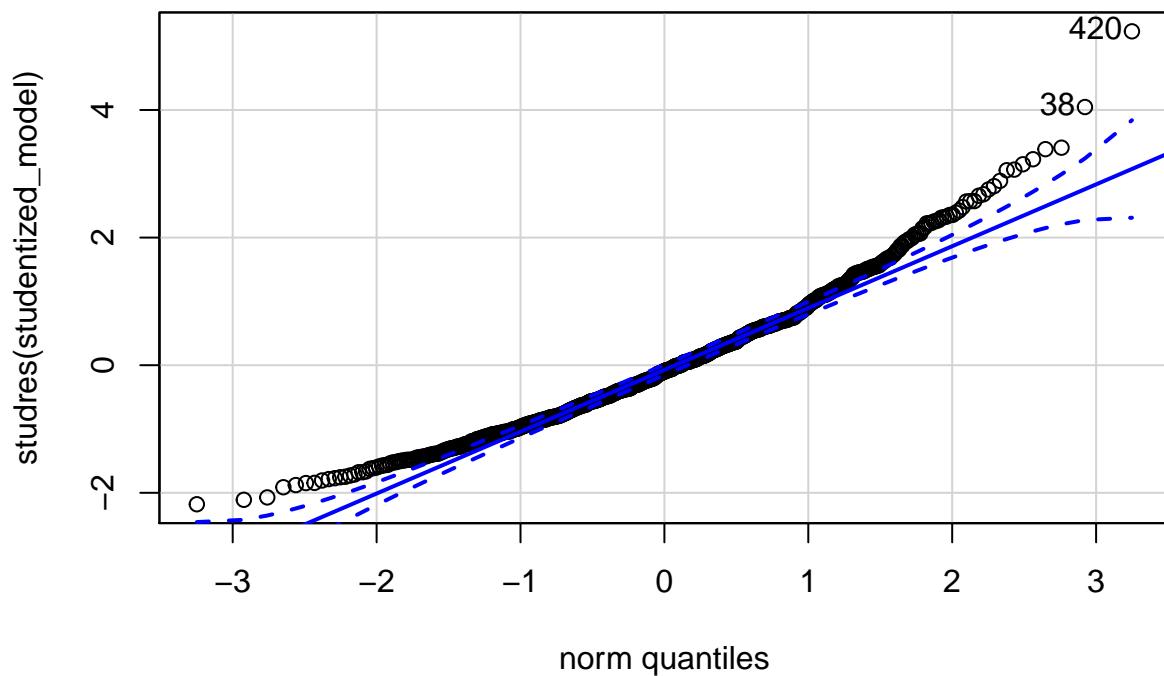


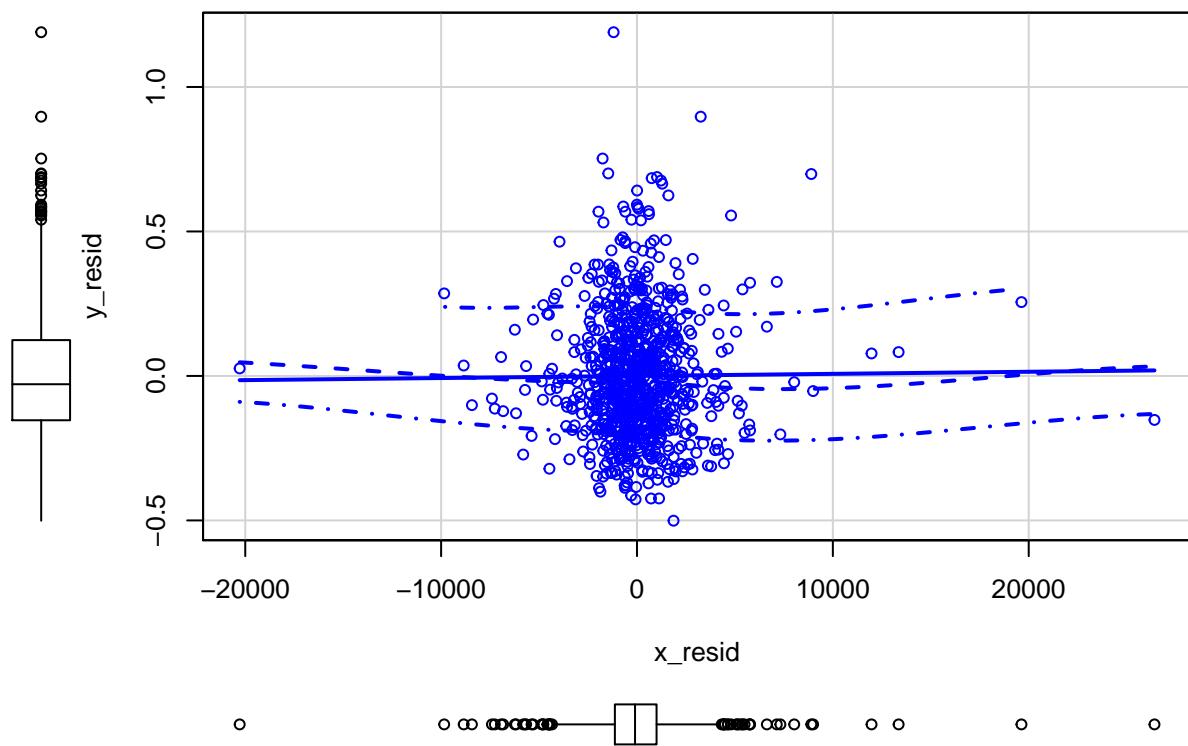
```

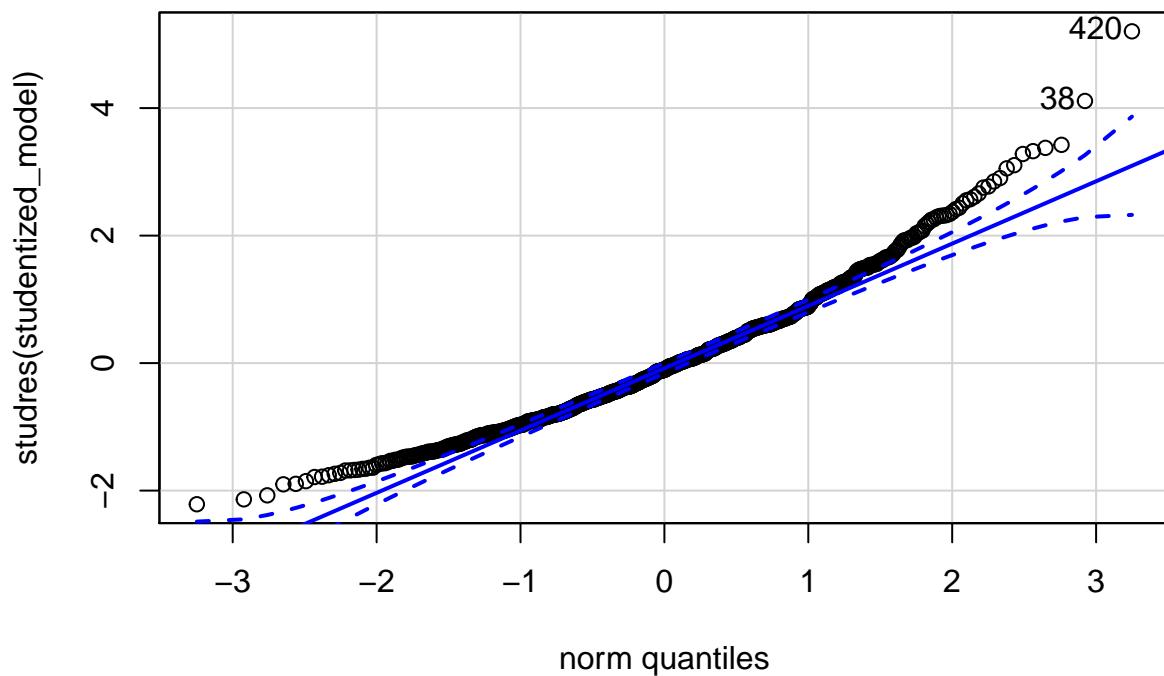
#find the linearity (residual y against residual x)
covariates = names(no_cat)
#remove response name
covariates = covariates[-1]
for (name in covariates){
  y_model = lm(paste("length", "~", ".", "-", name), data = pollutants)
  x_model = lm(paste(name, "~", ".", "- length"), data = pollutants)
  y_resid = resid(y_model)
  x_resid = resid(x_model)
  #residual QQ studentized - show the qqPlot of all covariates that is not categorical
  studentized_model = lm(paste("length", "~", name), data = pollutants)
  #linearity
  scatterplot(x_resid, y_resid)
  qqPlot(studres(studentized_model))
}

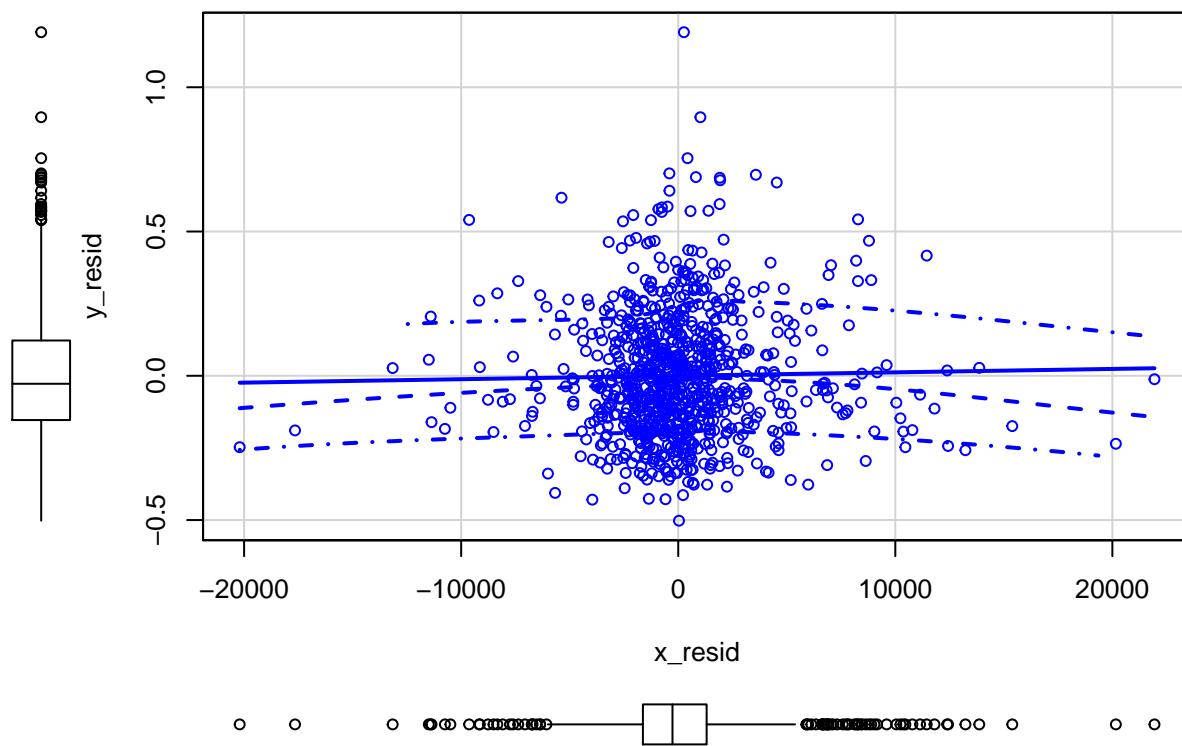
```

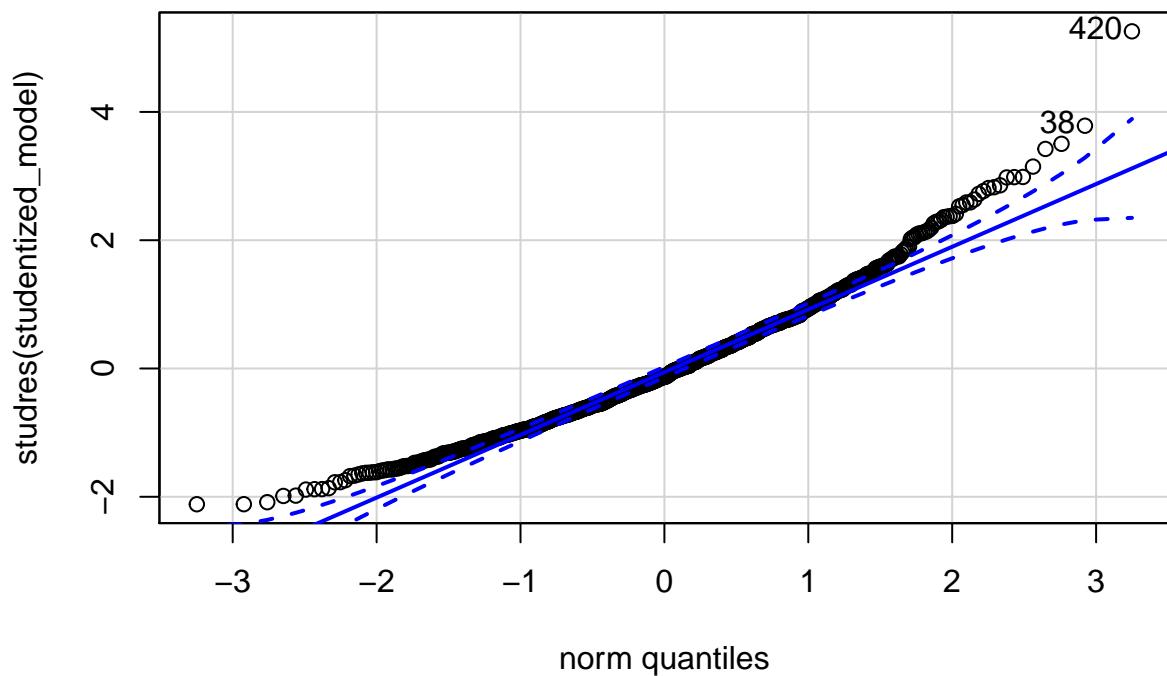


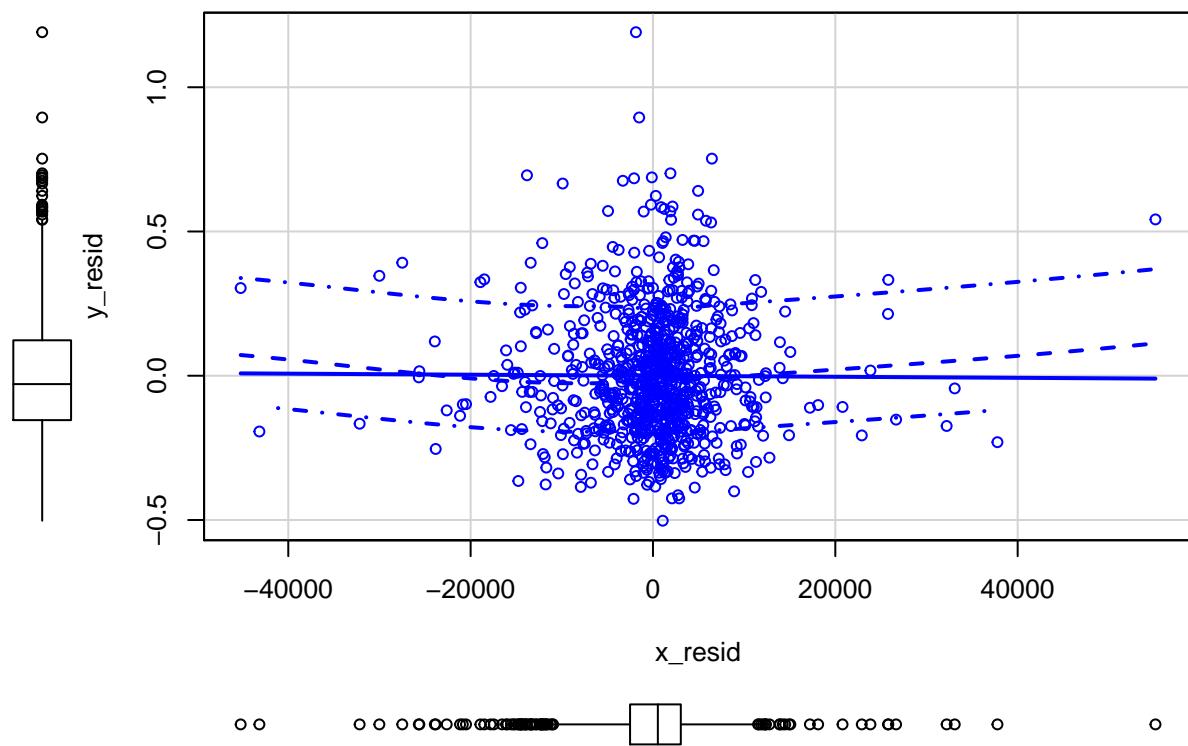


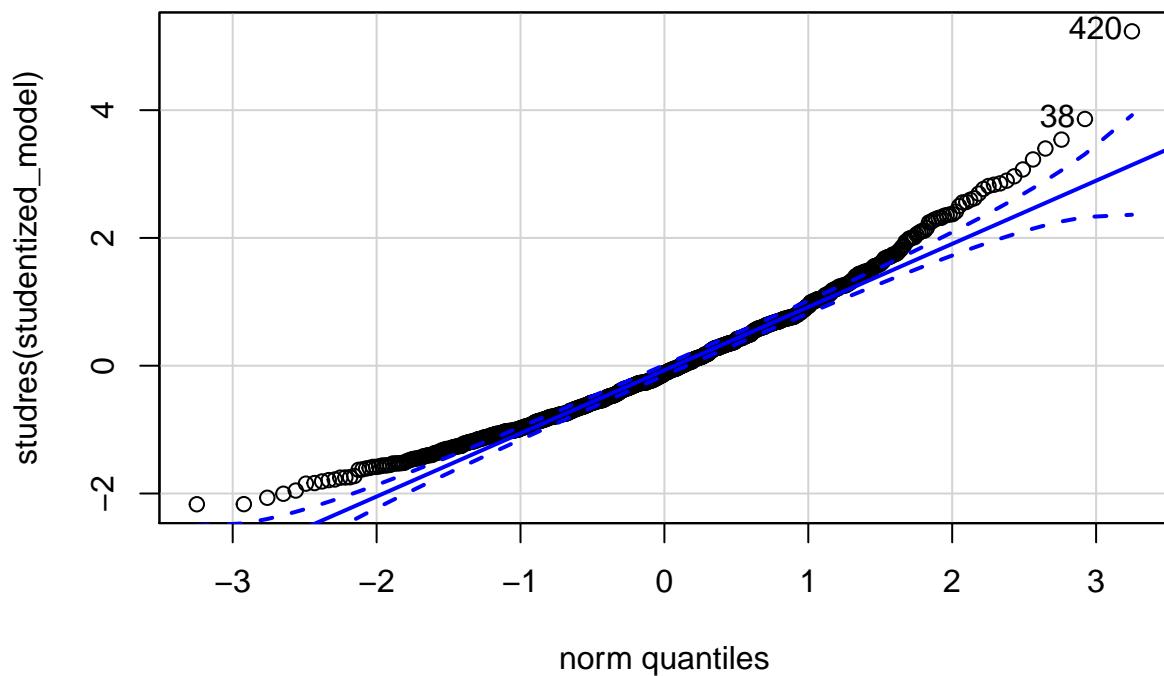


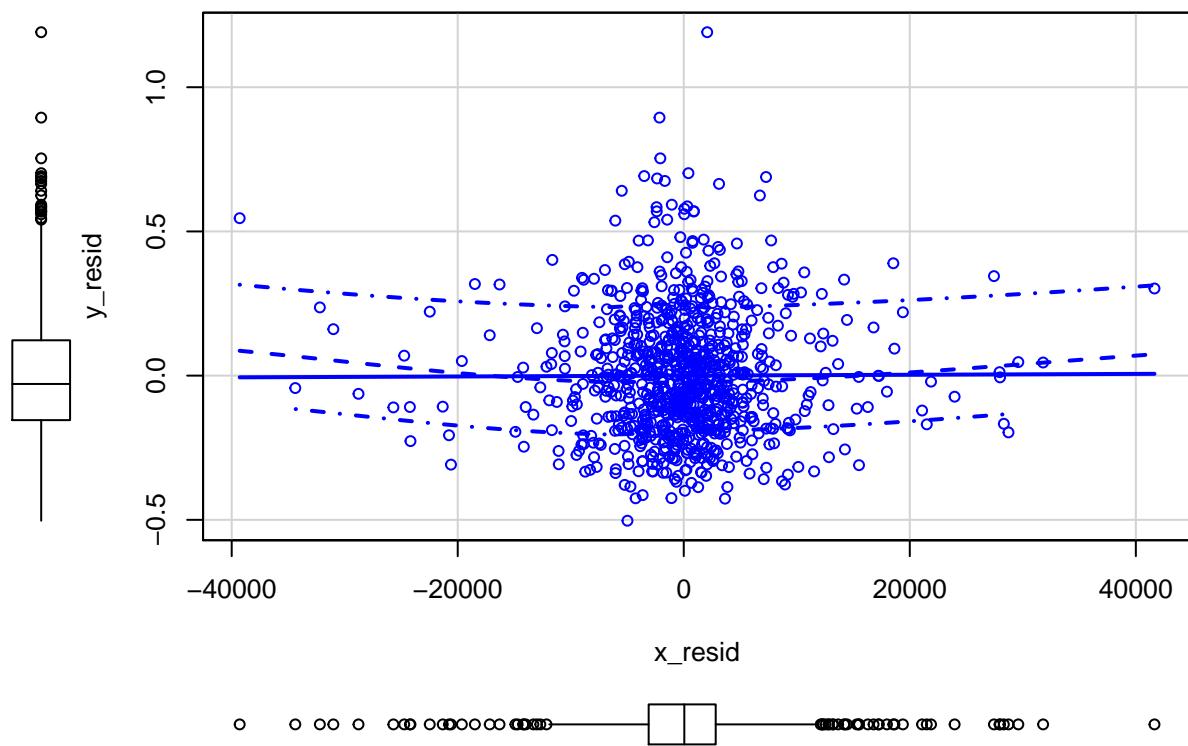


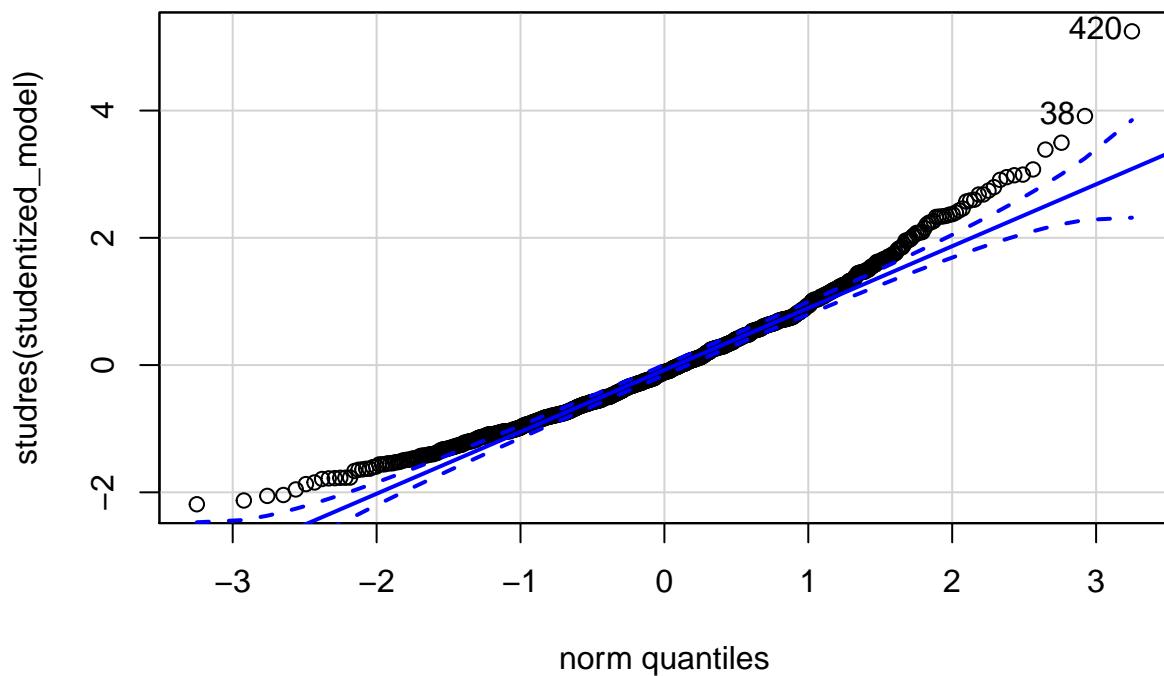


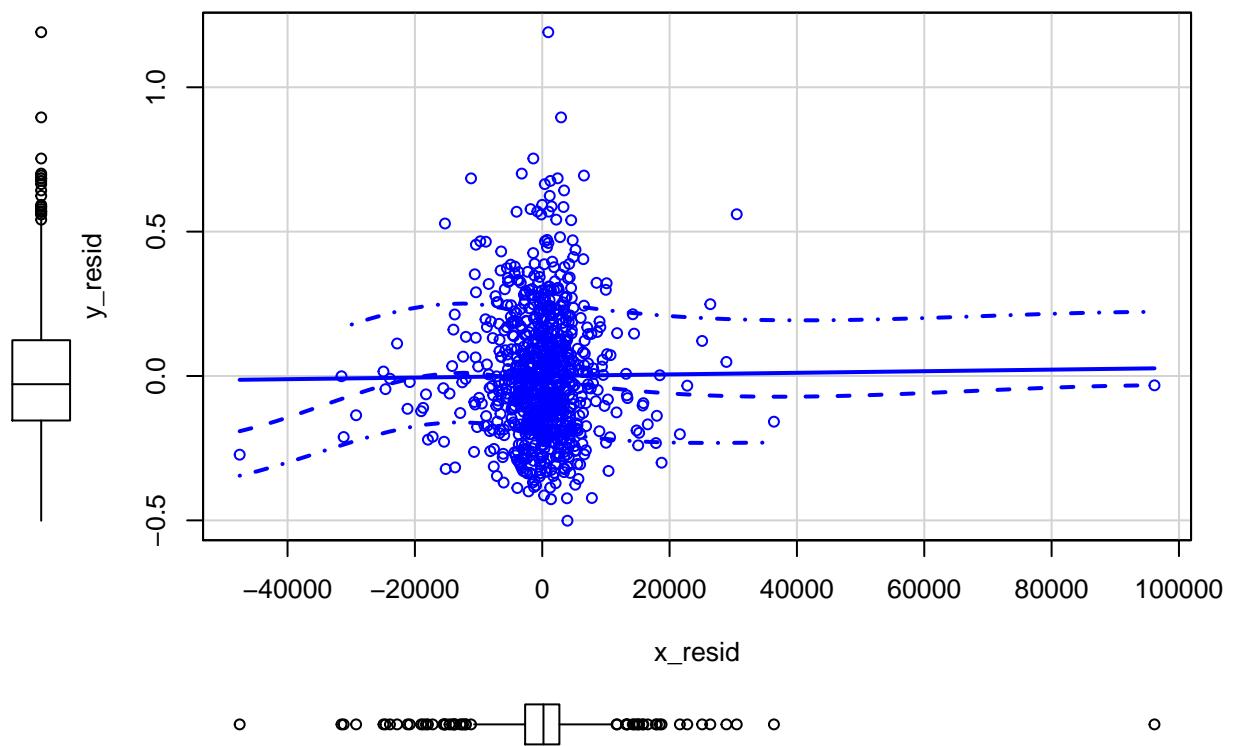


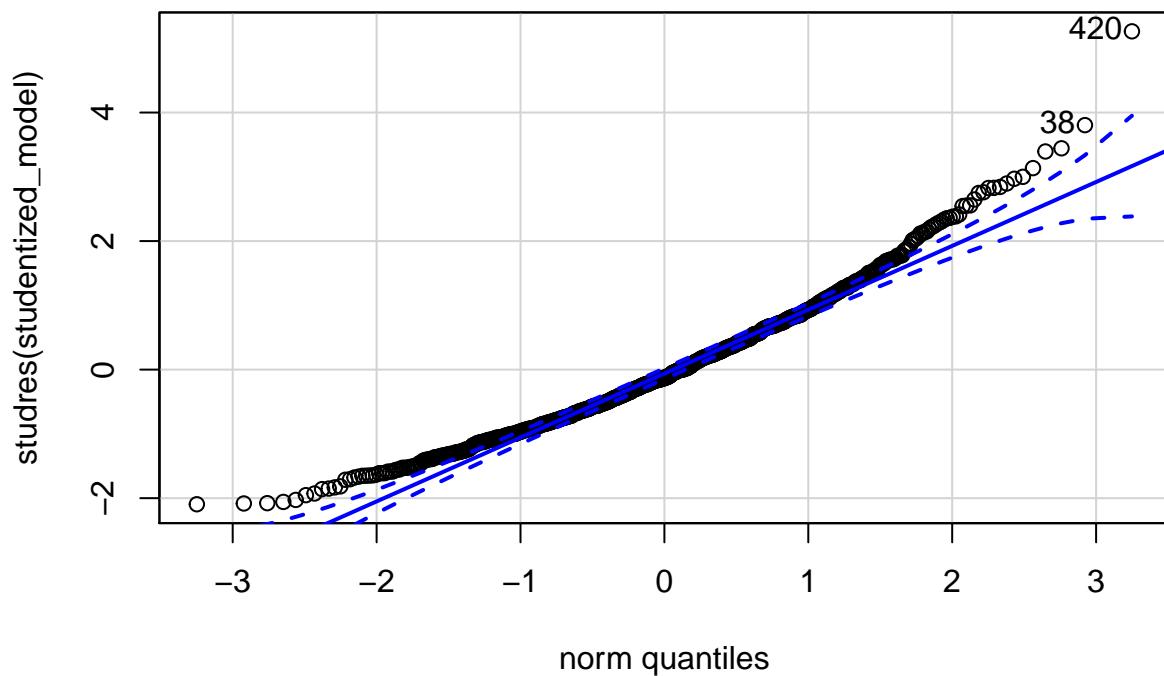


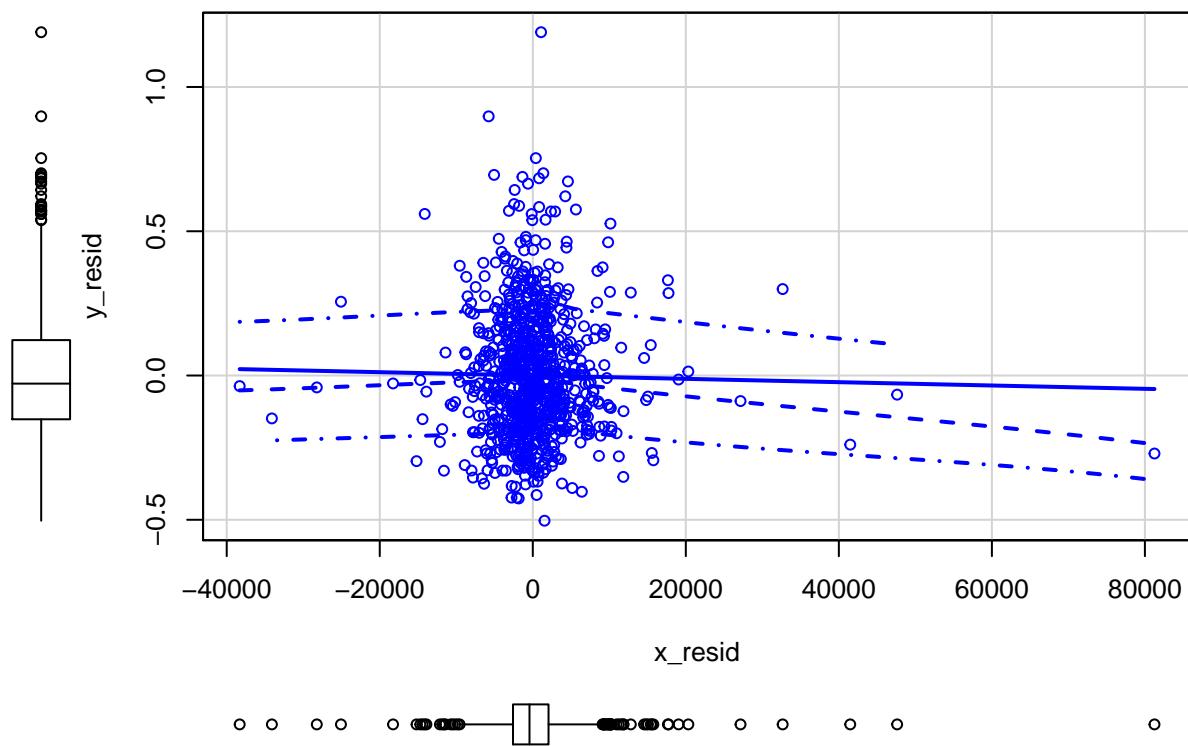


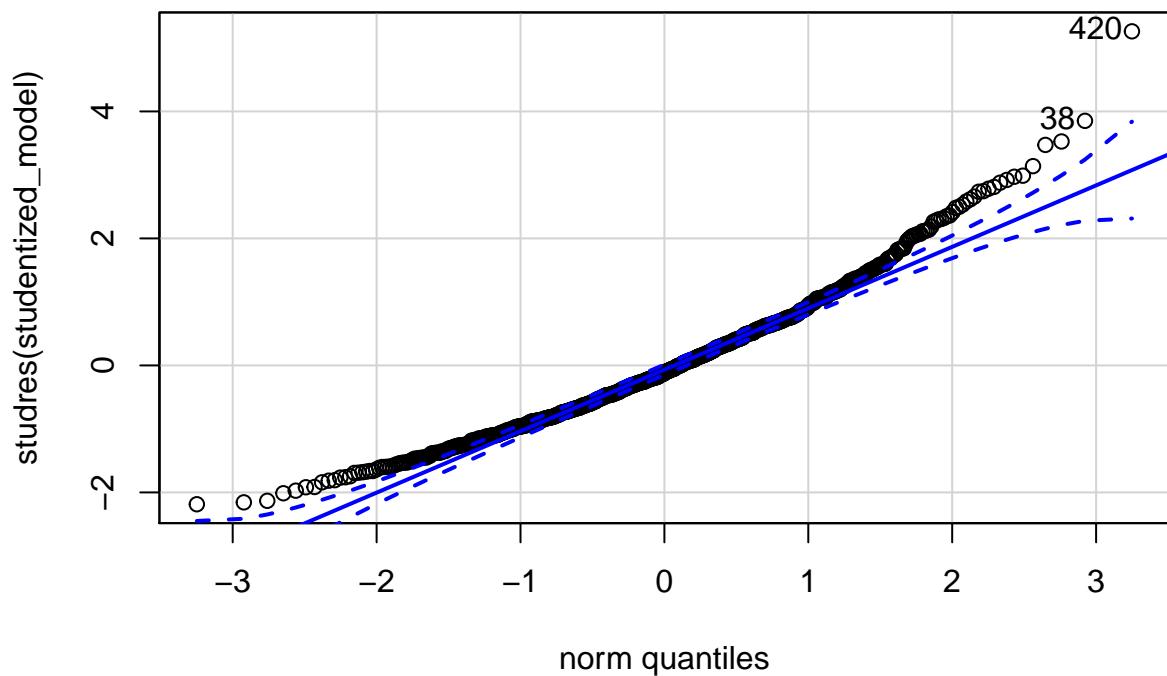


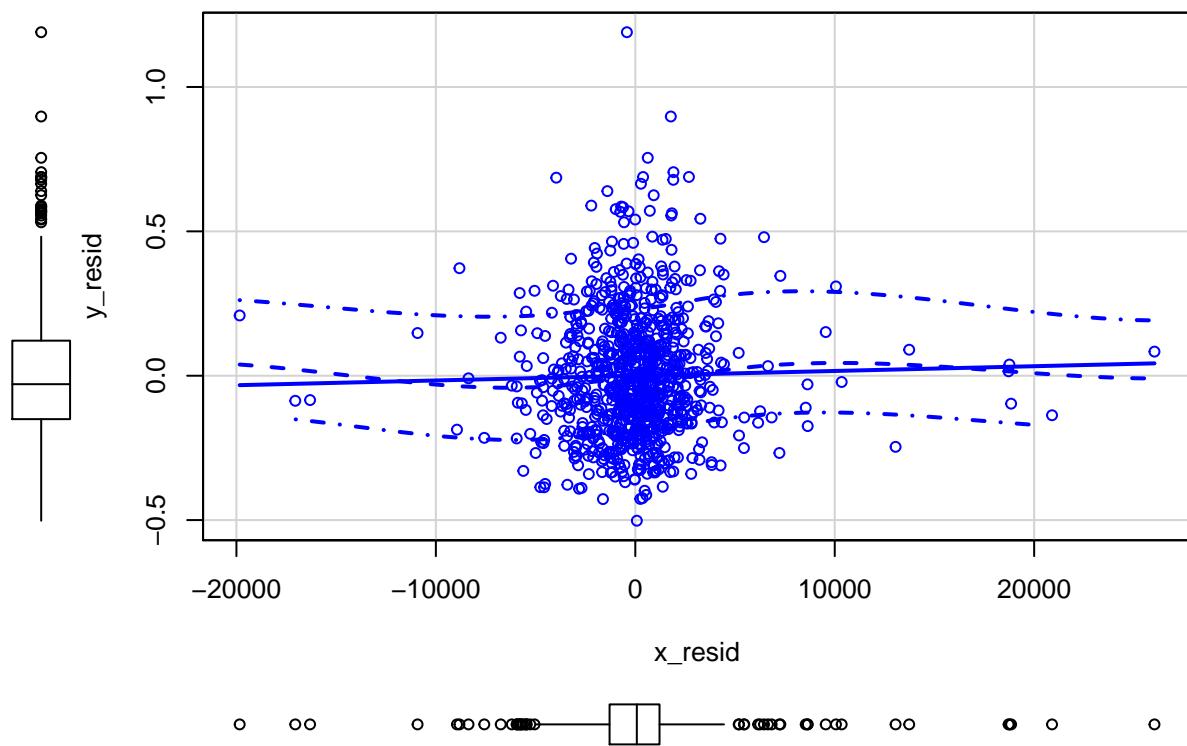


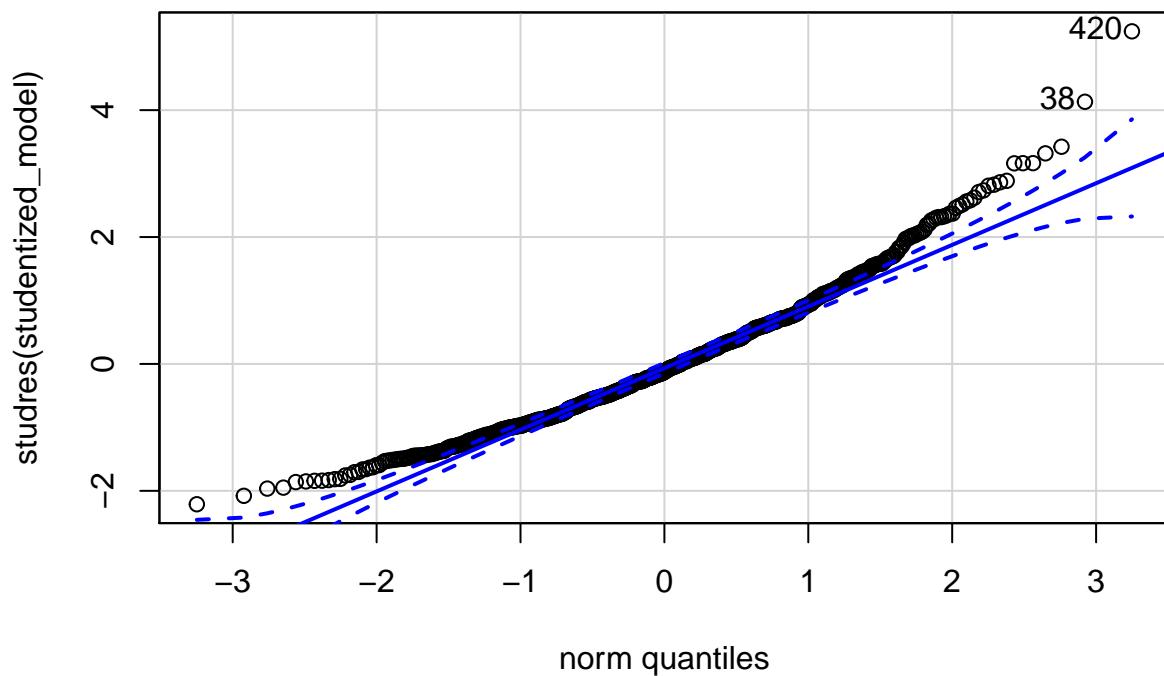


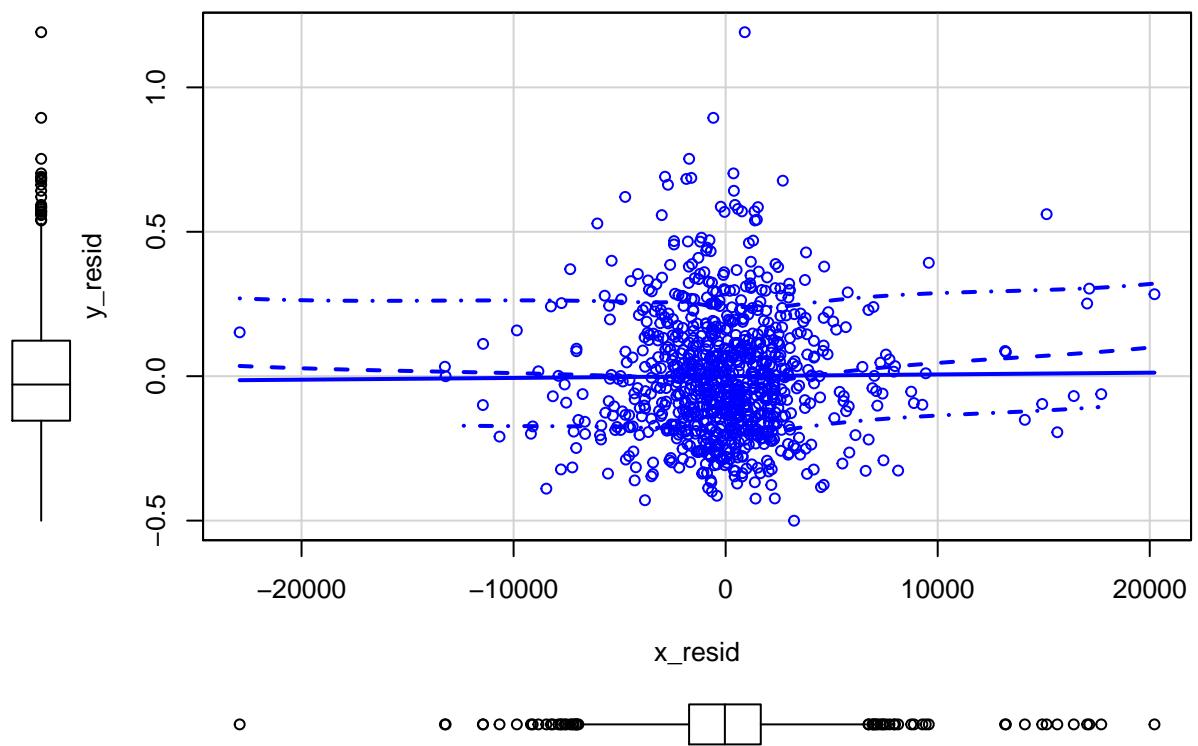


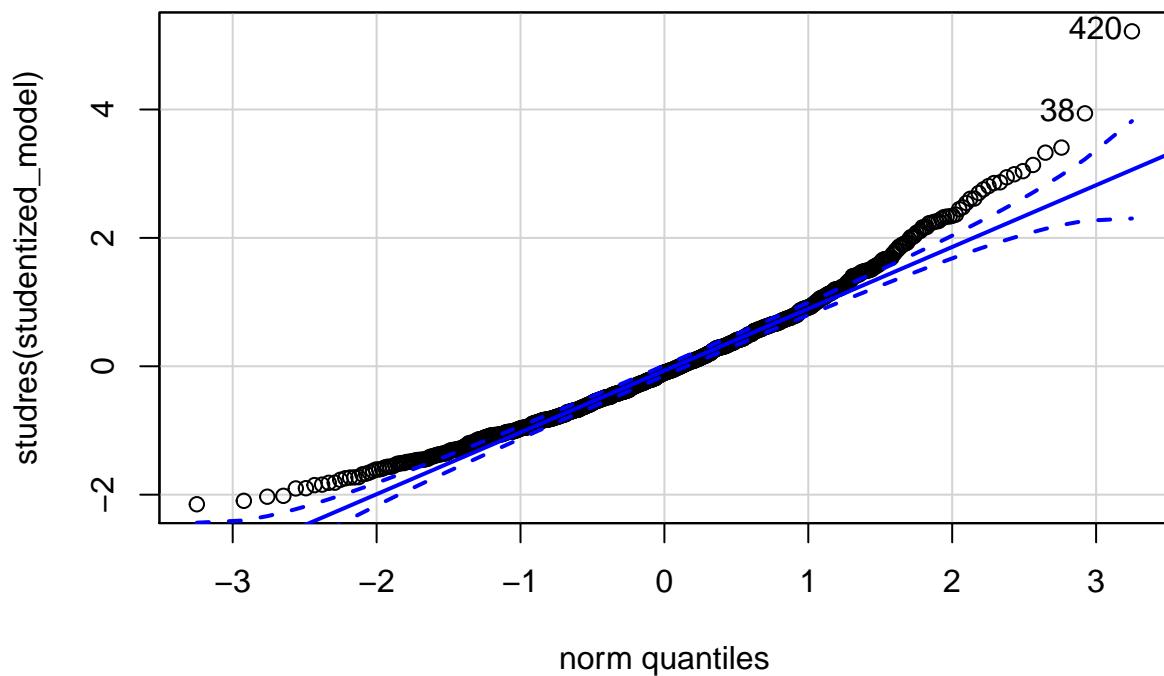


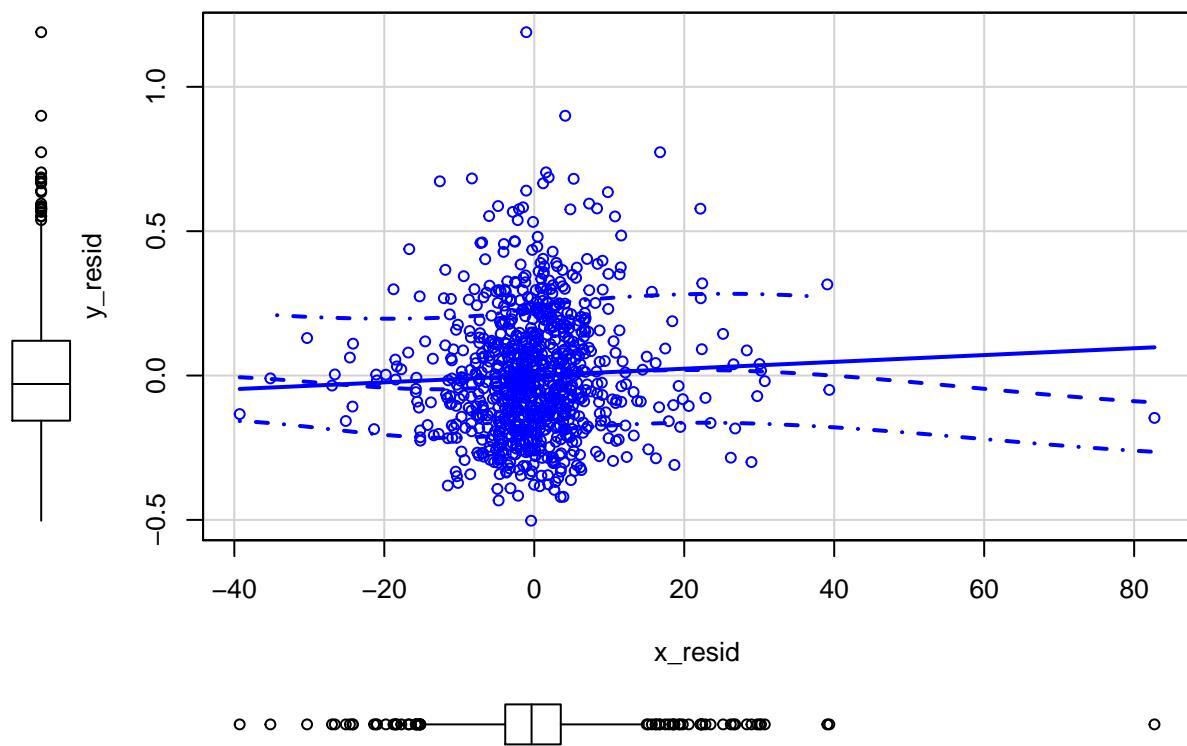


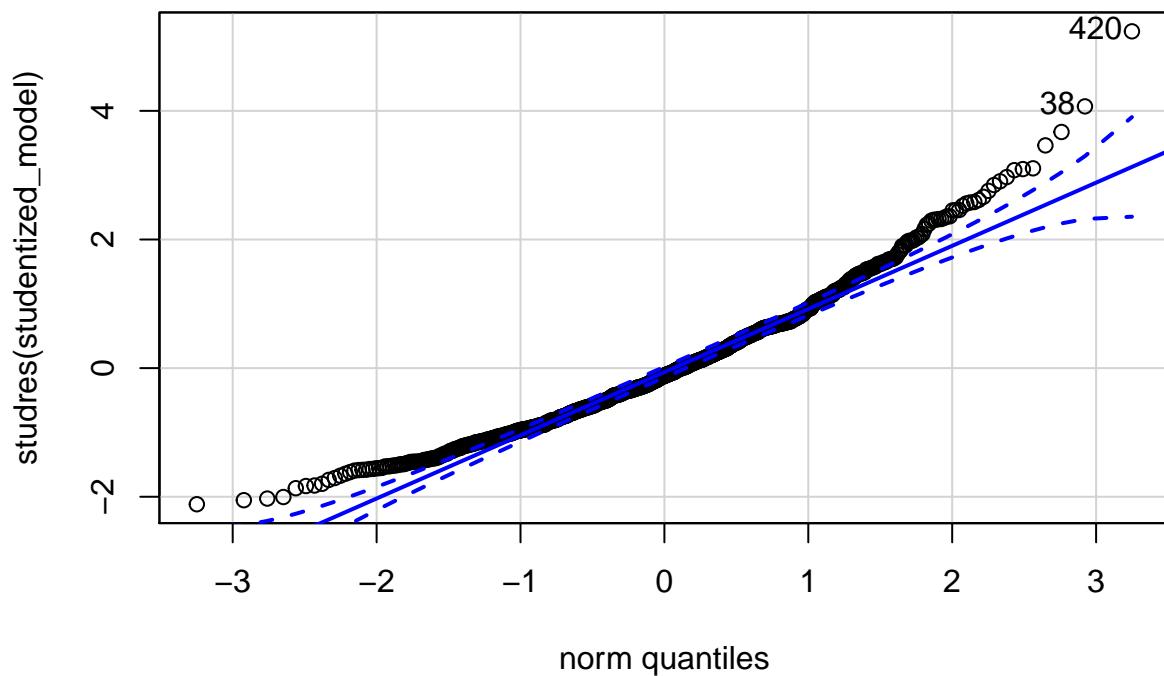


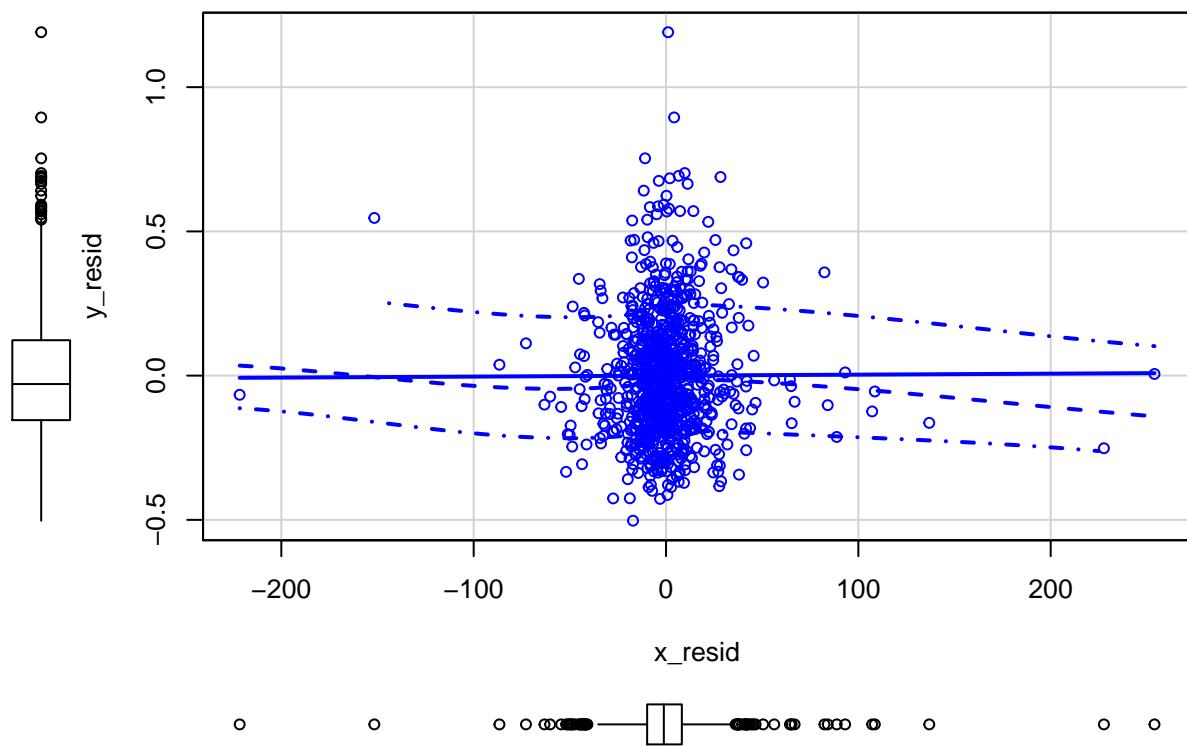


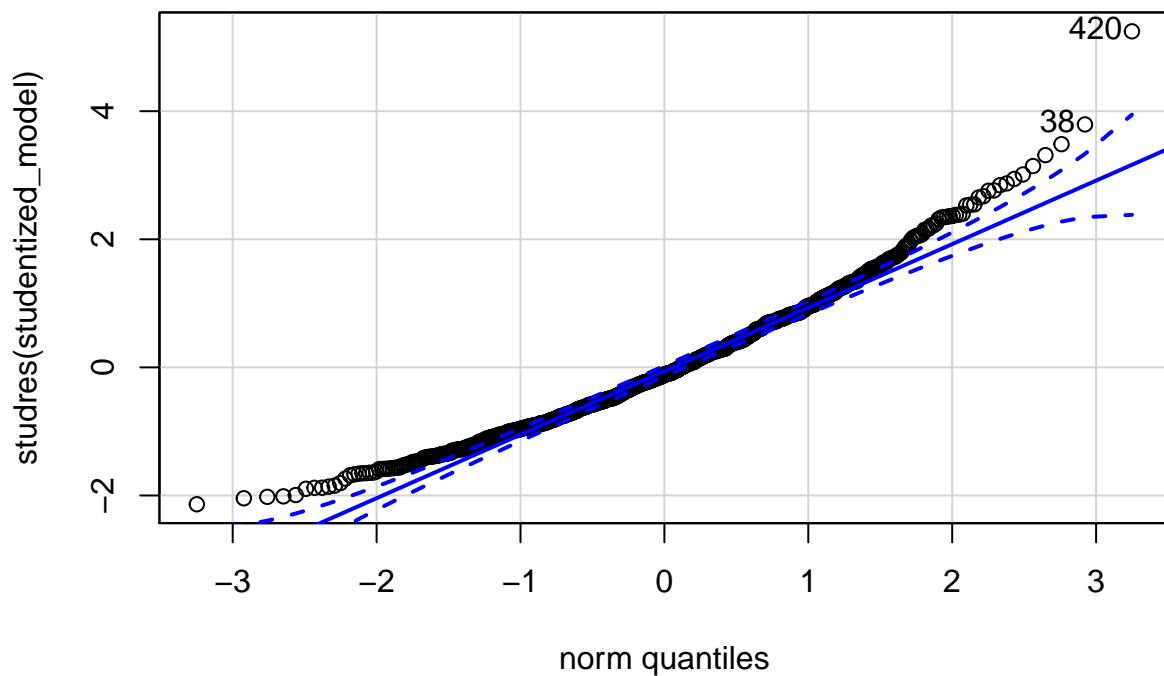


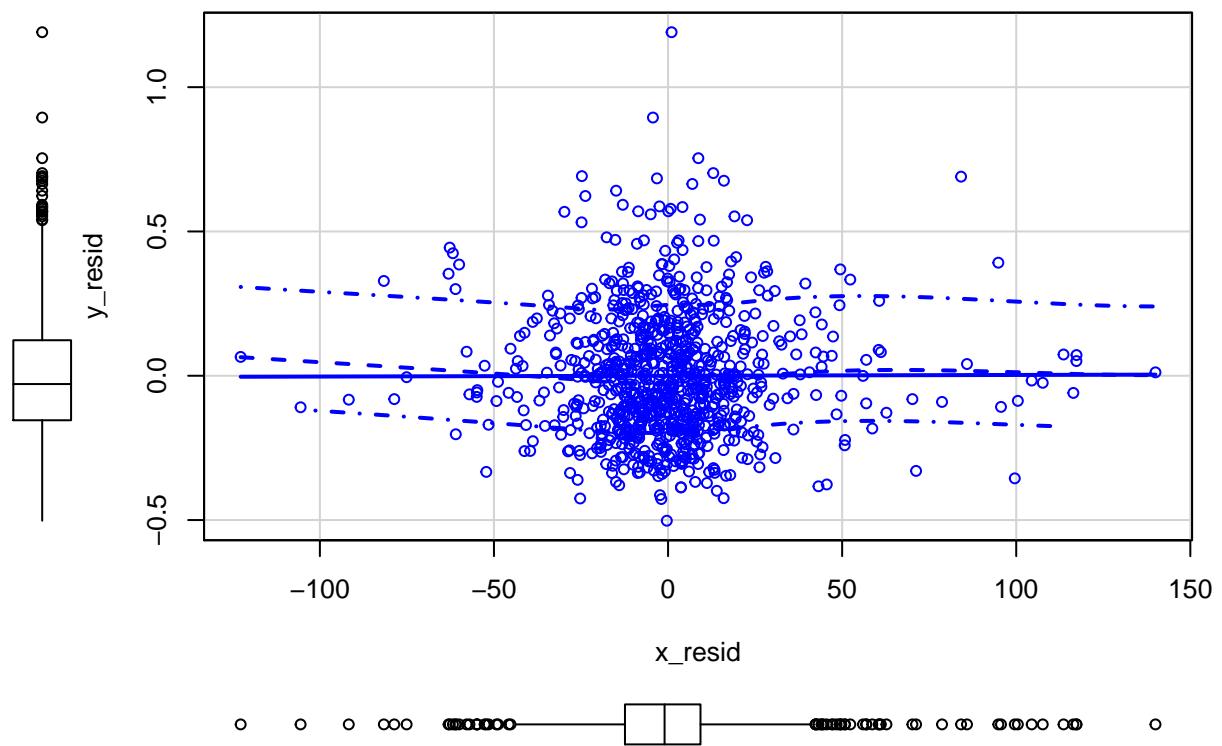


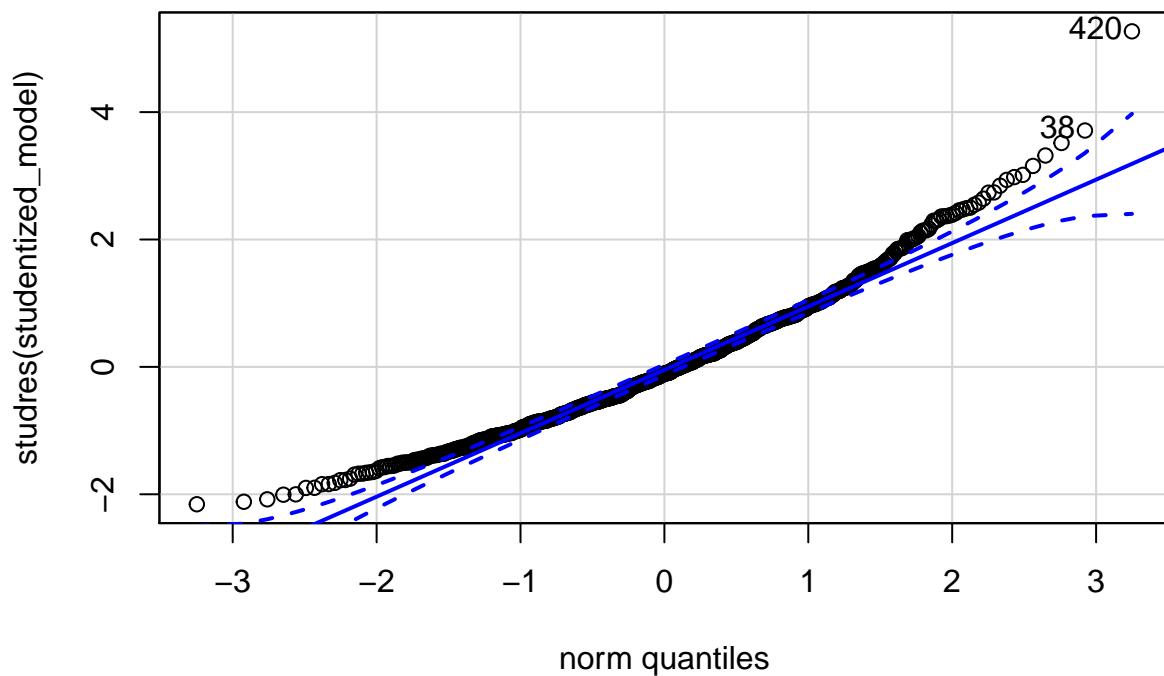


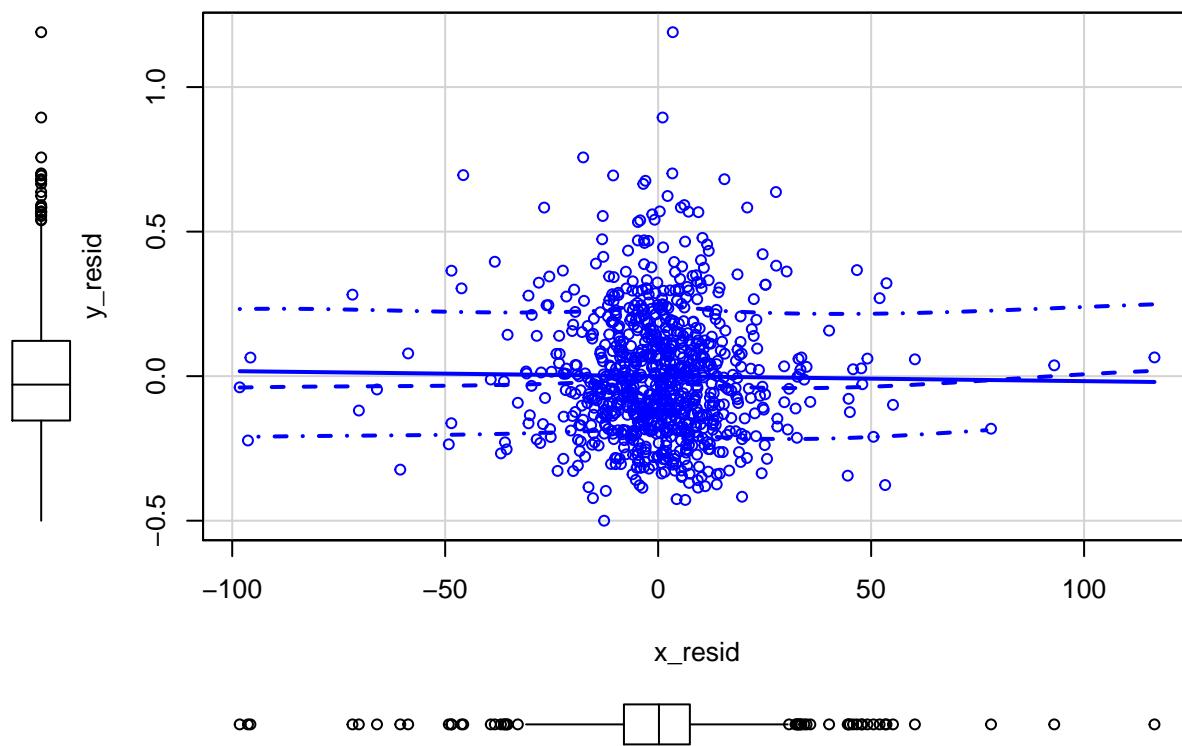


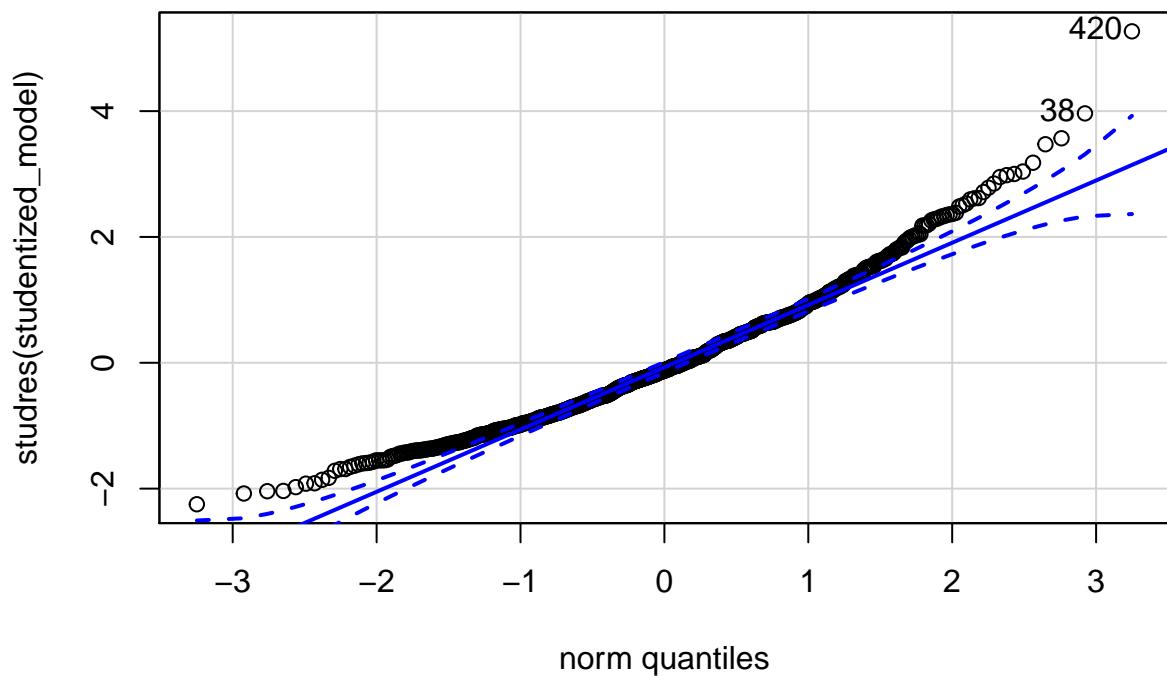


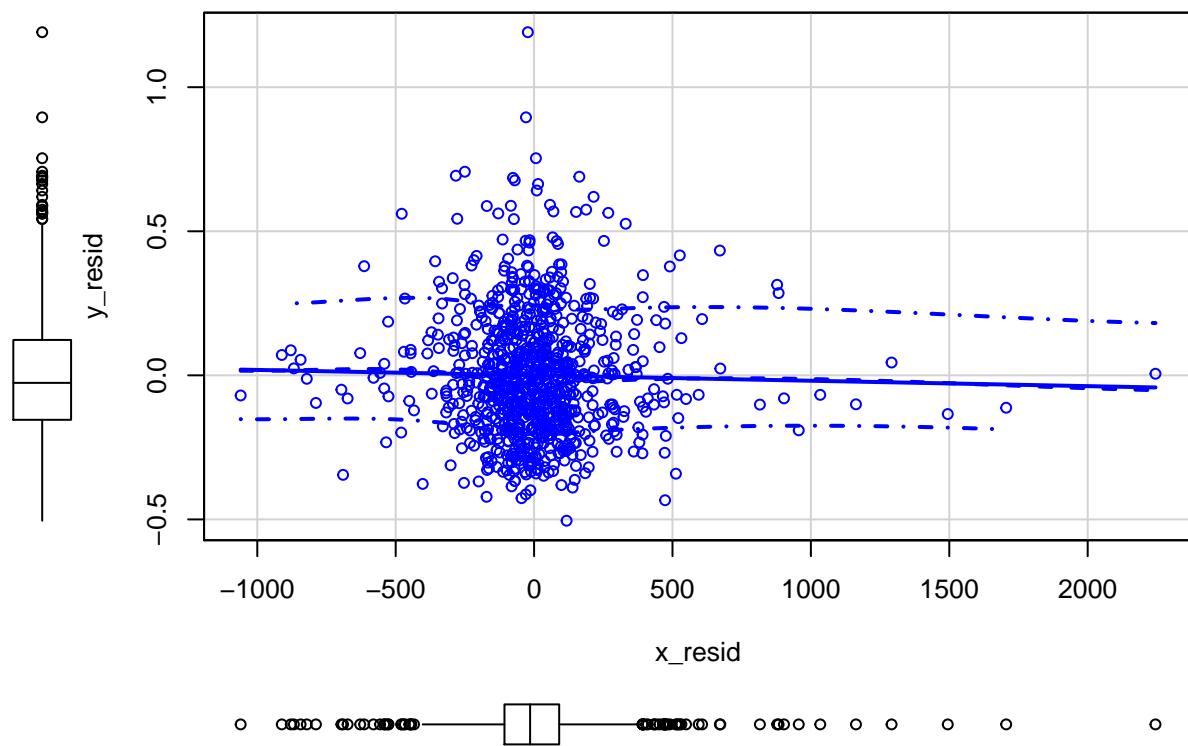


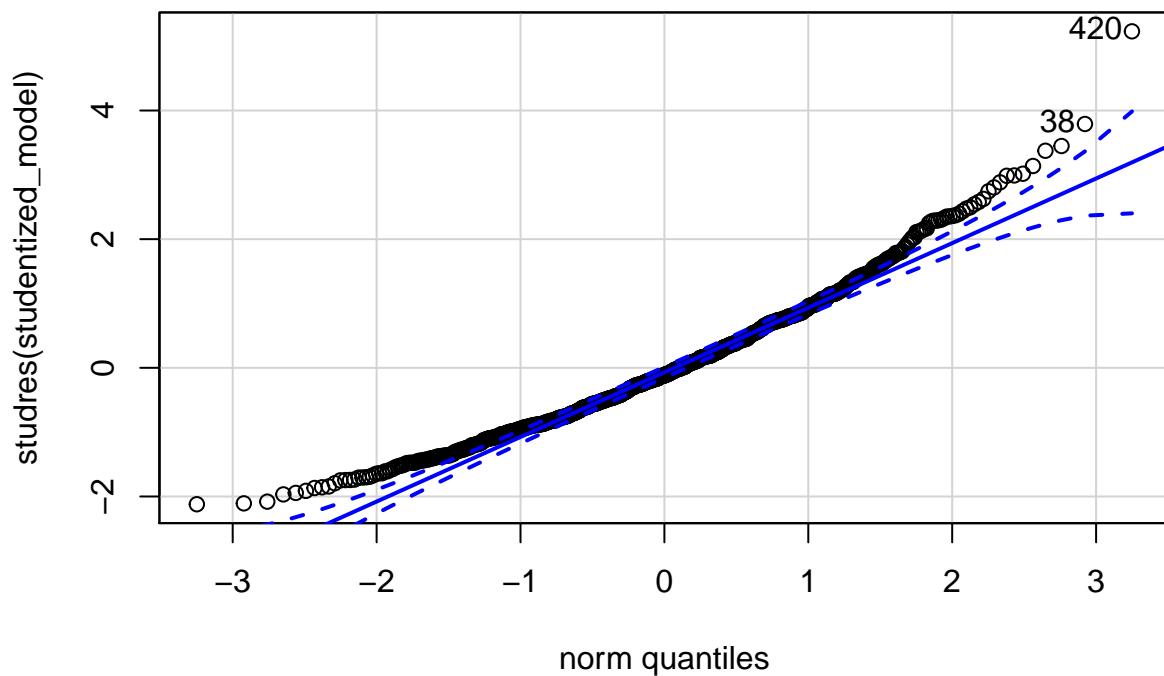


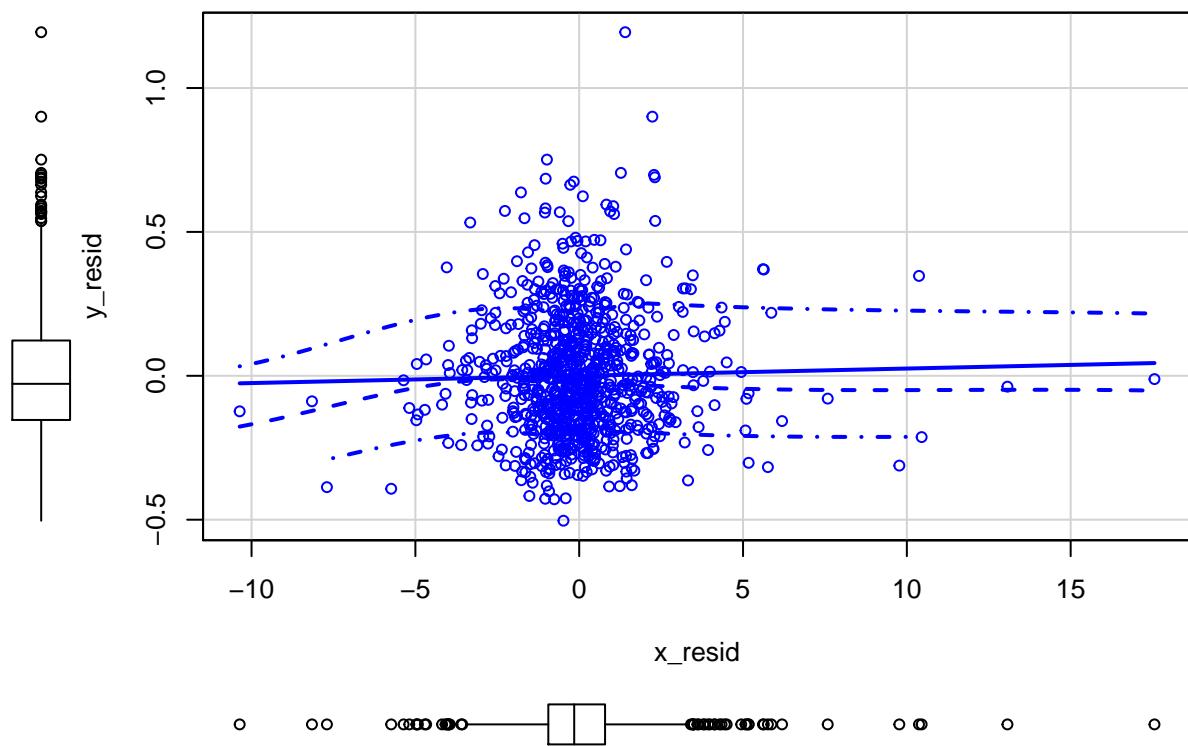


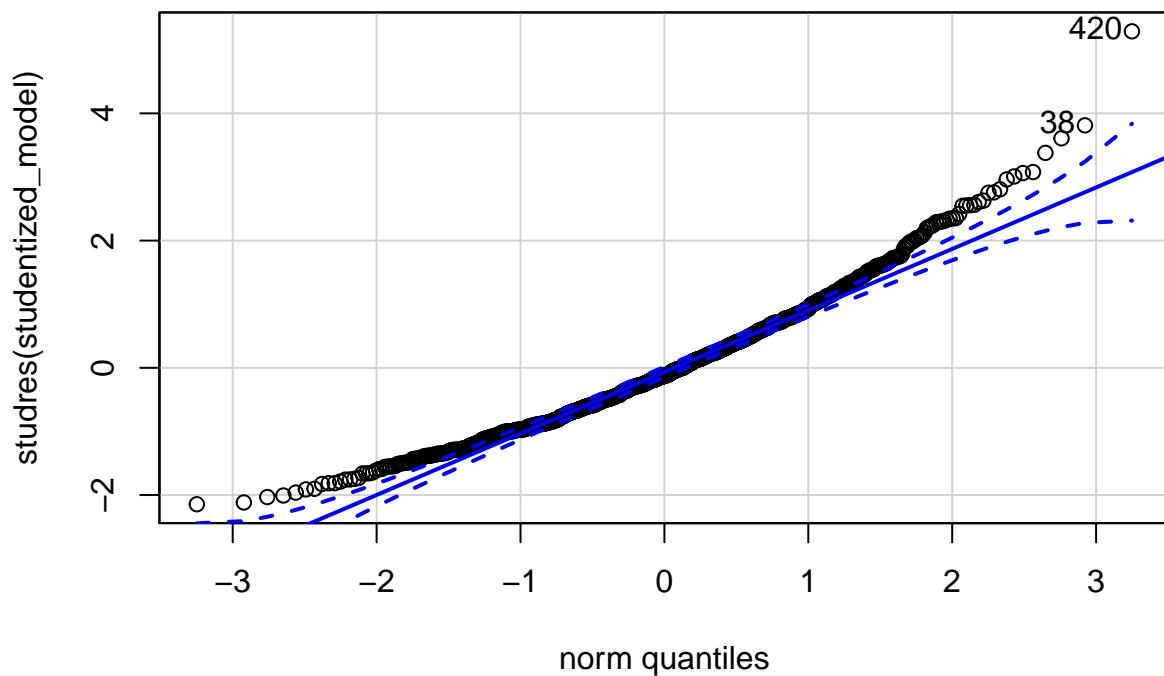


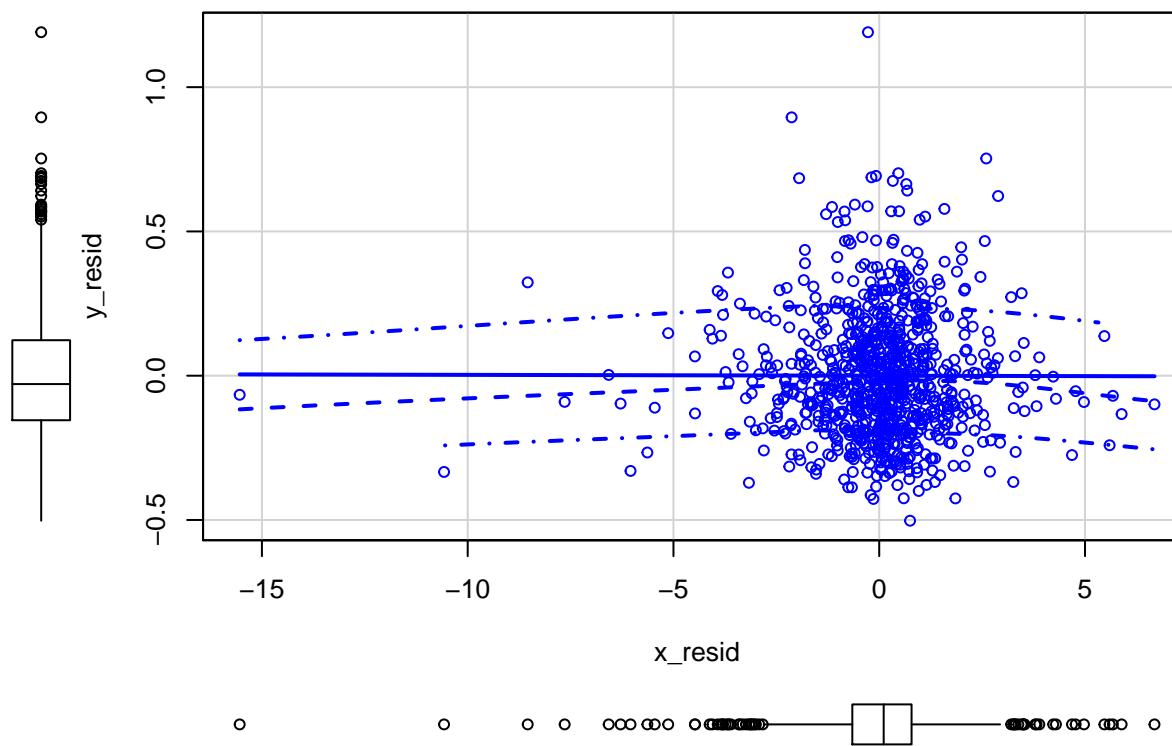


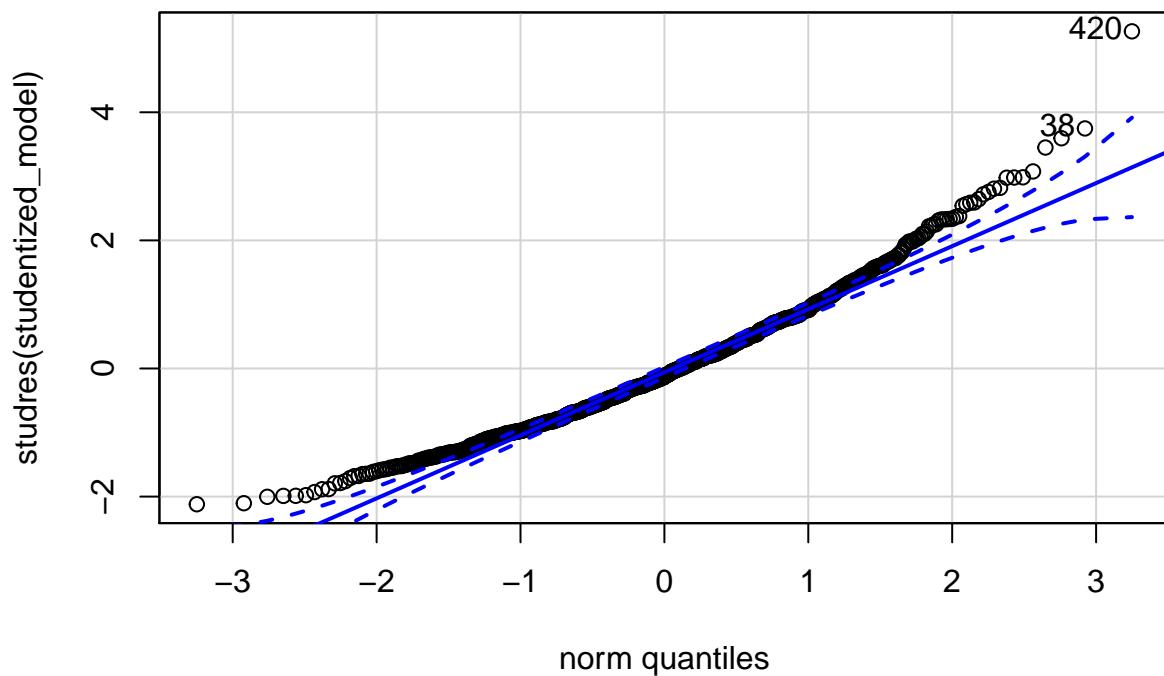


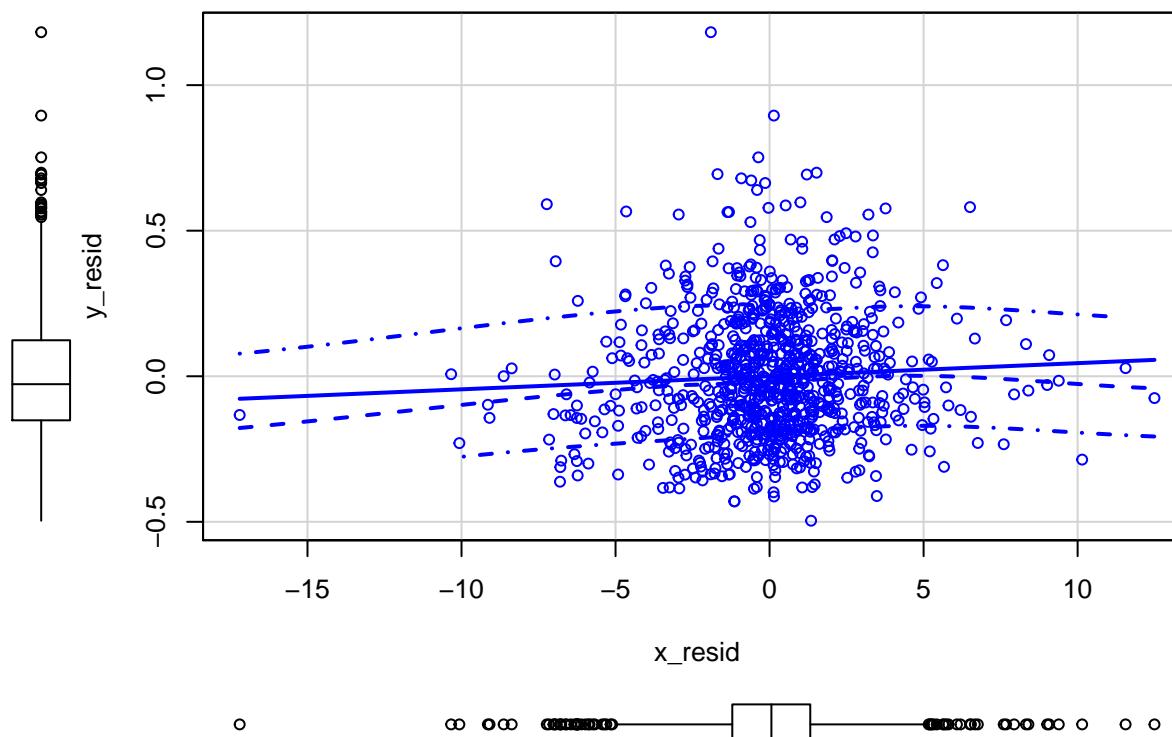


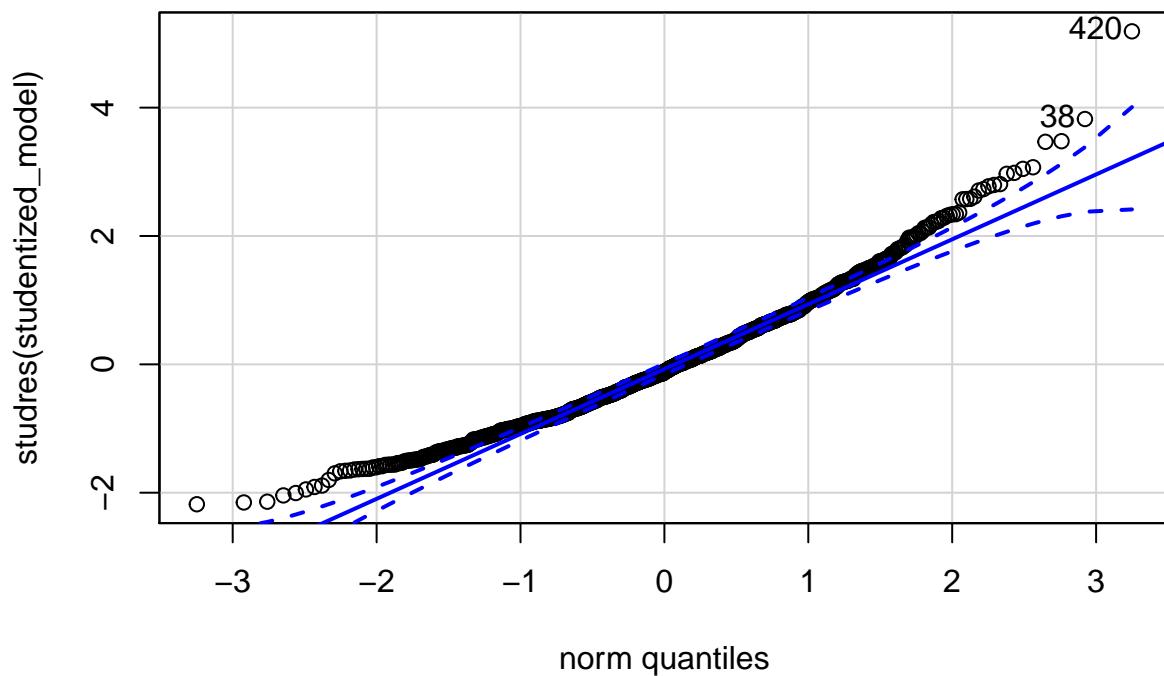


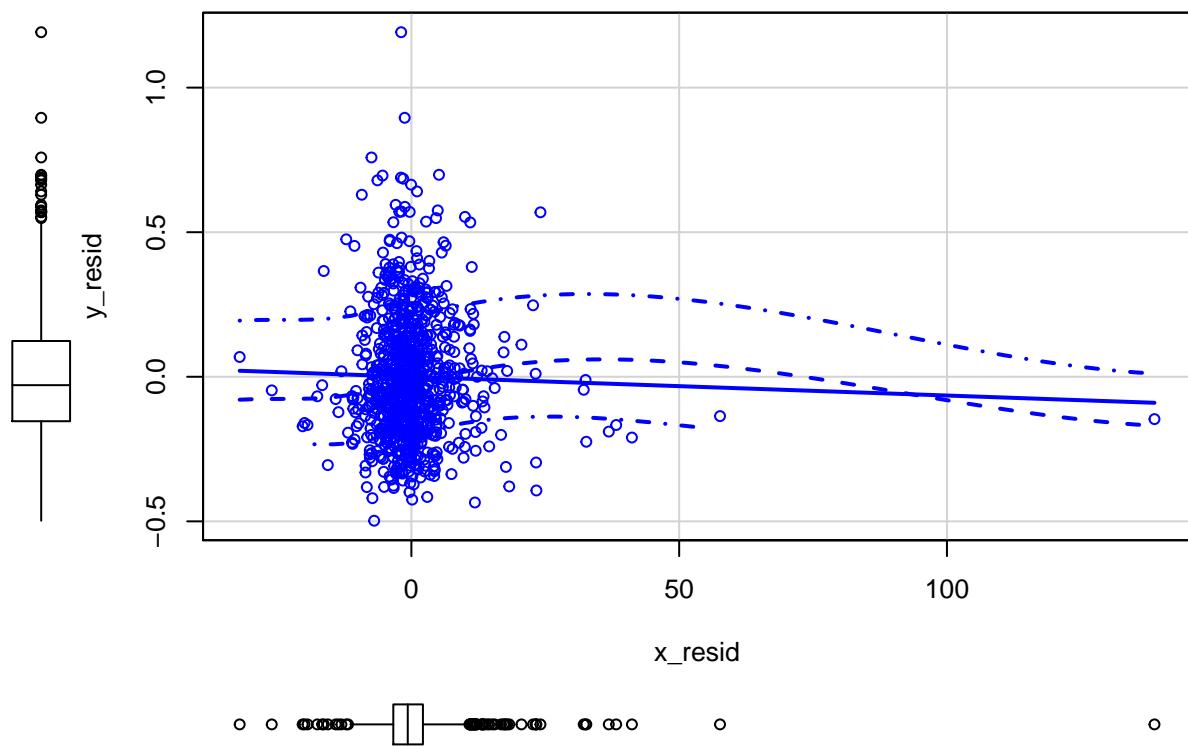


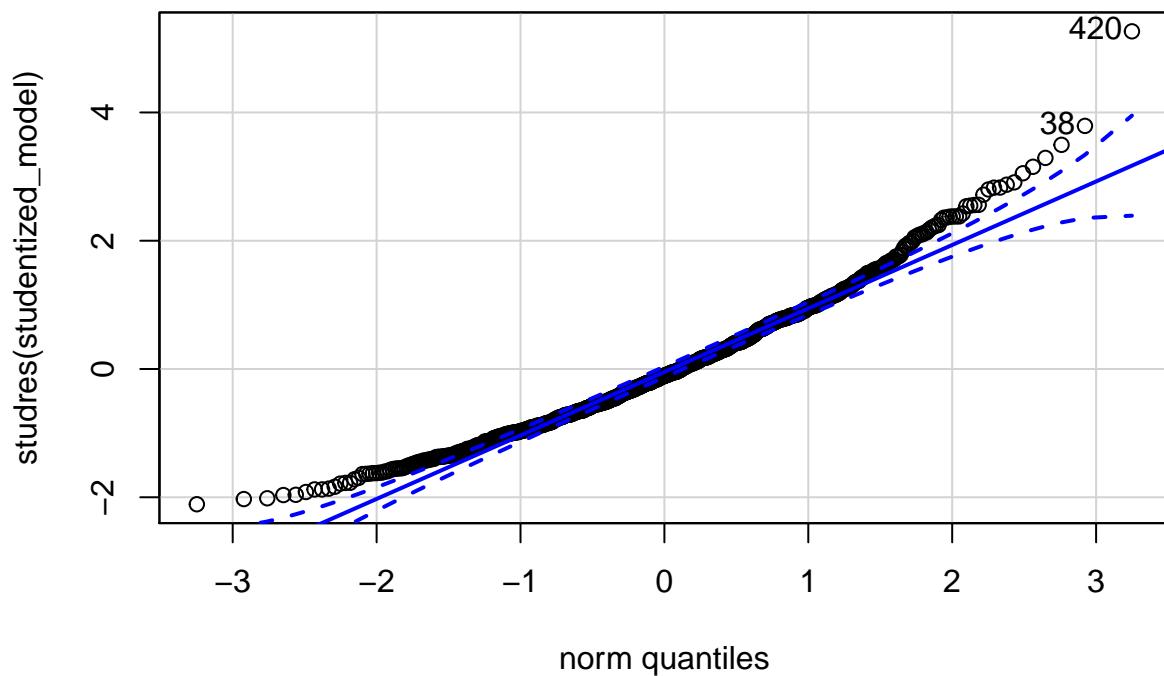


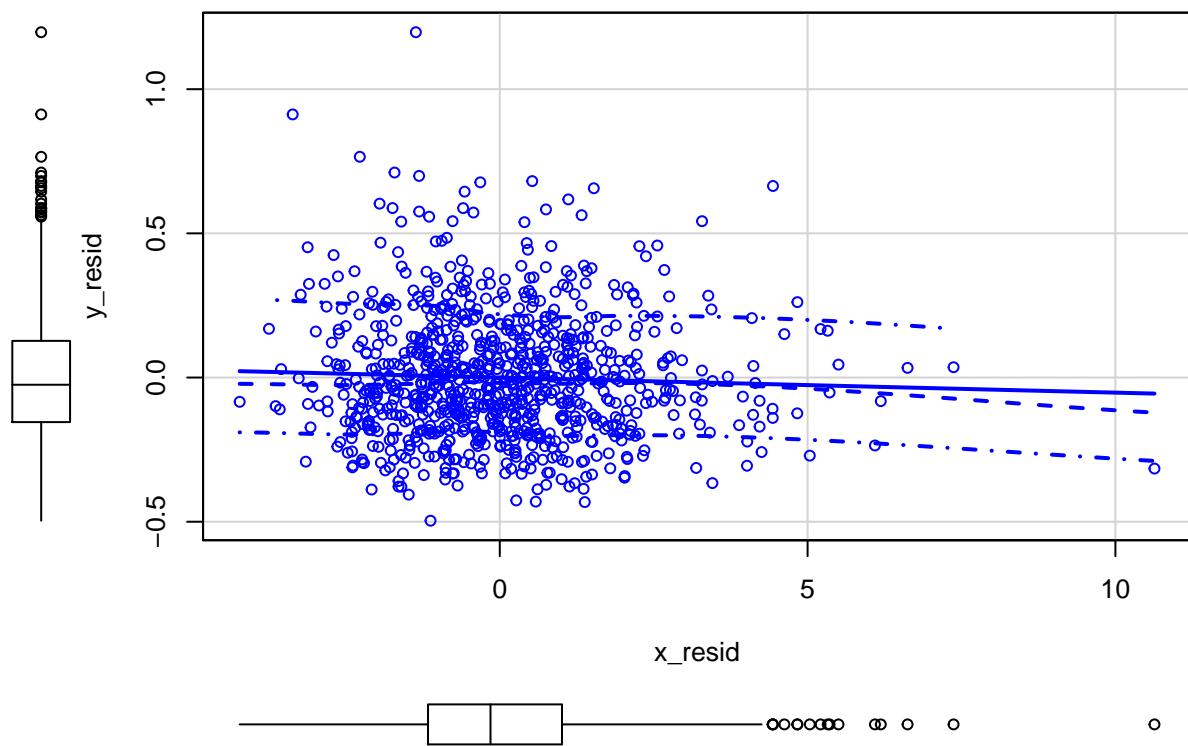


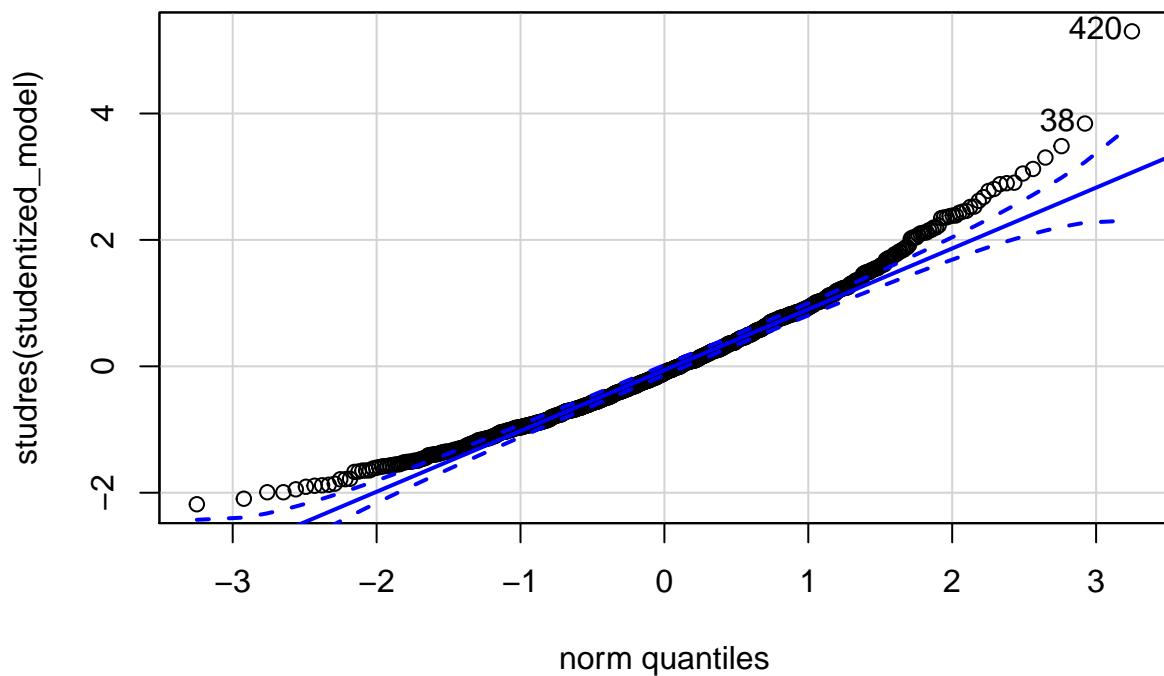


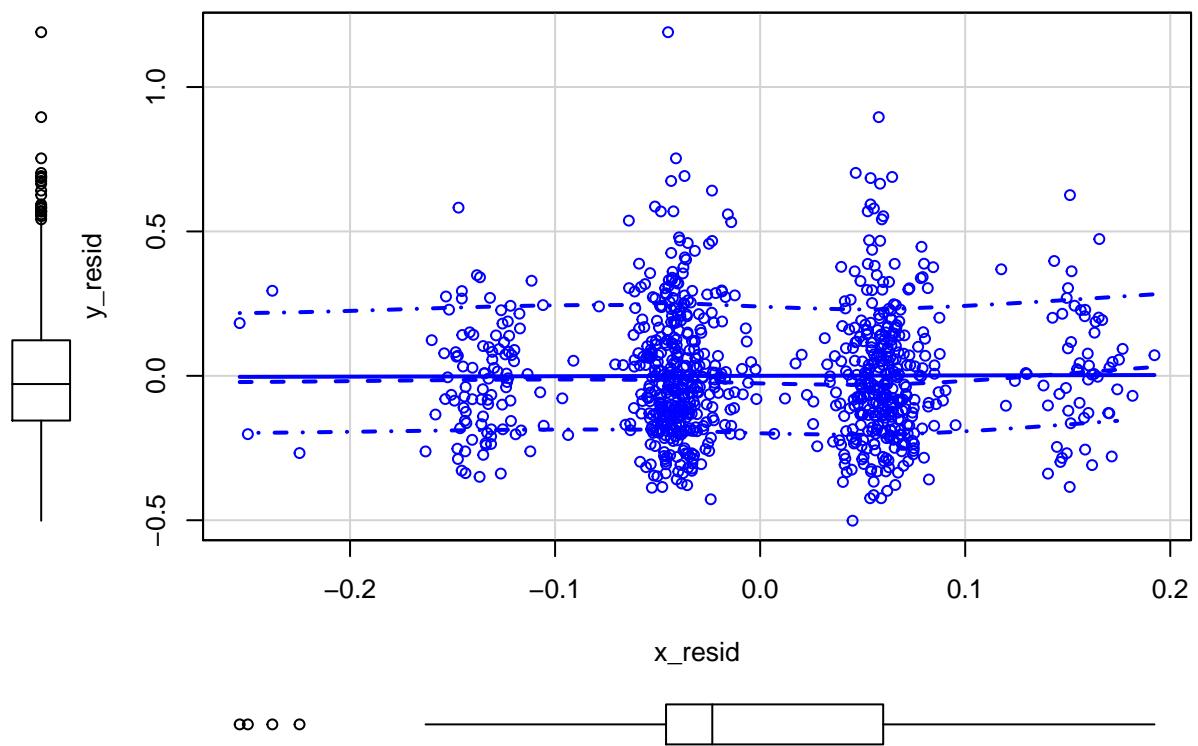


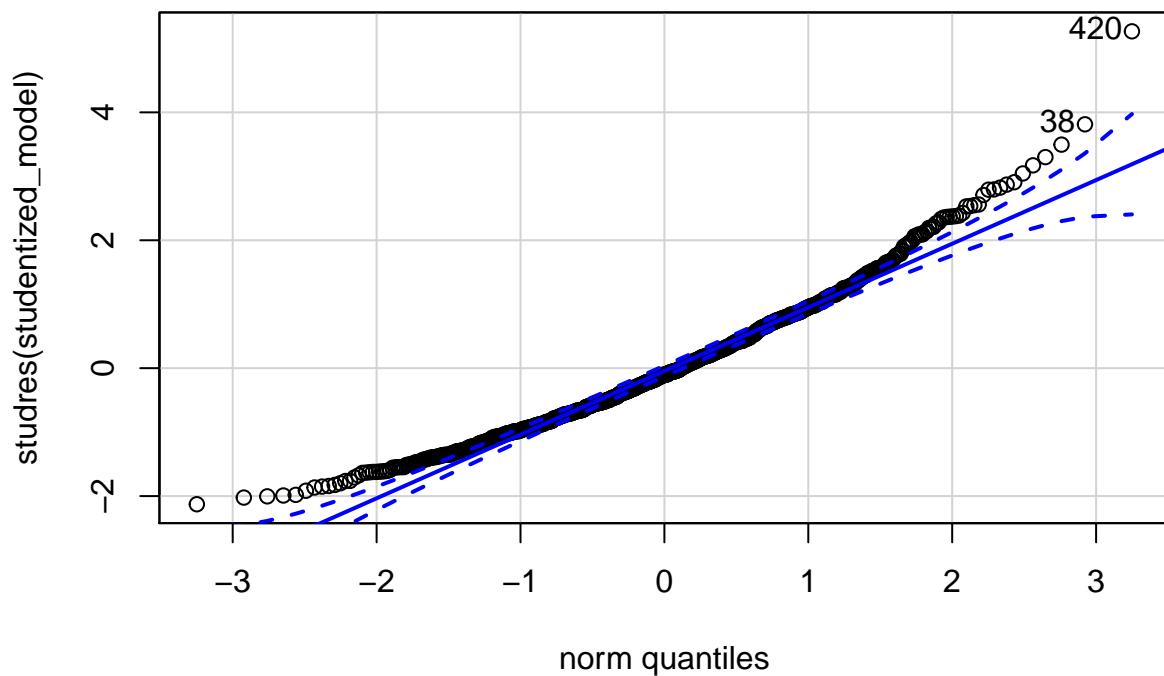


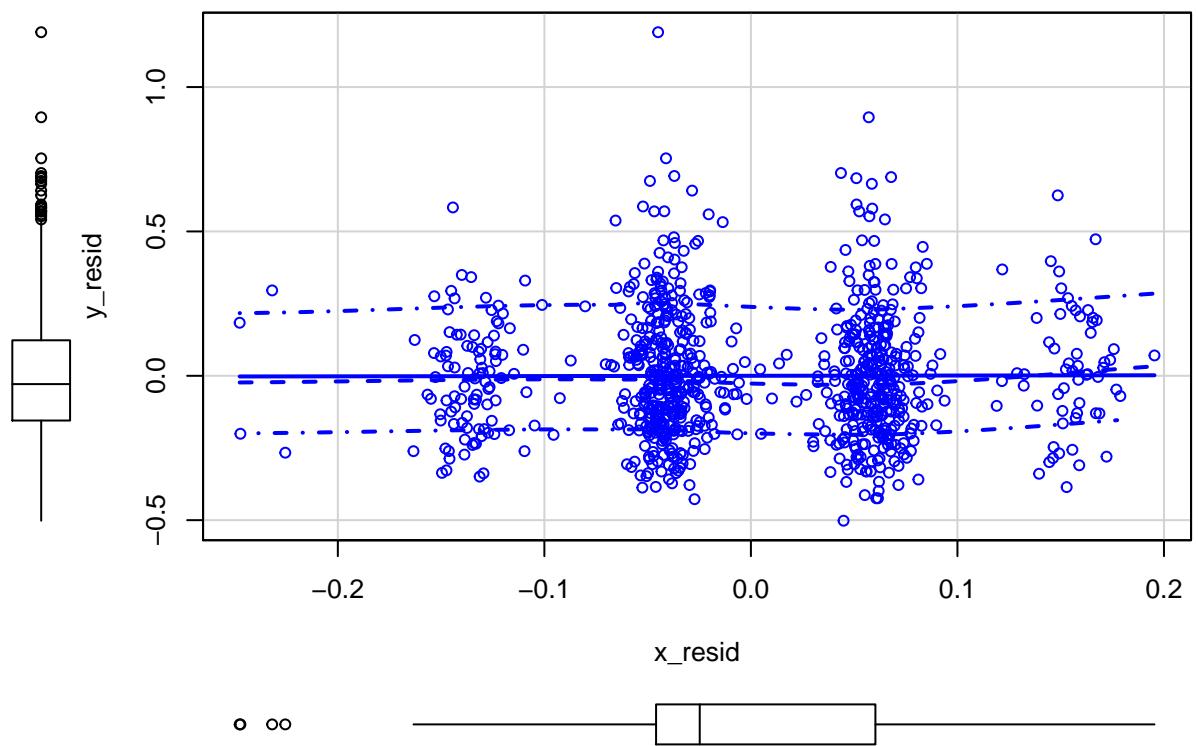


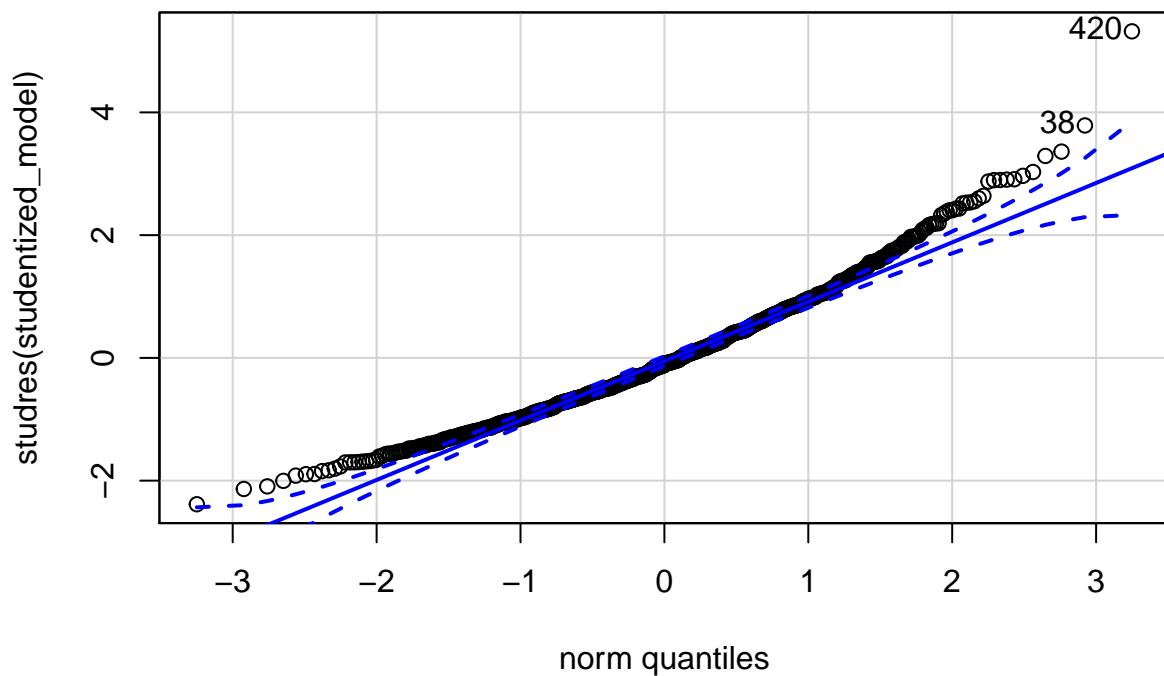


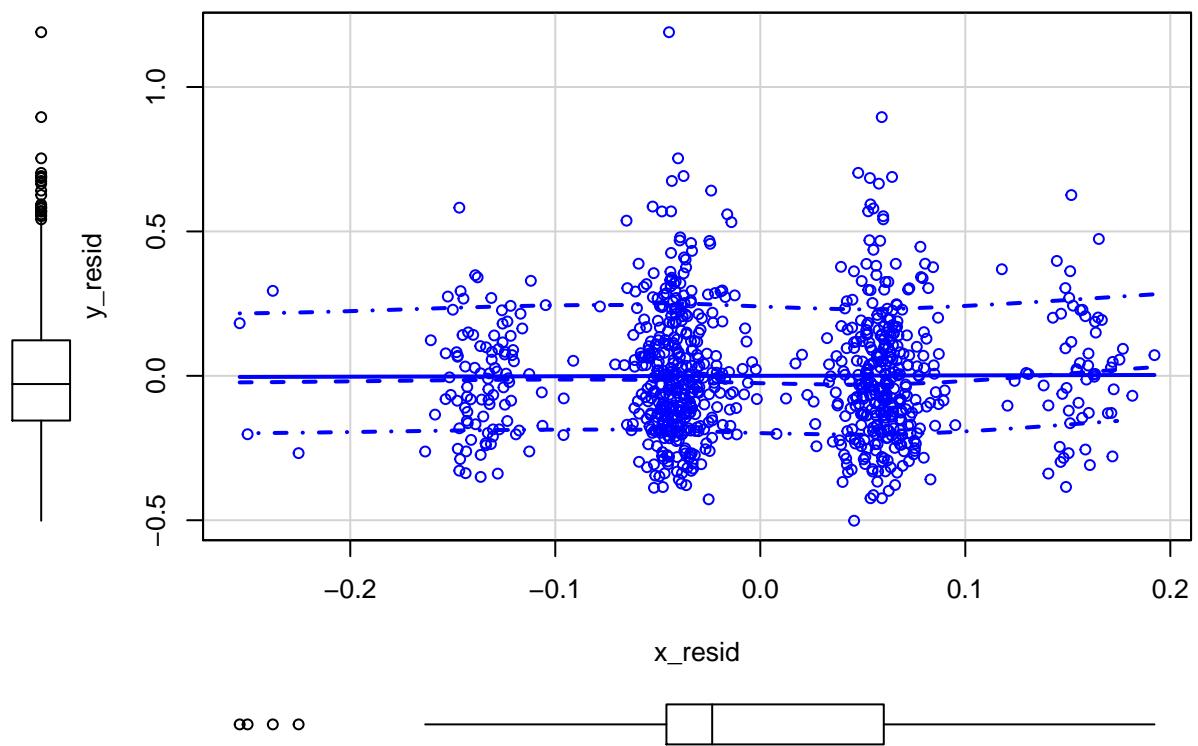


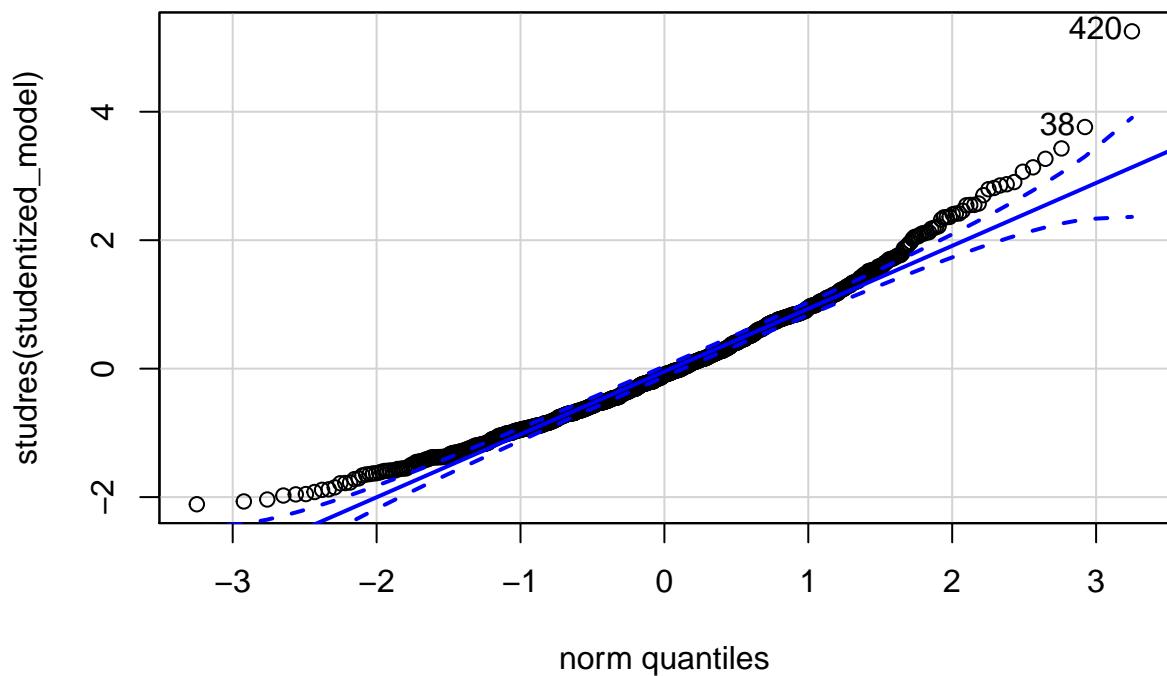


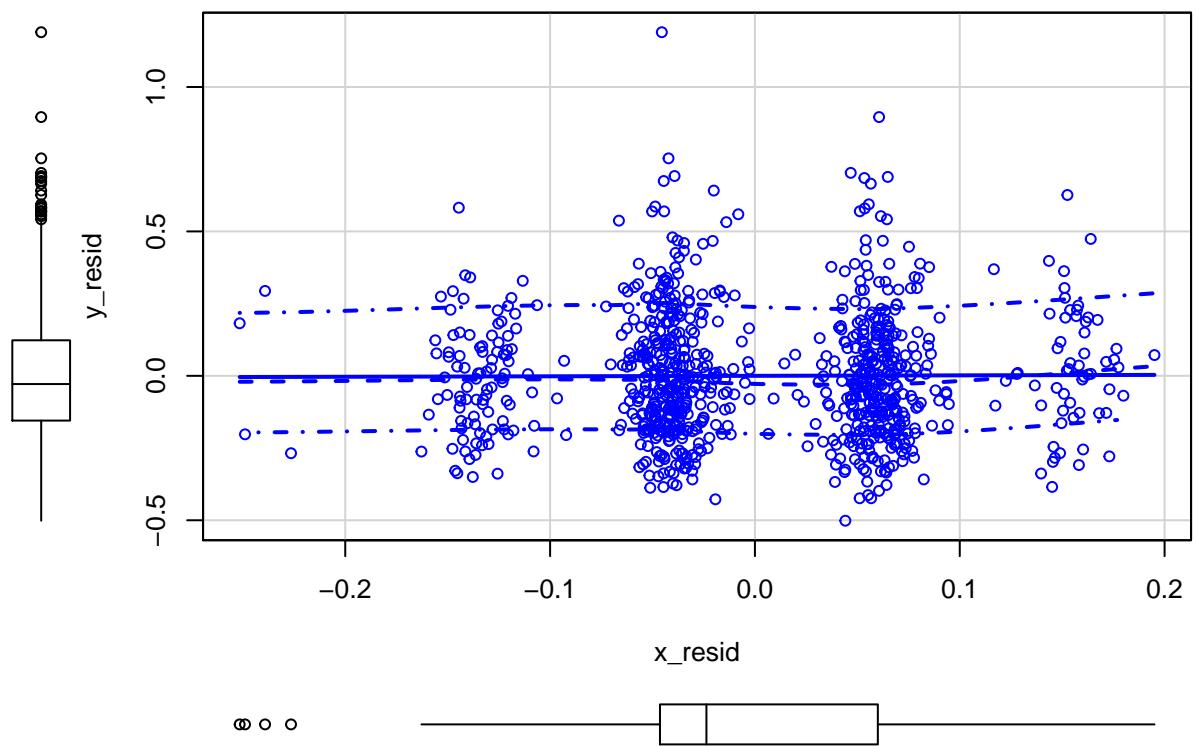


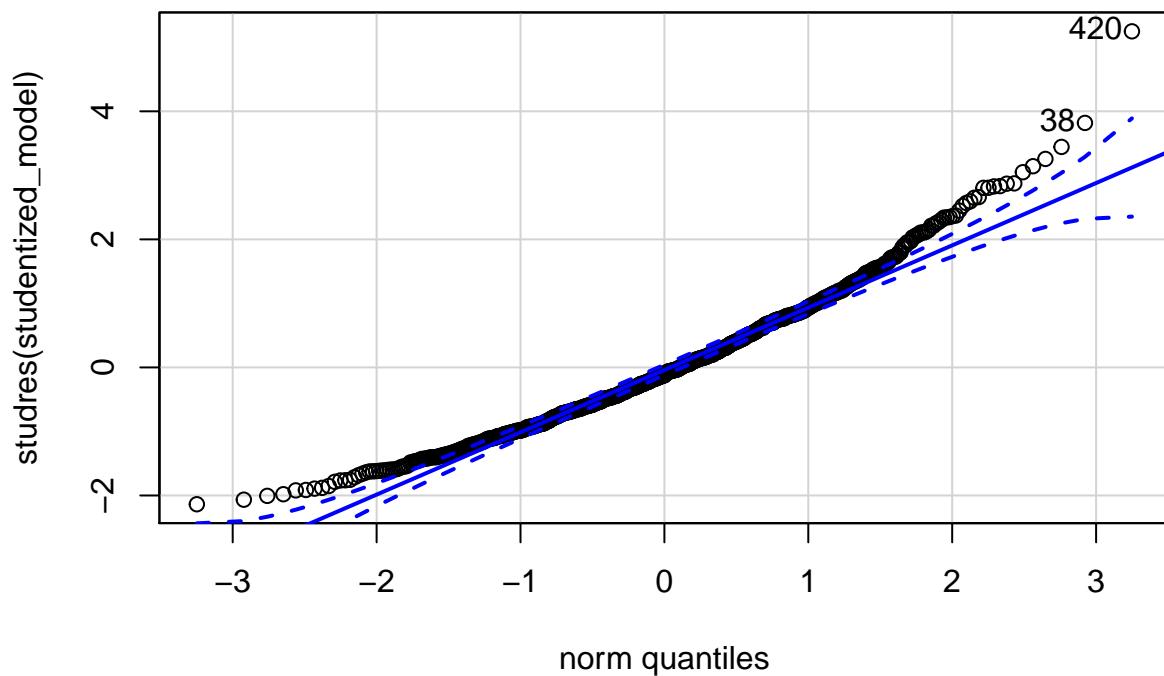


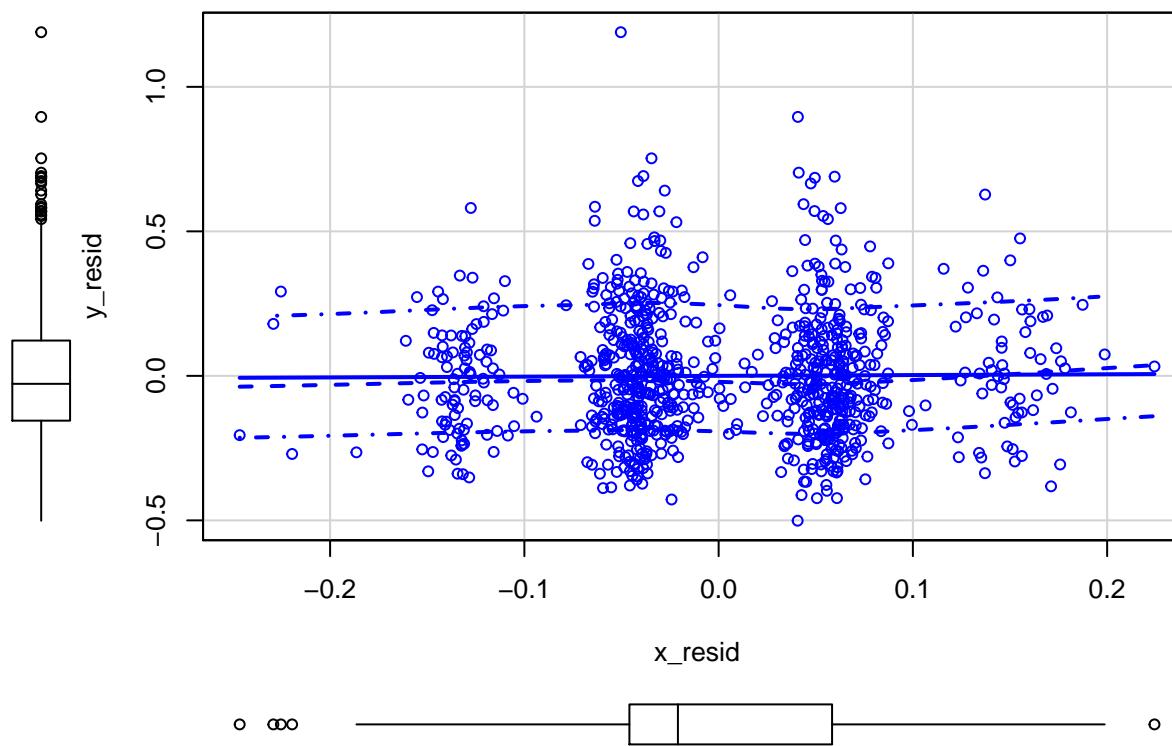


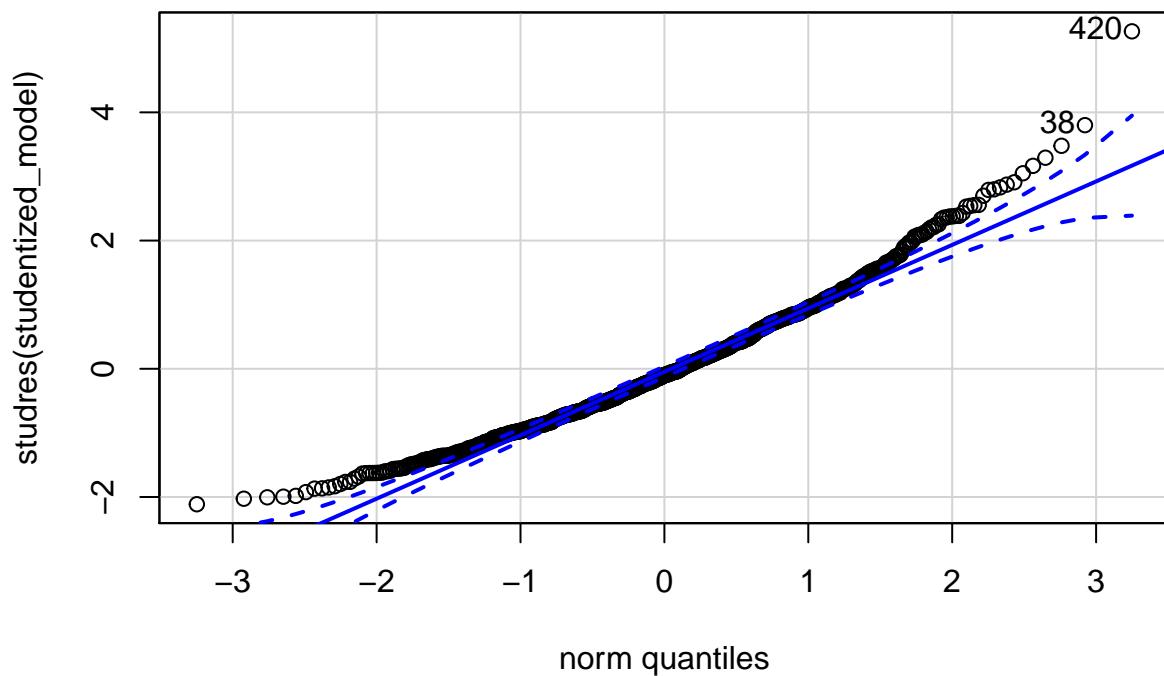


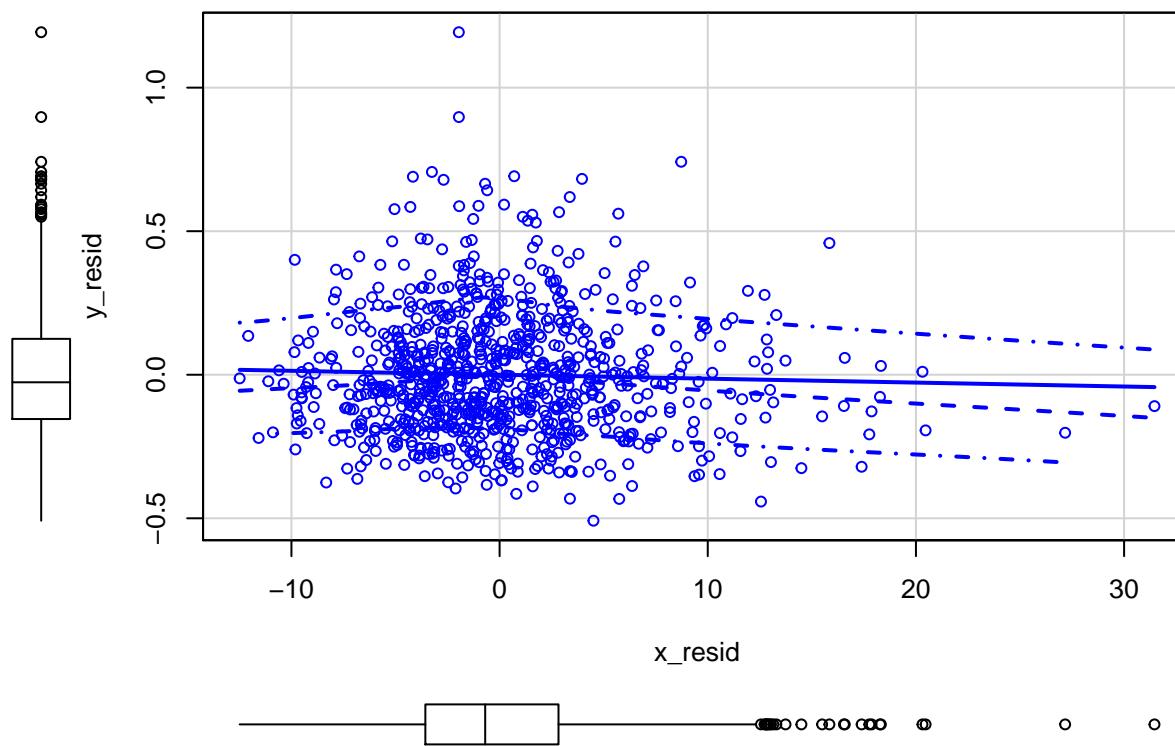


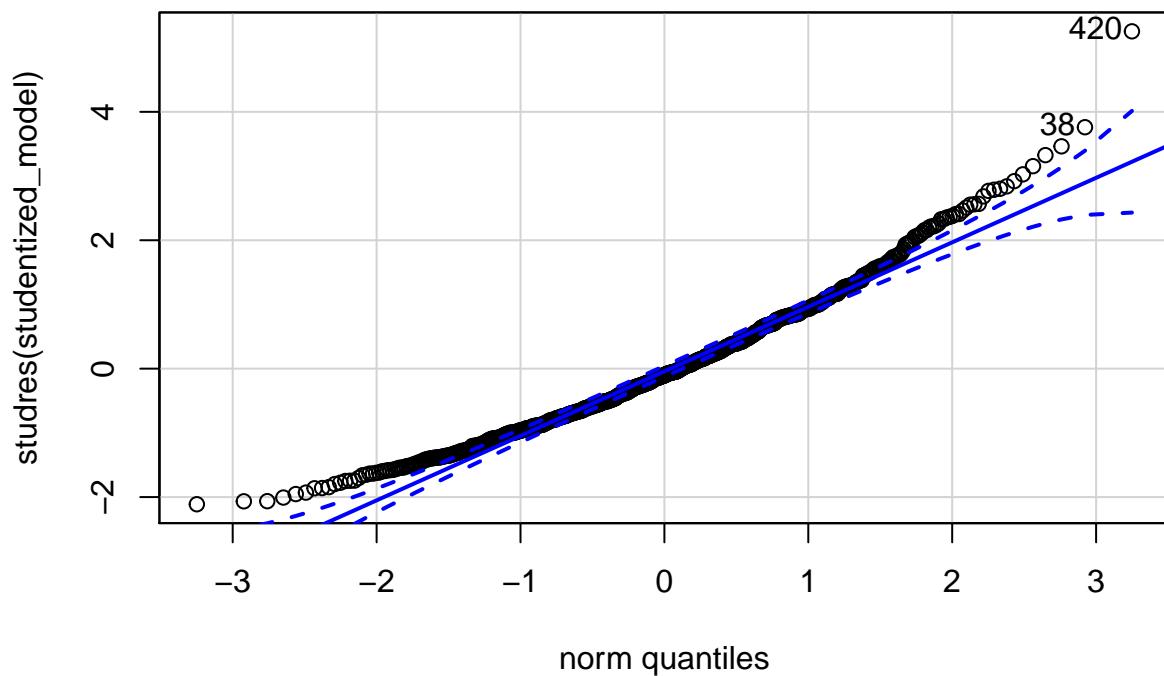


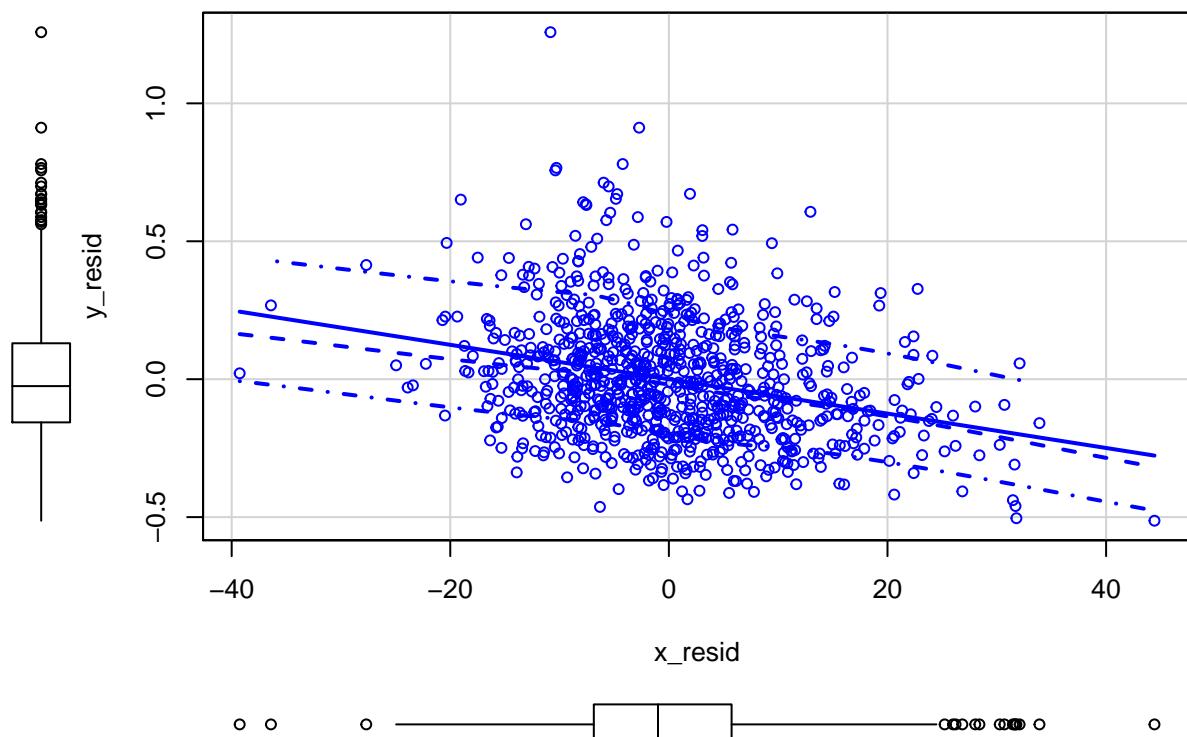


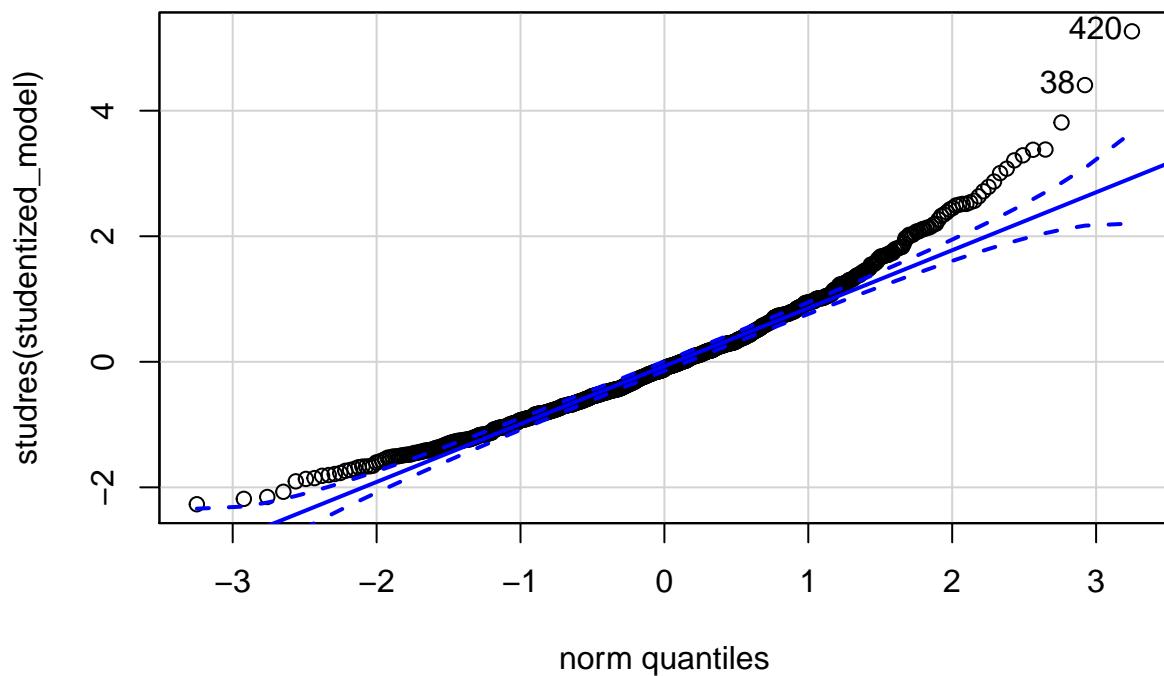


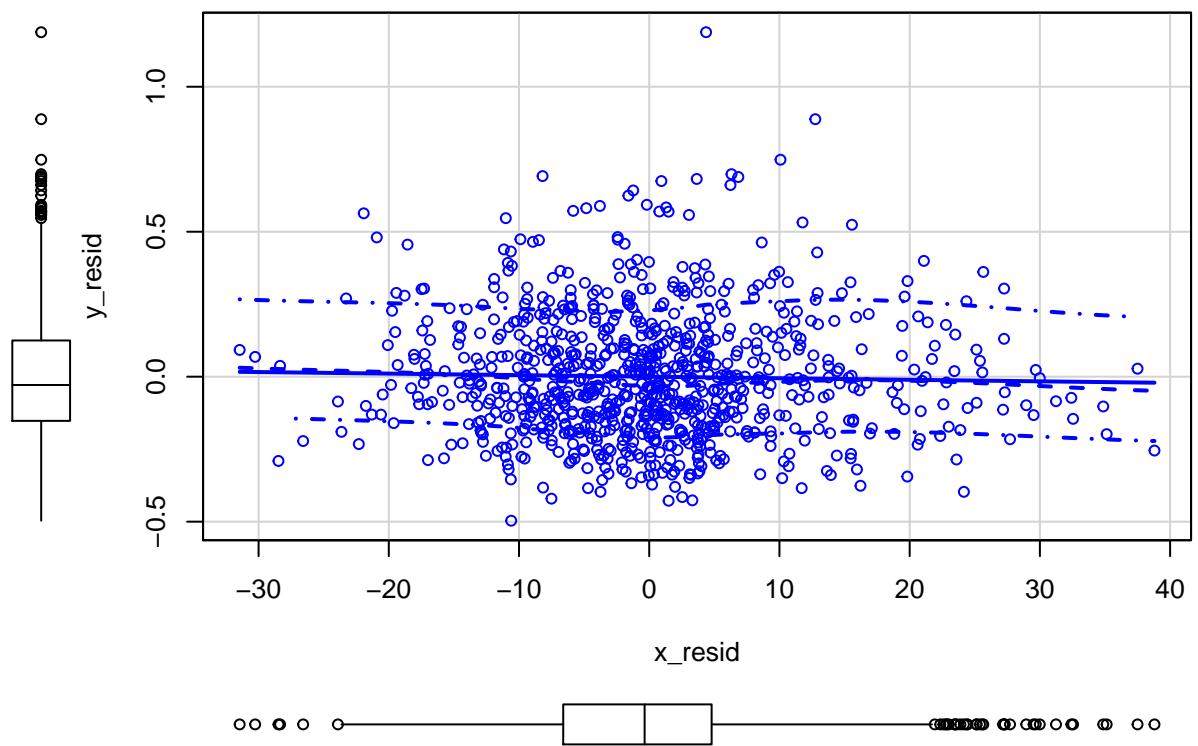


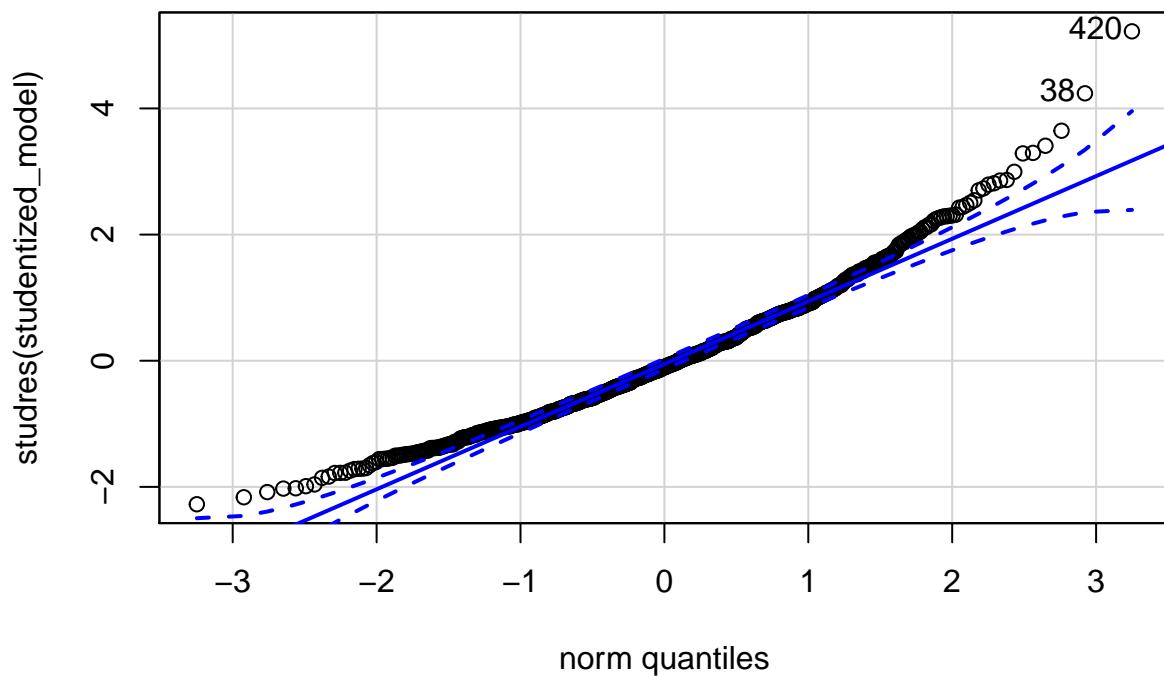


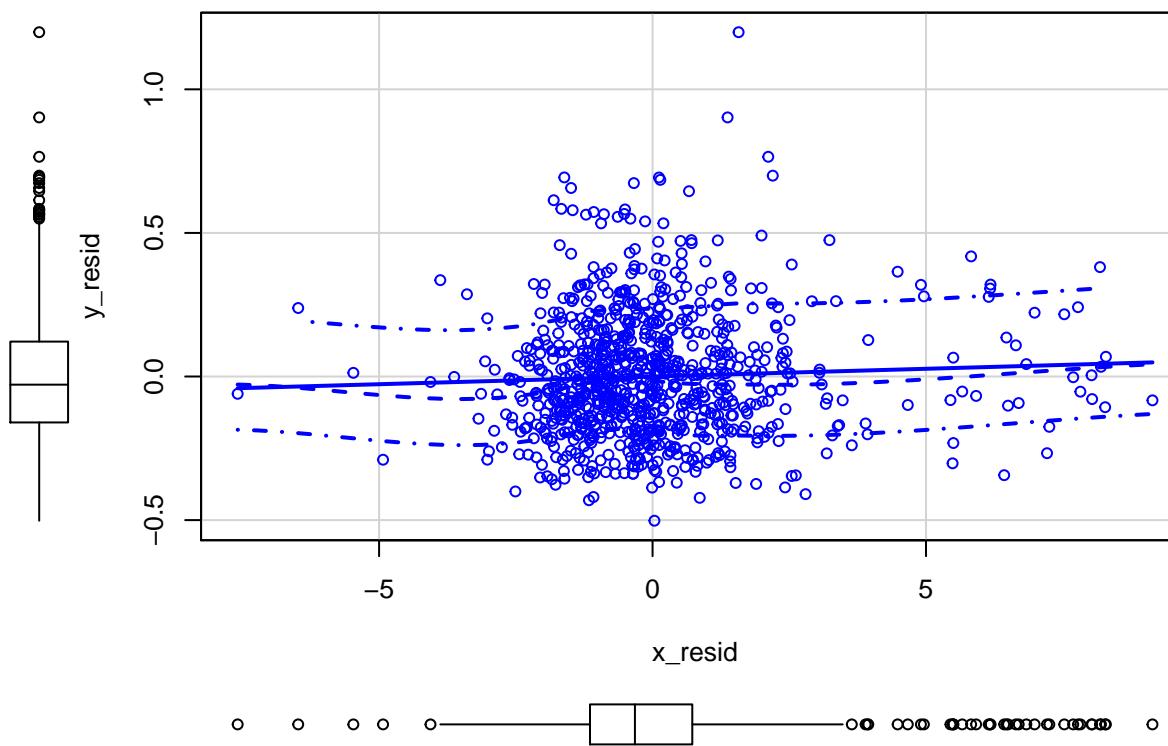


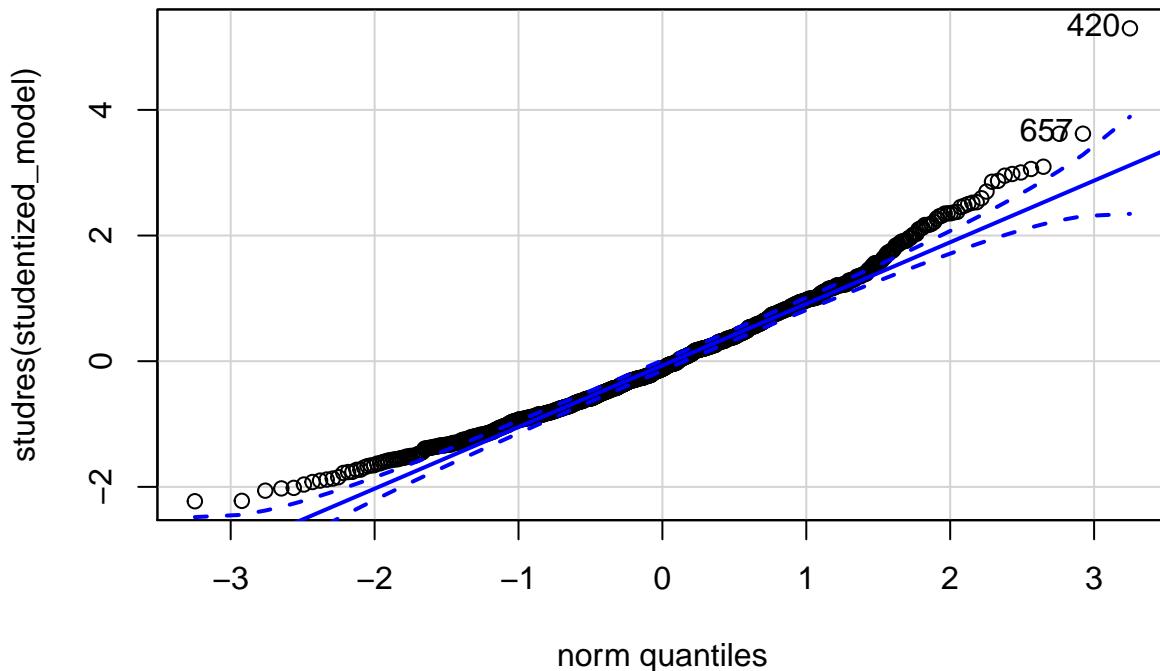












```
#remove covariates with VIF > 10 below:

#remove highest vif ~ 15047
pollutants$eosinophils_pct = NULL
#fit new model
model = lm(length ~ ., data = pollutants)
#summary
summary(model)

##
## Call:
## lm(formula = length ~ ., data = pollutants)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.50160 -0.15463 -0.02843  0.12293  1.18974 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.491e+00 8.425e-02 17.693 < 2e-16 ***
## POP_PCB1    -1.606e-06 1.075e-06 -1.494  0.1355    
## POP_PCB2     7.294e-07 3.021e-06  0.241  0.8093    
## POP_PCB3     1.191e-06 2.156e-06  0.552  0.5809    
## POP_PCB4    -1.774e-07 1.025e-06 -0.173  0.8627    
## POP_PCB5     1.403e-07 1.068e-06  0.131  0.8955    
## POP_PCB6     2.765e-07 1.058e-06  0.261  0.7939
```

```

## POP_PCB7      -5.739e-07  1.206e-06 -0.476  0.6343
## POP_PCB8      1.644e-06  2.446e-06  0.672  0.5016
## POP_PCB9      6.194e-07  2.111e-06  0.293  0.7693
## POP_PCB10     1.191e-03  8.890e-04  1.340  0.1806
## POP_PCB11     3.449e-05  3.078e-04  0.112  0.9108
## POP_dioxin1    3.072e-05  3.048e-04  0.101  0.9198
## POP_dioxin2   -1.744e-04  4.394e-04 -0.397  0.6916
## POP_dioxin3   -1.908e-05  3.018e-05 -0.632  0.5274
## POP_furan1     2.522e-03  3.844e-03  0.656  0.5119
## POP_furan2    -2.799e-04  4.500e-03 -0.062  0.9504
## POP_furan3     4.473e-03  2.756e-03  1.623  0.1049
## POP_furan4    -6.496e-04  9.195e-04 -0.706  0.4801
## whitecell_count -5.261e-03  4.404e-03 -1.195  0.2326
## lymphocyte_pct  -1.255e-03  1.030e-03 -1.219  0.2234
## monocyte_pct   -5.993e-03  4.041e-03 -1.483  0.1384
## basophils_pct   1.051e-03  3.475e-03  0.302  0.7624
## neutrophils_pct 1.213e-02  1.662e-02  0.730  0.4657
## BMI            -1.359e-03  1.409e-03 -0.964  0.3352
## edu_cat2       2.452e-02  2.215e-02  1.107  0.2687
## edu_cat3       4.787e-02  2.164e-02  2.212  0.0272 *
## edu_cat4       3.265e-02  2.555e-02  1.278  0.2016
## race_cat2      -2.195e-02  3.265e-02 -0.672  0.5016
## race_cat3      2.475e-02  3.370e-02  0.735  0.4628
## race_cat4      -3.473e-02  2.991e-02 -1.161  0.2460
## male1          -3.944e-02  1.771e-02 -2.227  0.0262 *
## ageyrs         -6.236e-03  7.442e-04 -8.380  2.25e-16 ***
## yrssmoke      -5.323e-04  7.271e-04 -0.732  0.4643
## smokenow1      2.172e-03  3.581e-02  0.061  0.9516
## ln_lbxcot      5.341e-03  3.922e-03  1.362  0.1736
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.222 on 828 degrees of freedom
## Multiple R-squared:  0.2448, Adjusted R-squared:  0.2129
## F-statistic: 7.668 on 35 and 828 DF,  p-value: < 2.2e-16
#show the VIF
vif(model)

```

```

##                  GVIF Df GVIF^(1/(2*Df))
## POP_PCB1        33.040956  1      5.748126
## POP_PCB2        34.276806  1      5.854640
## POP_PCB3        9.350969  1      3.057935
## POP_PCB4       31.734341  1      5.633324
## POP_PCB5       59.718357  1      7.727765
## POP_PCB6       11.386137  1      3.374335
## POP_PCB7        4.868942  1      2.206568
## POP_PCB8       12.982530  1      3.603128
## POP_PCB9       12.416573  1      3.523716
## POP_PCB10      5.988829  1      2.447208
## POP_PCB11      4.725385  1      2.173795
## POP_dioxin1     5.256334  1      2.292670
## POP_dioxin2     5.411476  1      2.326258
## POP_dioxin3     4.378774  1      2.092552
## POP_furan1      6.154213  1      2.480769

```

```

## POP_furan2      6.193725  1      2.488720
## POP_furan3      4.450557  1      2.109634
## POP_furan4      1.821773  1      1.349731
## whitecell_count 1.545798  1      1.243301
## lymphocyte_pct   1.382541  1      1.175815
## monocyte_pct     1.263611  1      1.124105
## basophils_pct    1.113199  1      1.055082
## neutrophils_pct  1.090031  1      1.044046
## BMI              1.261934  1      1.123358
## edu_cat          1.541083  3      1.074742
## race_cat          2.051619  3      1.127239
## male              1.350208  1      1.161984
## ageyrs            3.237762  1      1.799378
## yrssmoke          2.204134  1      1.484633
## smokenow          3.998531  1      1.999633
## ln_lbxcot         3.954234  1      1.988526

#remove highest vif eosinophils_pct ~ 15047 then POP_PCB5 ~ 59.718357
pollutants$POP_PCB5 = NULL
#fit new model
model = lm(length ~ ., data = pollutants)
#summary
summary(model)

##
## Call:
## lm(formula = length ~ ., data = pollutants)
##
## Residuals:
##       Min     1Q   Median     3Q    Max
## -0.50233 -0.15456 -0.02848  0.12242  1.19007
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.491e+00 8.419e-02 17.706 <2e-16 ***
## POP_PCB1    -1.584e-06 1.061e-06 -1.493  0.1359    
## POP_PCB2     8.659e-07 2.835e-06  0.305  0.7601    
## POP_PCB3     1.229e-06 2.135e-06  0.575  0.5652    
## POP_PCB4    -9.537e-08 8.129e-07 -0.117  0.9066    
## POP_PCB6     3.209e-07 1.002e-06  0.320  0.7489    
## POP_PCB7    -5.874e-07 1.201e-06 -0.489  0.6248    
## POP_PCB8     1.574e-06 2.386e-06  0.660  0.5095    
## POP_PCB9     7.223e-07 1.960e-06  0.369  0.7125    
## POP_PCB10    1.211e-03 8.759e-04  1.382  0.1672    
## POP_PCB11    2.293e-05 2.947e-04  0.078  0.9380    
## POP_dioxin1   2.999e-05 3.046e-04  0.098  0.9216    
## POP_dioxin2   -1.693e-04 4.374e-04 -0.387  0.6989    
## POP_dioxin3   -1.949e-05 3.001e-05 -0.650  0.5161    
## POP_furan1    2.449e-03 3.801e-03  0.644  0.5196    
## POP_furan2    -2.463e-04 4.490e-03 -0.055  0.9563    
## POP_furan3    4.477e-03 2.754e-03  1.626  0.1044    
## POP_furan4    -6.429e-04 9.176e-04 -0.701  0.4837    
## whitecell_count -5.303e-03 4.390e-03 -1.208  0.2274    
## lymphocyte_pct  -1.254e-03 1.029e-03 -1.219  0.2233    
## monocyte_pct   -5.978e-03 4.037e-03 -1.481  0.1390

```

```

## basophils_pct    1.012e-03  3.460e-03  0.292   0.7700
## neutrophils_pct 1.213e-02  1.661e-02  0.731   0.4653
## BMI            -1.357e-03  1.408e-03 -0.964   0.3356
## edu_cat2       2.450e-02  2.214e-02  1.107   0.2688
## edu_cat3       4.786e-02  2.163e-02  2.213   0.0272 *
## edu_cat4       3.284e-02  2.550e-02  1.288   0.1981
## race_cat2      -2.186e-02  3.263e-02 -0.670   0.5030
## race_cat3      2.503e-02  3.361e-02  0.745   0.4566
## race_cat4      -3.460e-02  2.988e-02 -1.158   0.2472
## male1          -3.945e-02  1.770e-02 -2.228   0.0261 *
## ageyrs         -6.237e-03  7.437e-04 -8.387 <2e-16 ***
## yrssmoke       -5.391e-04  7.248e-04 -0.744   0.4572
## smokenow1      2.165e-03  3.579e-02  0.061   0.9518
## ln_lbxcot      5.346e-03  3.919e-03  1.364   0.1729
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2219 on 829 degrees of freedom
## Multiple R-squared:  0.2448, Adjusted R-squared:  0.2138
## F-statistic: 7.902 on 34 and 829 DF,  p-value: < 2.2e-16

```

#show the VIF

```
vif(model)
```

	GVIF	Df	GVIF^(1/(2*Df))
## POP_PCB1	32.255513	1	5.679394
## POP_PCB2	30.220860	1	5.497350
## POP_PCB3	9.182804	1	3.030314
## POP_PCB4	19.965194	1	4.468243
## POP_PCB6	10.226099	1	3.197827
## POP_PCB7	4.833396	1	2.198498
## POP_PCB8	12.363256	1	3.516142
## POP_PCB9	10.707771	1	3.272273
## POP_PCB10	5.820661	1	2.412605
## POP_PCB11	4.338648	1	2.082942
## POP_dioxin1	5.254606	1	2.292293
## POP_dioxin2	5.368810	1	2.317069
## POP_dioxin3	4.332013	1	2.081349
## POP_furan1	6.023932	1	2.454370
## POP_furan2	6.173626	1	2.484678
## POP_furan3	4.450098	1	2.109526
## POP_furan4	1.816182	1	1.347658
## whitecell_count	1.537750	1	1.240060
## lymphocyte_pct	1.382511	1	1.175802
## monocyte_pct	1.262603	1	1.123656
## basophils_pct	1.105130	1	1.051251
## neutrophils_pct	1.090023	1	1.044042
## BMI	1.261809	1	1.123303
## edu_cat	1.533185	3	1.073822
## race_cat	2.042113	3	1.126367
## male	1.350206	1	1.161984
## ageyrs	3.237570	1	1.799325
## yrssmoke	2.192917	1	1.480850
## smokenow	3.998522	1	1.999630
## ln_lbxcot	3.953781	1	1.988412

```

#remove highest vif eosinophils_pct ~ 15047 then POP_PCB5 ~ 59.718357 then POP_PCB1 ~ 32.255513
pollutants$POP_PCB1 = NULL
#fit new model
model = lm(length ~ ., data = pollutants)
#summary
summary(model)

## 
## Call:
## lm(formula = length ~ ., data = pollutants)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.49329 -0.15025 -0.02981  0.12229  1.18766 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.505e+00 8.371e-02 17.977 <2e-16 ***
## POP_PCB2    -1.264e-06 2.451e-06 -0.516  0.6062    
## POP_PCB3    1.587e-06 2.123e-06  0.747  0.4551    
## POP_PCB4    -2.160e-07 8.094e-07 -0.267  0.7896    
## POP_PCB6    1.820e-07 9.987e-07  0.182  0.8555    
## POP_PCB7    -6.786e-07 1.200e-06 -0.565  0.5719    
## POP_PCB8    -1.708e-07 2.081e-06 -0.082  0.9346    
## POP_PCB9    -8.899e-09 1.899e-06 -0.005  0.9963    
## POP_PCB10   1.035e-03 8.686e-04  1.192  0.2337    
## POP_PCB11   9.381e-05 2.911e-04  0.322  0.7473    
## POP_dioxin1 3.693e-05 3.048e-04  0.121  0.9036    
## POP_dioxin2 -1.018e-04 4.354e-04 -0.234  0.8152    
## POP_dioxin3 -1.942e-05 3.003e-05 -0.647  0.5179    
## POP_furan1  2.488e-03 3.804e-03  0.654  0.5132    
## POP_furan2  -5.409e-04 4.489e-03 -0.120  0.9041    
## POP_furan3  4.497e-03 2.756e-03  1.632  0.1032    
## POP_furan4  -6.516e-04 9.182e-04 -0.710  0.4781    
## whitecell_count -5.516e-03 4.391e-03 -1.256  0.2094    
## lymphocyte_pct -1.382e-03 1.026e-03 -1.347  0.1784    
## monocyte_pct  -5.839e-03 4.039e-03 -1.446  0.1486    
## basophils_pct  1.159e-03 3.462e-03  0.335  0.7377    
## neutrophils_pct 1.127e-02 1.661e-02  0.678  0.4978    
## BMI          -1.358e-03 1.409e-03 -0.964  0.3354    
## edu_cat2     2.226e-02 2.211e-02  1.007  0.3143    
## edu_cat3     4.353e-02 2.145e-02  2.030  0.0427 *  
## edu_cat4     2.960e-02 2.542e-02  1.164  0.2447    
## race_cat2    -2.335e-02 3.263e-02 -0.716  0.4744    
## race_cat3    2.489e-02 3.363e-02  0.740  0.4596    
## race_cat4    -3.529e-02 2.990e-02 -1.181  0.2381    
## male1        -3.992e-02 1.771e-02 -2.254  0.0245 *  
## ageyrs       -6.298e-03 7.431e-04 -8.476 <2e-16 *** 
## yrssmoke     -4.629e-04 7.236e-04 -0.640  0.5225    
## smokenow1    7.390e-04 3.580e-02  0.021  0.9835    
## ln_lbxcot    5.387e-03 3.922e-03  1.374  0.1699    
## ---    
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##

```

```

## Residual standard error: 0.222 on 830 degrees of freedom
## Multiple R-squared:  0.2427, Adjusted R-squared:  0.2126
## F-statistic: 8.062 on 33 and 830 DF,  p-value: < 2.2e-16

#show the VIF
vif(model)

##                               GVIF Df GVIF^(1/(2*Df))
## POP_PCB2            22.564616  1      4.750223
## POP_PCB3             9.066898  1      3.011129
## POP_PCB4            19.767850  1      4.446105
## POP_PCB6            10.137933  1      3.184012
## POP_PCB7             4.820891  1      2.195653
## POP_PCB8             9.395135  1      3.065148
## POP_PCB9            10.038785  1      3.168404
## POP_PCB10            5.715628  1      2.390738
## POP_PCB11            4.226036  1      2.055733
## POP_dioxin1          5.253381  1      2.292026
## POP_dioxin2          5.311508  1      2.304671
## POP_dioxin3          4.332003  1      2.081346
## POP_furan1           6.023639  1      2.454310
## POP_furan2           6.161696  1      2.482276
## POP_furan3           4.449999  1      2.109502
## POP_furan4           1.816107  1      1.347630
## whitecell_count       1.536123  1      1.239404
## lymphocyte_pct        1.372891  1      1.171704
## monocyte_pct          1.261930  1      1.123356
## basophils_pct         1.104229  1      1.050823
## neutrophils_pct       1.088688  1      1.043402
## BMI                  1.261809  1      1.123303
## edu_cat              1.504890  3      1.070494
## race_cat              2.039232  3      1.126102
## male                 1.349779  1      1.161800
## ageyrs                3.227727  1      1.796588
## yrssmoke              2.182035  1      1.477171
## smokenow              3.995671  1      1.998918
## ln_lbxcot              3.953587  1      1.988363

#remove highest vif eosinophils_pct ~ 15047 then POP_PCB5 ~ 59.718357 then POP_PCB1 ~ 32.255513 then POP_PCB2 ~ 22.564616
pollutants$POP_PCB2 = NULL
#fit new model
model = lm(length ~ ., data = pollutants)
#summary
summary(model)

##
## Call:
## lm(formula = length ~ ., data = pollutants)
##
## Residuals:
##      Min      1Q      Median      3Q      Max 
## -0.49387 -0.15104 -0.02904  0.12181  1.18897 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.502e+00 8.354e-02 17.985   <2e-16 ***

```

```

## POP_PCB3      1.846e-06  2.062e-06  0.895  0.3709
## POP_PCB4     -4.363e-07  6.872e-07 -0.635  0.5256
## POP_PCB6      3.180e-07  9.628e-07  0.330  0.7413
## POP_PCB7     -7.809e-07  1.183e-06 -0.660  0.5094
## POP_PCB8     -8.255e-07  1.648e-06 -0.501  0.6166
## POP_PCB9     -3.387e-07  1.787e-06 -0.190  0.8497
## POP_PCB10    9.249e-04   8.414e-04  1.099  0.2720
## POP_PCB11    9.491e-05   2.909e-04  0.326  0.7444
## POP_dioxin1   1.933e-05  3.028e-04  0.064  0.9491
## POP_dioxin2   -1.104e-04  4.349e-04 -0.254  0.7997
## POP_dioxin3   -1.856e-05  2.997e-05 -0.619  0.5359
## POP_furan1    2.570e-03   3.799e-03  0.677  0.4989
## POP_furan2    -6.152e-04  4.485e-03 -0.137  0.8909
## POP_furan3    4.443e-03   2.753e-03  1.614  0.1069
## POP_furan4    -6.314e-04  9.170e-04 -0.689  0.4913
## whitecell_count -5.469e-03  4.388e-03 -1.246  0.2130
## lymphocyte_pct  -1.364e-03  1.025e-03 -1.330  0.1838
## monocyte_pct   -5.733e-03  4.032e-03 -1.422  0.1554
## basophils_pct   1.182e-03   3.460e-03  0.342  0.7326
## neutrophils_pct 1.103e-02   1.660e-02  0.664  0.5066
## BMI            -1.332e-03  1.408e-03 -0.946  0.3445
## edu_cat2       2.268e-02   2.208e-02  1.027  0.3046
## edu_cat3       4.375e-02   2.143e-02  2.041  0.0416 *
## edu_cat4       2.929e-02   2.541e-02  1.153  0.2494
## race_cat2      -2.311e-02  3.262e-02 -0.709  0.4788
## race_cat3      2.558e-02   3.359e-02  0.761  0.4466
## race_cat4      -3.492e-02  2.987e-02 -1.169  0.2428
## male1          -3.975e-02  1.770e-02 -2.246  0.0250 *
## ageyrs          -6.308e-03  7.425e-04 -8.495 <2e-16 ***
## yrssmoke       -4.532e-04  7.230e-04 -0.627  0.5310
## smokenow1      -5.910e-04  3.569e-02 -0.017  0.9868
## ln_lbxcot      5.473e-03   3.917e-03  1.397  0.1626
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2219 on 831 degrees of freedom
## Multiple R-squared:  0.2425, Adjusted R-squared:  0.2133
## F-statistic: 8.313 on 32 and 831 DF,  p-value: < 2.2e-16

#show the VIF
vif(model)

```

```

##                               GVIF Df GVIF^(1/(2*Df))
## POP_PCB3      8.558618  1      2.925512
## POP_PCB4     14.260201  1      3.776268
## POP_PCB6      9.430583  1      3.070925
## POP_PCB7      4.689281  1      2.165475
## POP_PCB8      5.899020  1      2.428790
## POP_PCB9      8.900140  1      2.983310
## POP_PCB10     5.368201  1      2.316938
## POP_PCB11     4.225811  1      2.055678
## POP_dioxin1   5.187456  1      2.277599
## POP_dioxin2   5.303754  1      2.302988
## POP_dioxin3   4.318541  1      2.078110
## POP_furan1    6.013167  1      2.452176

```

```

## POP_furan2      6.155357  1      2.480999
## POP_furan3      4.443638  1      2.107994
## POP_furan4      1.812808  1      1.346405
## whitecell_count 1.535444  1      1.239130
## lymphocyte_pct   1.371183  1      1.170975
## monocyte_pct     1.258625  1      1.121885
## basophils_pct    1.104048  1      1.050737
## neutrophils_pct  1.087848  1      1.042999
## BMI             1.260084  1      1.122535
## edu_cat          1.499070  3      1.069803
## race_cat          2.035520  3      1.125760
## male              1.349340  1      1.161611
## ageyrs            3.225616  1      1.796000
## yrssmoke          2.180551  1      1.476669
## smokenow          3.974936  1      1.993724
## ln_lbxcot         3.946444  1      1.986566

#remove highest vif eosinophils_pct ~ 15047 then POP_PCB5 ~ 59.718357 then POP_PCB1 ~ 32.255513 then PO
# then POP_PCB4 ~ 14.260201
pollutants$POP_PCB4 = NULL
#fit new model
model = lm(length ~ ., data = pollutants)
#summary
summary(model)

##
## Call:
## lm(formula = length ~ ., data = pollutants)
##
## Residuals:
##       Min     1Q     Median     3Q     Max 
## -0.49174 -0.15062 -0.02829  0.12117  1.19031 
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.505e+00 8.339e-02 18.048 <2e-16 ***
## POP_PCB3    1.039e-06 1.624e-06  0.640  0.5223    
## POP_PCB6    2.008e-07 9.446e-07  0.213  0.8317    
## POP_PCB7   -7.992e-07 1.182e-06 -0.676  0.4993    
## POP_PCB8   -7.951e-07 1.647e-06 -0.483  0.6294    
## POP_PCB9   -7.655e-07 1.655e-06 -0.463  0.6438    
## POP_PCB10   8.170e-04 8.238e-04  0.992  0.3216    
## POP_PCB11   8.365e-05 2.903e-04  0.288  0.7733    
## POP_dioxin1 2.403e-05 3.026e-04  0.079  0.9367    
## POP_dioxin2 -1.306e-04 4.336e-04 -0.301  0.7633    
## POP_dioxin3 -1.933e-05 2.993e-05 -0.646  0.5185    
## POP_furan1   2.682e-03 3.793e-03  0.707  0.4797    
## POP_furan2  -5.840e-04 4.483e-03 -0.130  0.8964    
## POP_furan3   4.589e-03 2.742e-03  1.673  0.0946 .  
## POP_furan4  -6.331e-04 9.166e-04 -0.691  0.4900    
## whitecell_count -5.564e-03 4.384e-03 -1.269  0.2047    
## lymphocyte_pct  -1.355e-03 1.025e-03 -1.323  0.1863    
## monocyte_pct   -5.859e-03 4.025e-03 -1.456  0.1459    
## basophils_pct   1.356e-03 3.448e-03  0.393  0.6941    
## neutrophils_pct 1.038e-02 1.656e-02  0.627  0.5311

```

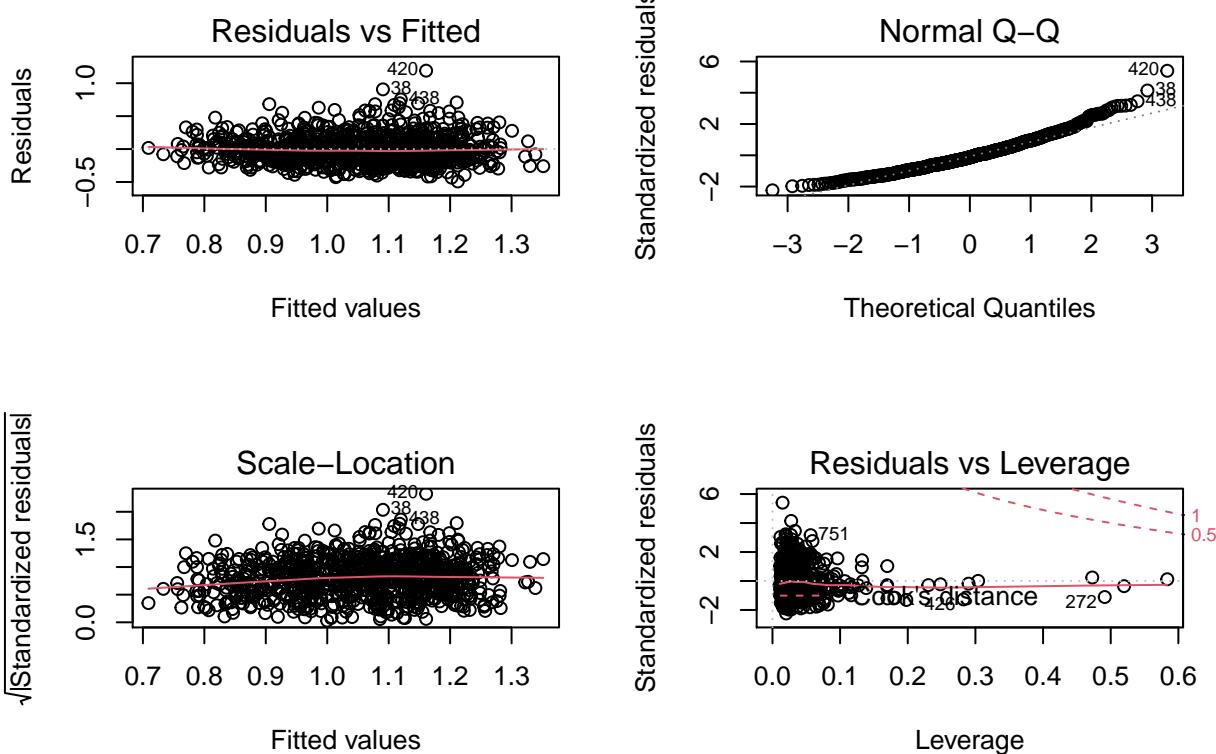
```

## BMI           -1.292e-03 1.406e-03 -0.919  0.3585
## edu_cat2    2.258e-02 2.207e-02  1.023  0.3066
## edu_cat3    4.383e-02 2.143e-02  2.046  0.0411 *
## edu_cat4    2.954e-02 2.539e-02  1.163  0.2450
## race_cat2   -2.332e-02 3.260e-02 -0.715  0.4746
## race_cat3   2.375e-02 3.346e-02  0.710  0.4780
## race_cat4   -3.549e-02 2.985e-02 -1.189  0.2348
## male1        -4.034e-02 1.767e-02 -2.282  0.0227 *
## ageyrs       -6.317e-03 7.421e-04 -8.512 <2e-16 ***
## yrssmoke    -5.096e-04 7.173e-04 -0.710  0.4776
## smokenow1   4.148e-04 3.564e-02  0.012  0.9907
## ln_lbxcot   5.455e-03 3.915e-03  1.393  0.1639
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2219 on 832 degrees of freedom
## Multiple R-squared: 0.2421, Adjusted R-squared: 0.2139
## F-statistic: 8.574 on 31 and 832 DF, p-value: < 2.2e-16
#show the VIF
vif(model)

##          GVIF Df GVIF^(1/(2*Df))
## POP_PCB3      5.310340 1     2.304417
## POP_PCB6      9.083828 1     3.013939
## POP_PCB7      4.686485 1     2.164829
## POP_PCB8      5.894052 1     2.427767
## POP_PCB9      7.640480 1     2.764142
## POP_PCB10     5.149483 1     2.269247
## POP_PCB11     4.210120 1     2.051858
## POP_dioxin1   5.184345 1     2.276916
## POP_dioxin2   5.275271 1     2.296796
## POP_dioxin3   4.311410 1     2.076394
## POP_furan1    6.000097 1     2.449509
## POP_furan2    6.154621 1     2.480851
## POP_furan3    4.412739 1     2.100652
## POP_furan4    1.812793 1     1.346400
## whitecell_count 1.5333642 1     1.238403
## lymphocyte_pct 1.370966 1     1.170882
## monocyte_pct   1.255543 1     1.120510
## basophils_pct  1.097132 1     1.047441
## neutrophils_pct 1.083675 1     1.040997
## BMI            1.257562 1     1.121411
## edu_cat        1.498239 3     1.069704
## race_cat       2.012804 3     1.123657
## male           1.345703 1     1.160045
## ageyrs         3.224432 1     1.795670
## yrssmoke       2.147610 1     1.465473
## smokenow       3.967106 1     1.991759
## ln_lbxcot      3.946223 1     1.986510

#find the model fit for homoscedasticity before remove multicolinearity
par(mfrow=c(2,2))
plot(model)

```



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

# get set a dataset with no categorical covariates
no_cat = pollutants
no_cat$edu_cat = NULL
no_cat$race_cat = NULL
no_cat$male = NULL
no_cat$smokenow = NULL
#summary of the dataset
summary(no_cat)
```

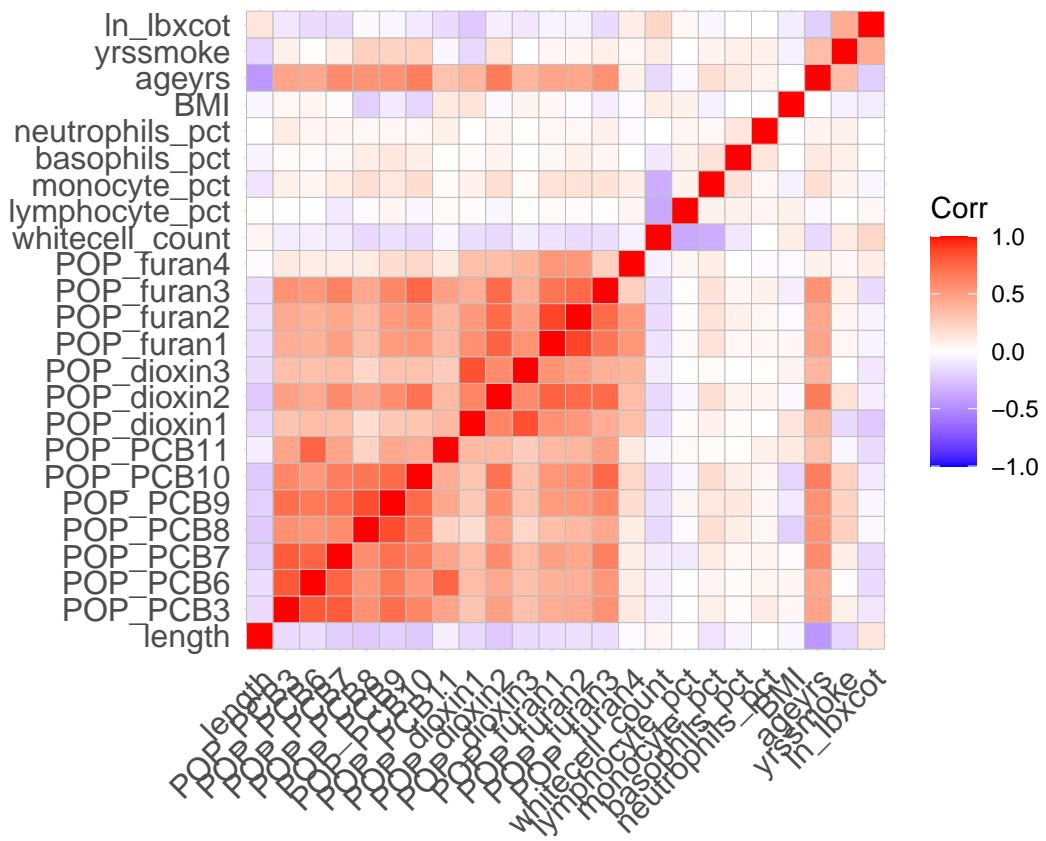
```
##      length      POP_PCB3      POP_PCB6      POP_PCB7
##  Min.   :0.5266   Min.   : 2000   Min.   : 2000   Min.   : 1100
##  1st Qu.:0.8754   1st Qu.: 3700   1st Qu.: 4400   1st Qu.: 4000
##  Median :1.0286   Median : 6200   Median : 9400   Median : 7450
##  Mean   :1.0543   Mean   :10158   Mean   :16820   Mean   :12682
##  3rd Qu.:1.2095   3rd Qu.:12000   3rd Qu.:19500   3rd Qu.:15625
##  Max.   :2.3512   Max.   :123000  Max.   :319000  Max.   :144000
##      POP_PCB8      POP_PCB9      POP_PCB10     POP_PCB11
##  Min.   : 1100   Min.   : 1100   Min.   : 1.70   Min.   : 1.30
##  1st Qu.: 3800   1st Qu.: 3900   1st Qu.: 9.10   1st Qu.: 14.80
##  Median : 6950   Median : 8050   Median : 18.35   Median : 24.50
##  Mean   :10530   Mean   :12220   Mean   :24.49   Mean   :38.15
##  3rd Qu.:14425   3rd Qu.:16025  3rd Qu.:34.90   3rd Qu.:42.95
##  Max.   :187000  Max.   :144000  Max.   :172.00  Max.   :845.00
##      POP_dioxin1    POP_dioxin2    POP_dioxin3    POP_furan1
```

```

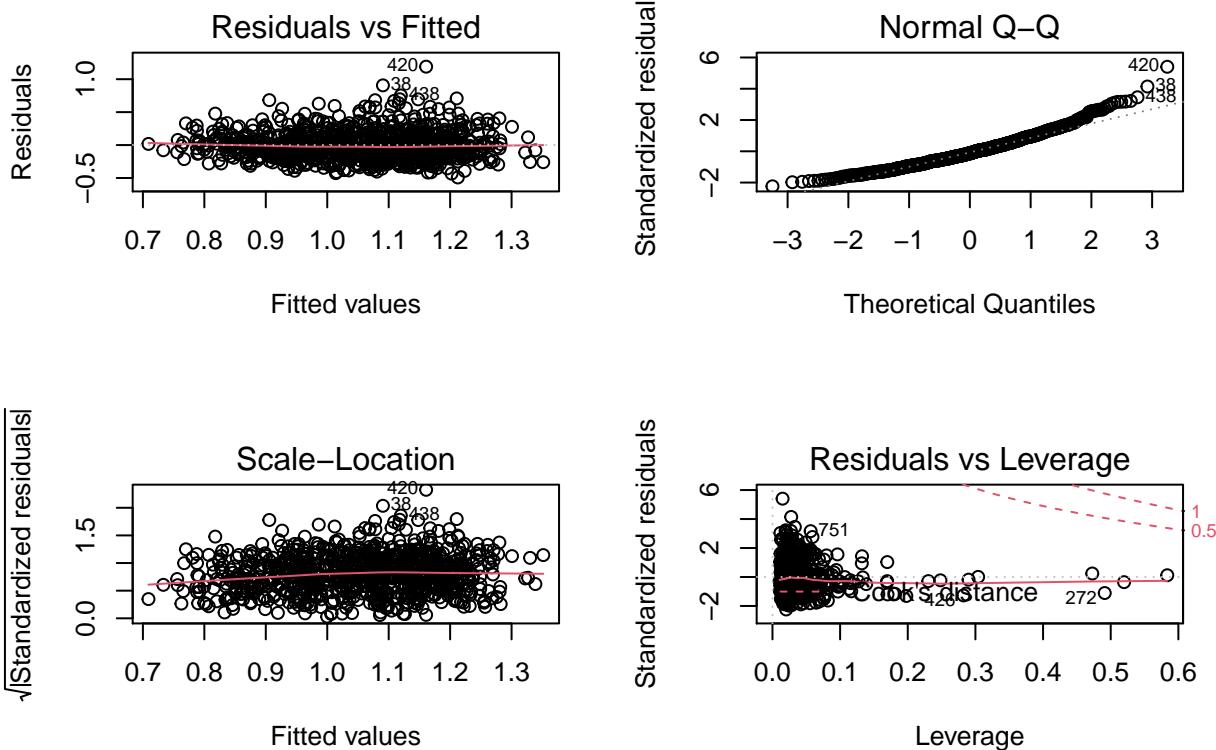
## Min. : 1.90 Min. : 1.40 Min. : 36.8 Min. : 1.000
## 1st Qu.: 23.90 1st Qu.: 21.27 1st Qu.: 197.0 1st Qu.: 3.200
## Median : 41.35 Median : 37.80 Median : 342.5 Median : 5.200
## Mean : 57.65 Mean : 47.81 Mean : 494.4 Mean : 6.371
## 3rd Qu.: 71.62 3rd Qu.: 62.42 3rd Qu.: 603.0 3rd Qu.: 7.700
## Max. :760.00 Max. :281.00 Max. :8190.0 Max. :44.400
## POP_furan2 POP_furan3 POP_furan4 whitecell_count
## Min. : 0.800 Min. : 0.700 Min. : 0.90 Min. : 2.300
## 1st Qu.: 2.600 1st Qu.: 2.200 1st Qu.: 6.40 1st Qu.: 5.600
## Median : 4.200 Median : 5.050 Median : 9.65 Median : 6.900
## Mean : 5.390 Mean : 6.669 Mean : 11.54 Mean : 7.191
## 3rd Qu.: 6.825 3rd Qu.: 9.300 3rd Qu.: 14.00 3rd Qu.: 8.300
## Max. :33.500 Max. :38.300 Max. :234.00 Max. :20.100
## lymphocyte_pct monocyte_pct basophils_pct neutrophils_pct
## Min. : 5.80 Min. : 1.600 Min. : 0.000 Min. : 0.0000
## 1st Qu.:24.00 1st Qu.: 6.600 1st Qu.: 1.500 1st Qu.: 0.4000
## Median :28.95 Median : 7.700 Median : 2.300 Median : 0.6000
## Mean :29.92 Mean : 7.936 Mean : 2.903 Mean : 0.6669
## 3rd Qu.:35.42 3rd Qu.: 9.100 3rd Qu.: 3.700 3rd Qu.: 0.8000
## Max. :73.40 Max. :23.800 Max. :28.200 Max. :5.5000
## BMI ageyrs yrssmoke ln_lbxcot
## Min. :16.16 Min. :20.00 Min. : 0.0 Min. :-4.5099
## 1st Qu.:23.88 1st Qu.:34.00 1st Qu.: 0.0 1st Qu.: -4.0745
## Median :27.38 Median :46.00 Median : 0.0 Median : -2.7334
## Mean :28.09 Mean :48.36 Mean :10.6 Mean : -0.9804
## 3rd Qu.:31.17 3rd Qu.:63.00 3rd Qu.:20.0 3rd Qu.: 2.8000
## Max. :62.99 Max. :85.00 Max. :69.0 Max. : 6.5848

#calculate correlation matrix
corr_matrix = cor(no_cat)
#graph colored corr matrix
ggcorrplot(corr_matrix)

```



```
#find the model fit for homoscedasticity after removing multicollinearity  
par(mfrow=c(2,2))  
plot(model)
```



```

par(mfrow=c(1,1))

#set up train and test model
data = model.matrix(length ~ ., data = pollutants)
n = nrow(data)
#set seed for sample and
set.seed(331)
#get index for random train and test index (90% train, 10% test)
train_row = sample(1:n, 0.9*n)
#train set
x_matrix = data[,-1]
y_matrix = pollutants$length

#get the y values
train_y = y_matrix[train_row]
#get the x values
train_x = x_matrix[train_row,]
#get the y values
test_y = y_matrix[-train_row]
#get the x values
test_x = x_matrix[-train_row,]

eval_results <- function(true, predicted, df){
  SSE <- sum((predicted - true)^2)
  SST <- sum((true - mean(true))^2)
  R_square <- 1 - SSE / SST
}

```

```

RMSE = sqrt(SSE/nrow(df))

# Model performance metrics
data.frame(
  RMSE = RMSE,
  Rsquare = R_square)
}

#k fold lasso
lambdas <- 10^seq(2, -4, by = -.0001)
lasso = cv.glmnet(train_x, train_y, alpha = 1, lambda = lambdas)
best_lam = lasso$lambda.min
best_lam

## [1] 0.008386872

#use lasso with best lambda
best_lasso = glmnet(train_x, train_y, alpha = 1, lambda = best_lam)
#predict with lasso result with training set
pred_train = predict(best_lasso, s = best_lam, newx = train_x)
eval_results(train_y, pred_train, train_x)

##          RMSE      Rsquare
## 1 0.2201948 0.2180372

#predict with lasso result with test set
pred_train = predict(best_lasso, s = best_lam, newx = test_x)
eval_results(test_y, pred_train, test_x)

##          RMSE      Rsquare
## 1 0.2303651 0.2104174

#training data set
train_data = pollutants[train_row,]
#training model
train_model = lm(length ~ ., data = train_data)

#step wise AIC on training data
step_aic = step(train_model, direction = "both", trace = FALSE)
summary(step_aic)

## 
## Call:
## lm(formula = length ~ POP_furan3 + male + ageyrs + ln_lbxcot,
##     data = train_data)
## 
## Residuals:
##       Min     1Q   Median     3Q    Max 
## -0.5045 -0.1510 -0.0268  0.1211  1.1946 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.3701058  0.0232028 59.049 < 2e-16 ***
## POP_furan3  0.0058730  0.0016414  3.578 0.000368 ***  
## male1        -0.0421775  0.0161389 -2.613 0.009139 ** 
## 
```

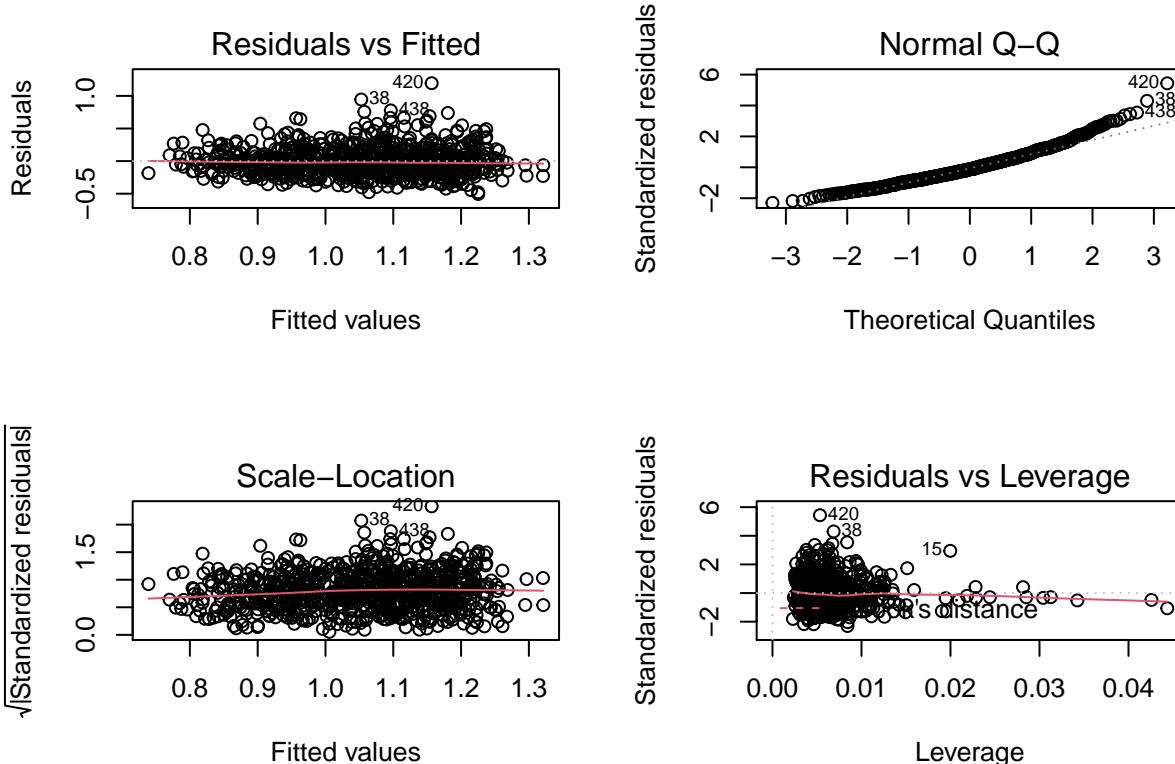
```

## ageyrs      -0.0068654  0.0005261 -13.049 < 2e-16 ***
## ln_lbxcot    0.0034173  0.0021327   1.602 0.109501
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2203 on 772 degrees of freedom
## Multiple R-squared:  0.2223, Adjusted R-squared:  0.2183
## F-statistic: 55.17 on 4 and 772 DF,  p-value: < 2.2e-16
aic_pred = predict(step_aic, newdata = pollutants[-train_row,])

#RMSE AIC
aic_true = pollutants$length[-train_row]
aic_sd = sum((aic_true - aic_pred)^2)
msd_aic = aic_sd / length(aic_true)
rmse_aic = sqrt(msd_aic)
rmse_aic

## [1] 0.2295366
#find the AIC model fit for homoscedasticity after removing multicollinearity
par(mfrow=c(2,2))
plot(step_aic)

```



```

par(mfrow=c(1,1))

#step wise BIC on training data
step_bic = step(train_model, direction = "both", trace = FALSE, k = log(nrow(train_x)))

```

```

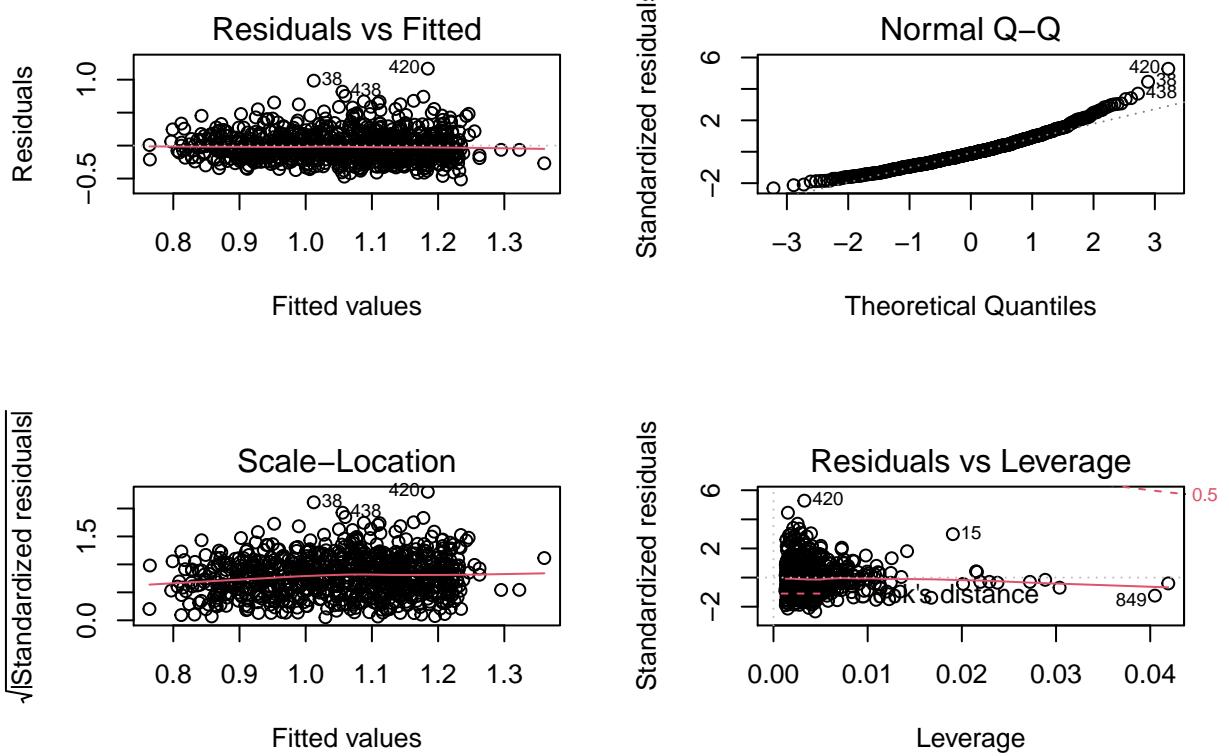
summary(step_bic)

##
## Call:
## lm(formula = length ~ POP_furan3 + ageyrs, data = train_data)
##
## Residuals:
##       Min     1Q Median     3Q    Max
## -0.51364 -0.15642 -0.02609  0.12092  1.16684
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.359962  0.022585 60.215 < 2e-16 ***
## POP_furan3  0.006055  0.001644  3.682 0.000247 ***
## ageyrs      -0.007125  0.000518 -13.755 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2212 on 774 degrees of freedom
## Multiple R-squared:  0.2137, Adjusted R-squared:  0.2116
## F-statistic: 105.2 on 2 and 774 DF,  p-value: < 2.2e-16
bic_pred = predict(step_bic, newdata = pollutants[-train_row,])

#RMSE AIC
bic_true = pollutants$length[-train_row]
bic_sd = sum((bic_true - bic_pred)^2)
msd_bic = bic_sd / length(bic_true)
rmse_bic = sqrt(msd_bic)
rmse_bic

## [1] 0.2298015
#find the BIC model fit for homoscedasticity after removing multicollinearity
par(mfrow=c(2,2))
plot(step_bic)

```



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