Lista 04

Matheus Cougias e Klysman Rezende

27/08/2020

Leitura de arquivo

Realiza a leitura dos dados presentes no arquivo Base_DEA_valores_medios_2014-2016.csv. A função read.csv2 foi utilizada devido ao tipo de separador presente no arquivo e algumas das colunas foram retiradas pois não são de interesse para análise.

```
dados <- read.csv2('Base_DEA_valores_medios_2014-2016.csv')
dados <- dados[4:11]
```

Análise exploratória dos dados

Como início da análise exploratória dos dados, foi feito um histograma para ter ideia de como funciona a distribuição da variável resposta. Pode-se perceber uma tendência de distribuição exponencial nos dados, pois os as colunas do histograma decrescem de tamanho da direção do eixo X positivo. Uma teoria provável para a próxima etapa será aplicar um logaritmo na variável PMSO, para tentar corrigir esse comportamento exponencial da variável.

Para facilitar uma análise das possíveis situações dos dados, foram anexadas à base de dados colunas que representam o logaritmo das variáveis preditoras, e também foi feita uma cópia da base de dados, alterando os valores do PMSO para o logaritmo da mesma. A partir da análise dos gráficos gerados comparando tanto as situações da variável resposta quanto as variáveis preditoras em logaritmo, é perceptivel uma maior linearidade entre os dados quando ambos os lados estão em logaritmo, exceto na variável rsub, onde a maior linearidade dos dados está na situação onde a variável preditora está em escala log e PMSO está na escala original. Por questão de estética, foram deixados somente os gráficos onde a distribuição se mostrou mais padronizada.

```
require(packHV)
```

```
## Loading required package: packHV
```

```
## Loading required package: survival
```

```
require(exploreR)
```

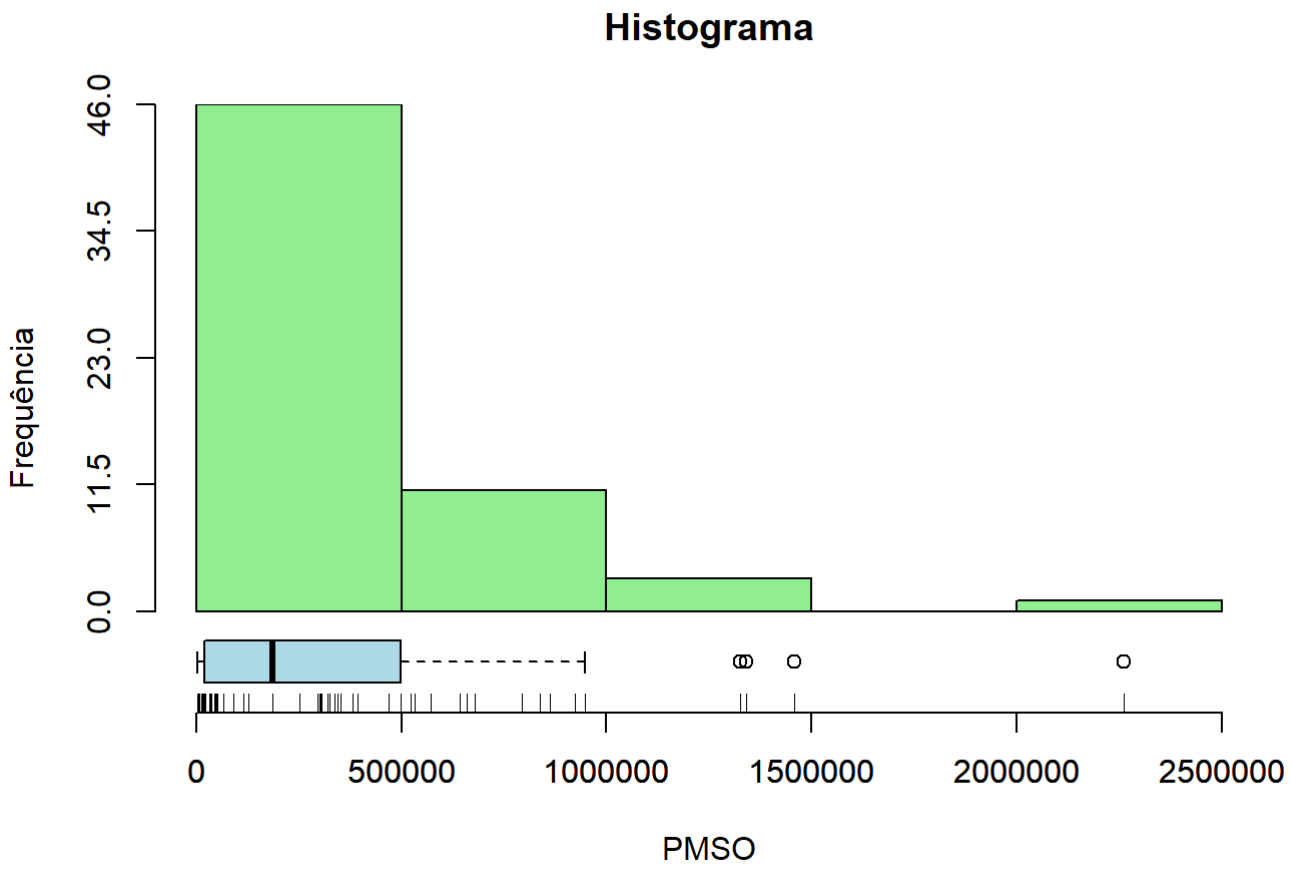
```
## Loading required package: exploreR
```

```
require(corrplot)
```

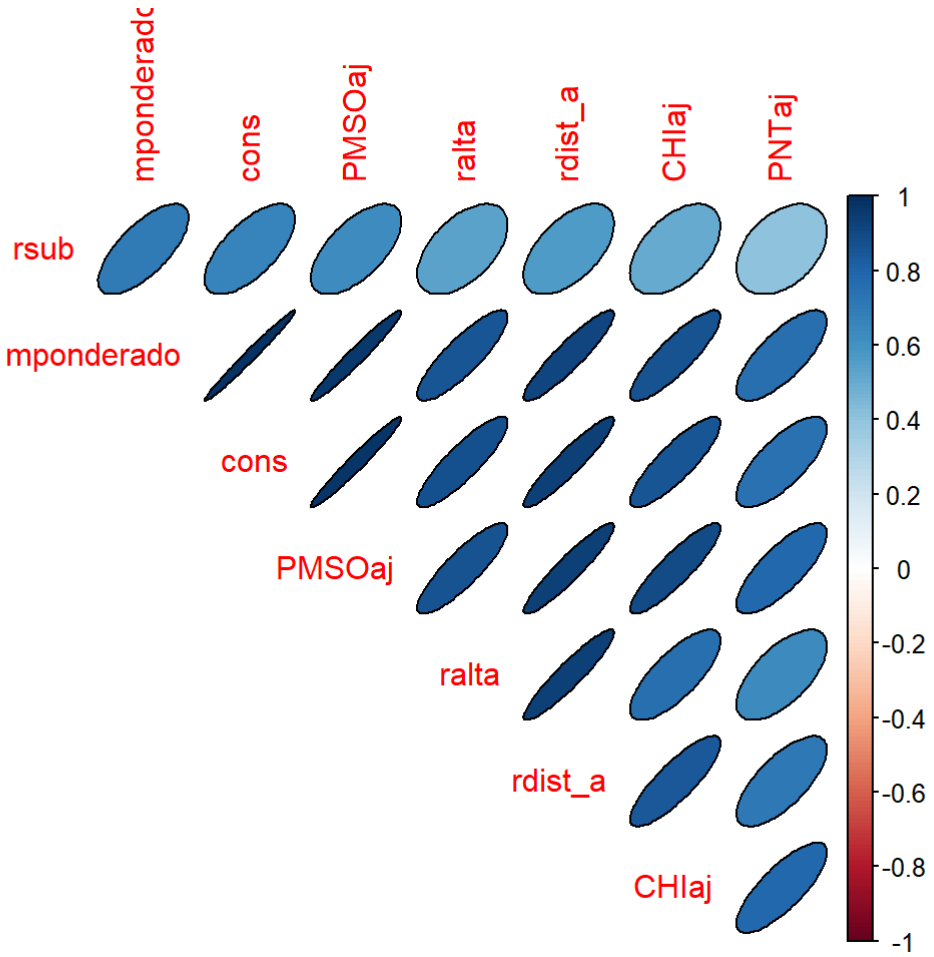
```
## Loading required package: corrplot
```

```
## corrplot 0.84 loaded
```

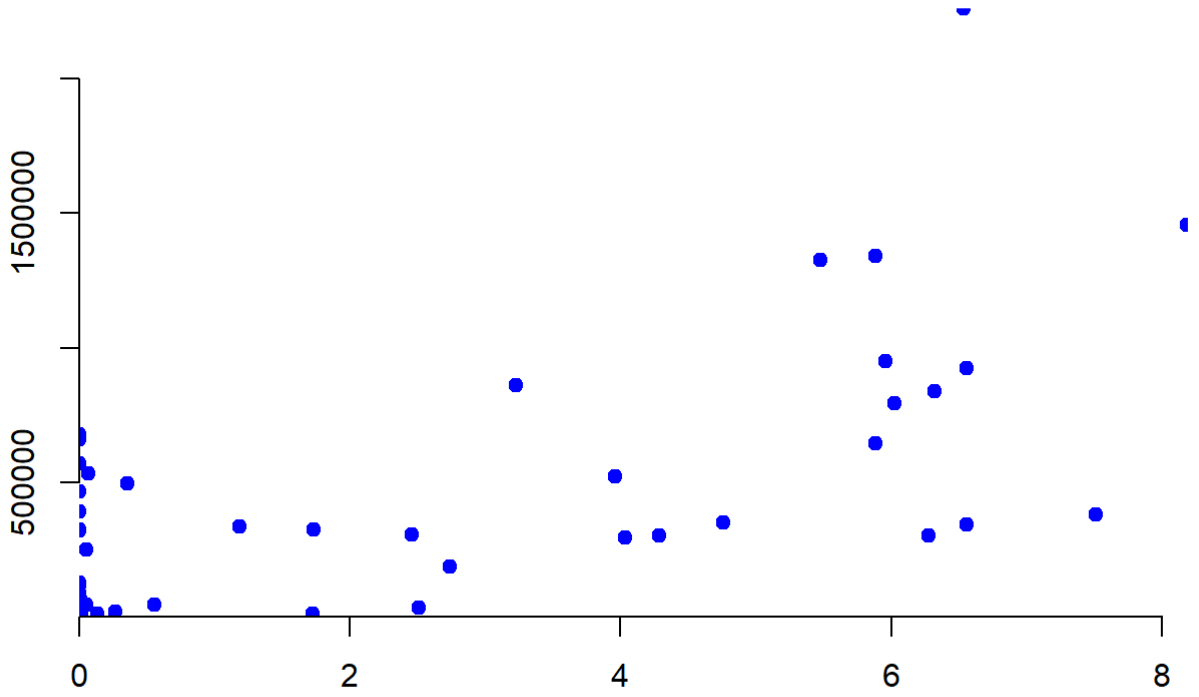
```
hist_boxplot(dados$PMSOaj, main="Histograma", xlab="PMSO", ylab="Frequência", col="light gree
n")
rug(dados$PMSOaj)
```



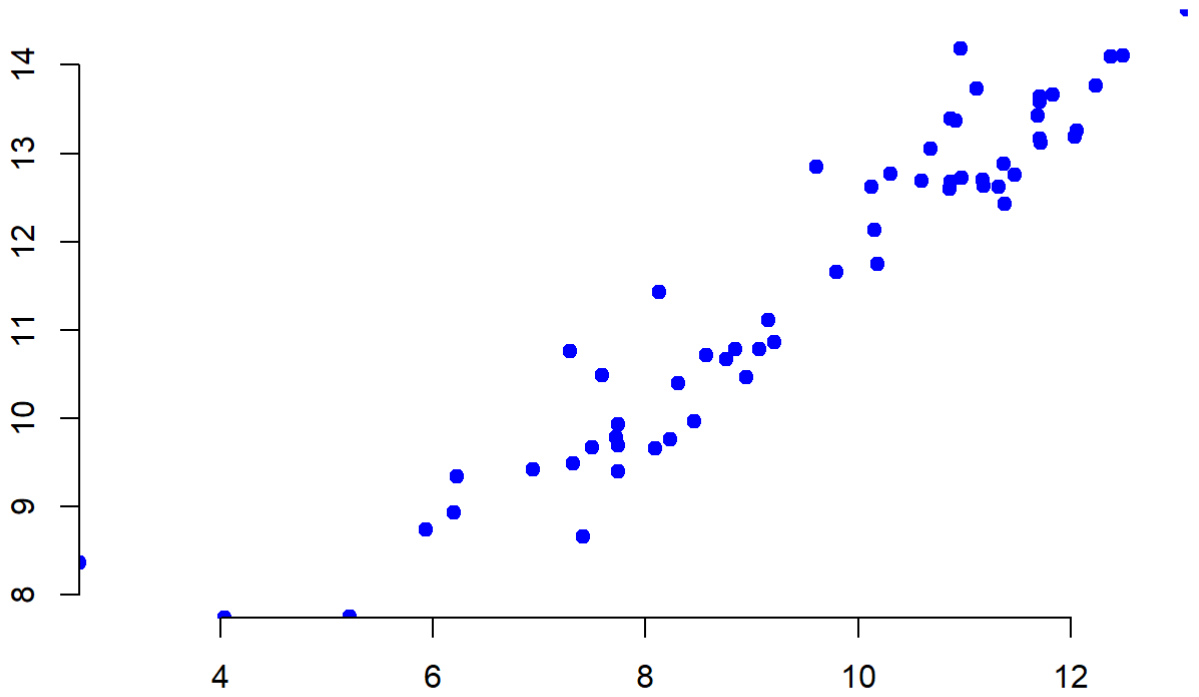

```
corMat <- cor(dados, method="spearman")
corrplot(corMat, method = "ellipse", type="upper", order="AOE",
          diag=FALSE, addgrid.col=NA, outline=TRUE)
```



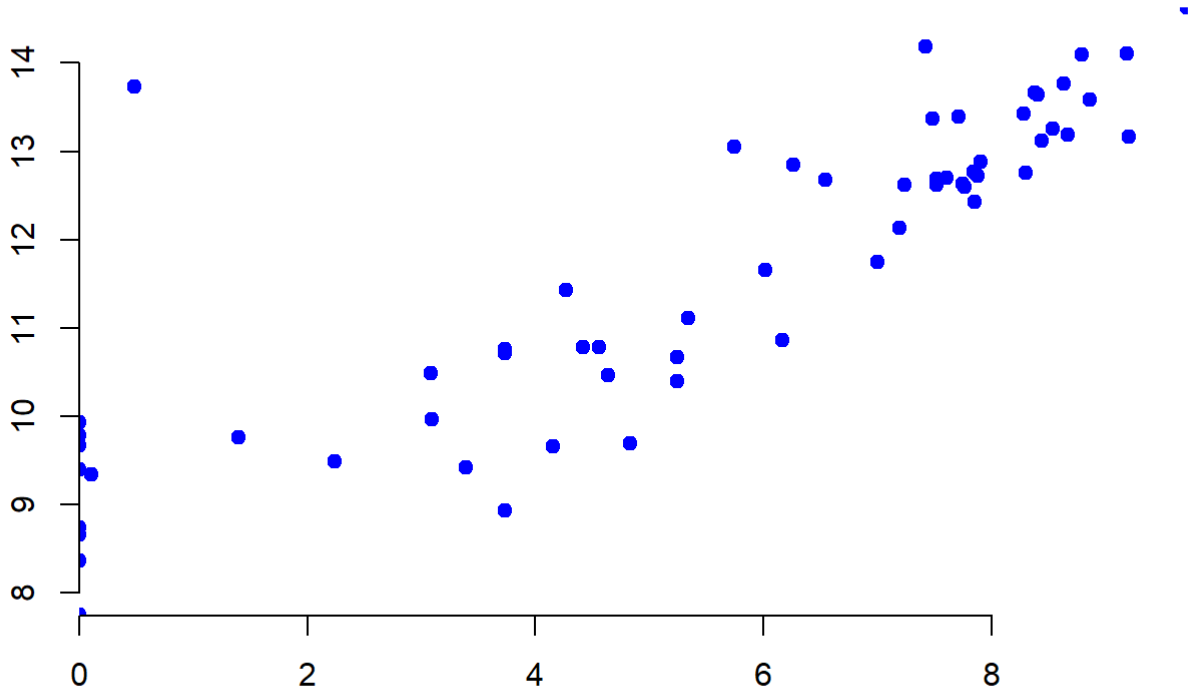

```
dados$logrsub    <- log(dados$rsub+1)
dados$logrdist_a    <- log(dados$rdist_a)
dados$logralta    <- log(dados$ralta+1)
dados$logmponderado    <- log(dados$mponderado)
dados$logcons    <- log(dados$cons)
dados$logCHIaj    <- log(dados$CHIaj+1)
dados$logPNTaj    <- log(dados$PNTaj+1)

dados2 <- dados[2:15]
dados2$logPMSO <- log(dados$PMSOaj)


plot(PMSOaj ~ logrsub, data=dados, pch=19, col="blue")
```



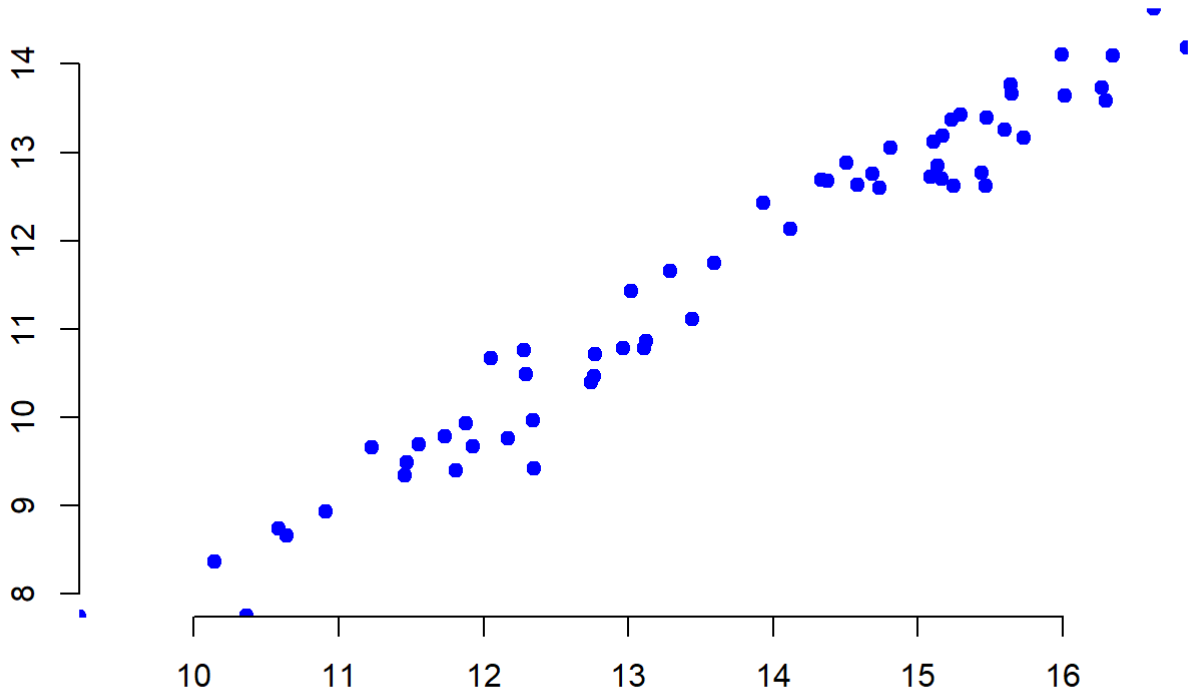

```
plot(logPMSO ~ logrdist_a, data=dados2, pch=19, col="blue")
```



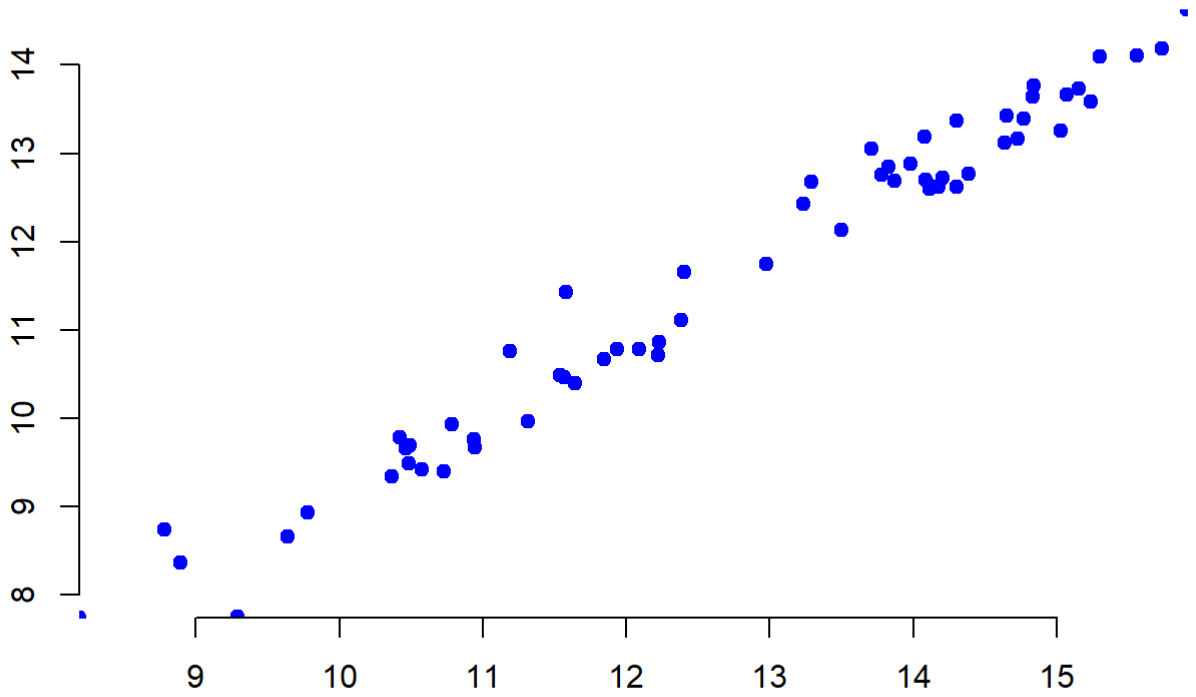
```
plot(logPMSO ~ logralta, data=dados2, pch=19, col="blue")
```



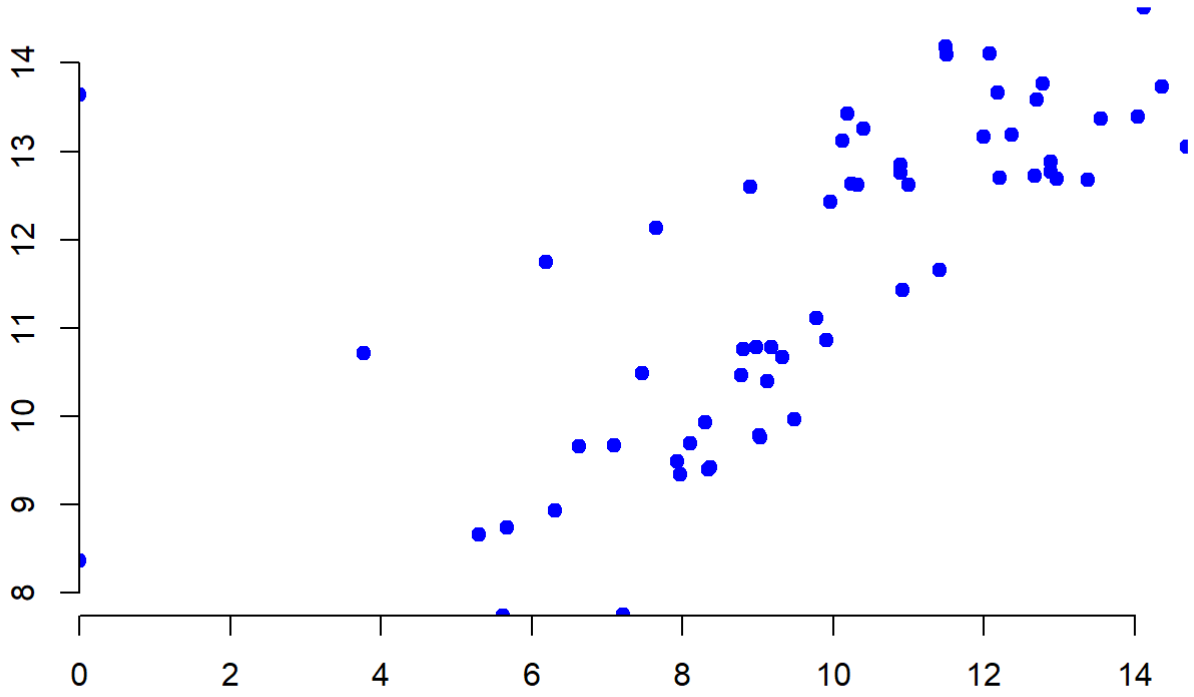

```
plot(logPMSO ~ logmponderado, data=dados2, pch=19, col="blue")
```



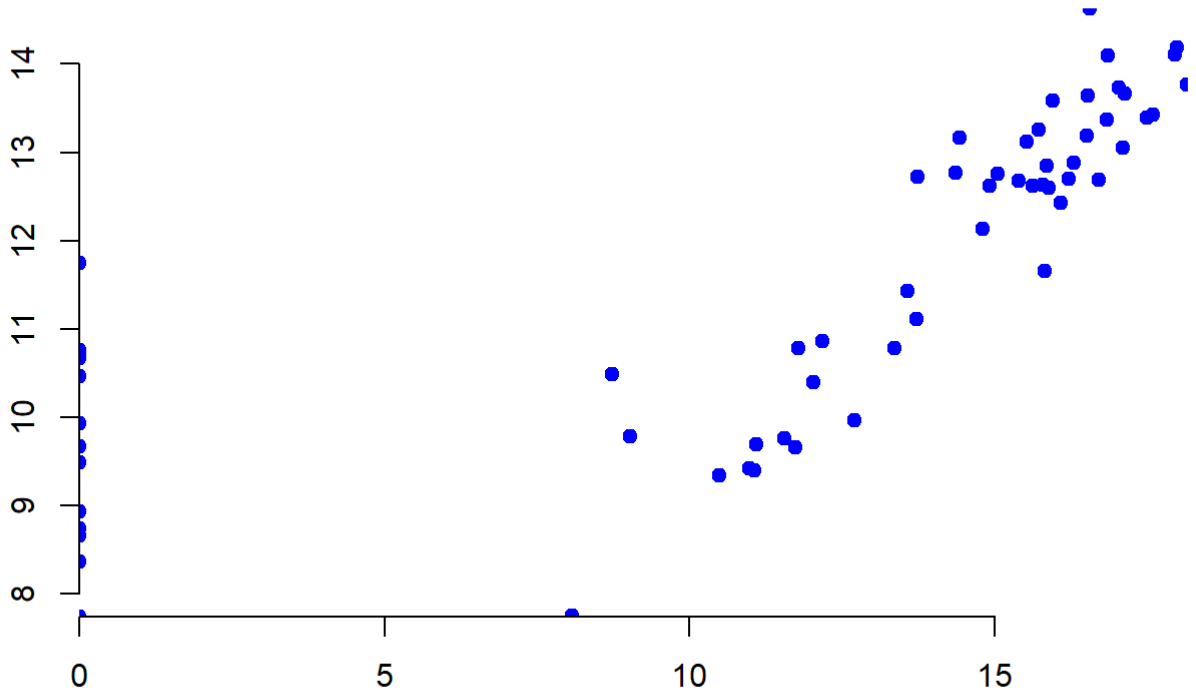

```
plot(logPMSO ~ logcons, data=dados2, pch=19, col="blue")
```



```
plot(logPMSO ~ logPNTaj, data=dados2, pch=19, col="blue")
```




```
plot(logPMSO ~ logCHIaj, data=dados2, pch=19, col="blue")
```



Modelo de regressão linear múltipla

Ainda com a ideia de que tanto a base onde a variável resposta está na base original, quanto na escala log devem ser testadas, foi realizada uma regrassão linear múltipla, no objetivo de identificar quais realmente seriam as variáveis que melhor representam os dados originais. Quanto comparados os resíduos gerados pelas bases, a que o PMSO está em escala original, os resíduos geram uma distribuição totalmente confusa, onde existem diversos pontos fora da normal tanto em valores pequenos quanto em valores mais altos no gráfico.

Já no caso da regressão linear múltipla aplicada sobre a base de dados com log de PMSO, os resultados se mostraram mais satisfatórios, mantendo o nível de R^2 próximo de 0,99 e normalizando um pouco mais os resíduos gerados e deixando mais padronizada a curva desses resíduos. Dessa maneira, o modelo a ser utilizado tomará como variáveis: logPMSO, rsub, rdist_a, cons, PNTaj, CHIaj, logralta, logmponderado e logcons.

```
require(car)
```

```
## Loading required package: car
```

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

```
modelo <- lm(PMSOaj ~ ., data = dados)
modelo <- step(modelo)
```

```
## Start:  AIC=1352.27
## PMSOaj ~ rsub + rdist_a + ralta + mponderado + cons + PNTaj +
##     CHIaj + logrsub + logrdist_a + logralta + logmponderado +
##     logcons + logCHIaj + logPNTaj
##
##                 Df  Sum of Sq        RSS    AIC
## - logcons        1 6.4785e+07 1.5831e+11 1350.3
## - logrdist_a     1 3.4280e+08 1.5858e+11 1350.4
## - logCHIaj       1 3.8795e+08 1.5863e+11 1350.4
## - logmponderado  1 4.0415e+08 1.5864e+11 1350.4
## - logPNTaj       1 1.7826e+09 1.6002e+11 1351.0
## - logralta       1 3.1900e+09 1.6143e+11 1351.5
## - logrsub        1 5.0475e+09 1.6329e+11 1352.2
## <none>                        1.5824e+11 1352.3
## - ralta          1 1.0349e+10 1.6859e+11 1354.1
## - rsub           1 1.1008e+10 1.6925e+11 1354.4
## - mponderado     1 1.4737e+10 1.7298e+11 1355.7
## - cons           1 2.3743e+10 1.8198e+11 1358.8
## - CHIaj          1 3.1841e+10 1.9008e+11 1361.5
## - PNTaj          1 6.9621e+10 2.2786e+11 1372.5
## - rdist_a        1 1.6890e+11 3.2714e+11 1394.6
##
## Step:  AIC=1350.29
## PMSOaj ~ rsub + rdist_a + ralta + mponderado + cons + PNTaj +
##     CHIaj + logrsub + logrdist_a + logralta + logmponderado +
##     logCHIaj + logPNTaj
##
##                 Df  Sum of Sq        RSS    AIC
## - logrdist_a     1 2.7807e+08 1.5858e+11 1348.4
## - logCHIaj       1 3.4726e+08 1.5865e+11 1348.4
## - logmponderado  1 6.7444e+08 1.5898e+11 1348.5
## - logPNTaj       1 1.9810e+09 1.6029e+11 1349.0
## - logralta       1 3.7364e+09 1.6204e+11 1349.7
## - logrsub        1 4.9842e+09 1.6329e+11 1350.2
## <none>                        1.5831e+11 1350.3
## - ralta          1 1.0559e+10 1.6886e+11 1352.2
## - rsub           1 1.0992e+10 1.6930e+11 1352.4
## - mponderado     1 1.7476e+10 1.7578e+11 1354.7
## - CHIaj          1 3.1878e+10 1.9018e+11 1359.5
## - cons           1 3.5840e+10 1.9415e+11 1360.7
## - PNTaj          1 7.0662e+10 2.2897e+11 1370.8
## - rdist_a        1 1.7756e+11 3.3587e+11 1394.2
##
## Step:  AIC=1348.4
## PMSOaj ~ rsub + rdist_a + ralta + mponderado + cons + PNTaj +
##     CHIaj + logrsub + logralta + logmponderado + logCHIaj + logPNTaj
##
##                 Df  Sum of Sq        RSS    AIC
## - logCHIaj       1 4.0693e+08 1.5899e+11 1346.6
## - logPNTaj       1 2.4642e+09 1.6105e+11 1347.3
## - logmponderado  1 2.9044e+09 1.6149e+11 1347.5
## - logralta       1 3.4979e+09 1.6208e+11 1347.7
## <none>                        1.5858e+11 1348.4
## - logrsub        1 5.8211e+09 1.6440e+11 1348.6
## - ralta          1 1.0306e+10 1.6889e+11 1350.2
## - rsub           1 1.1022e+10 1.6961e+11 1350.5
## - mponderado     1 1.7787e+10 1.7637e+11 1352.9
```

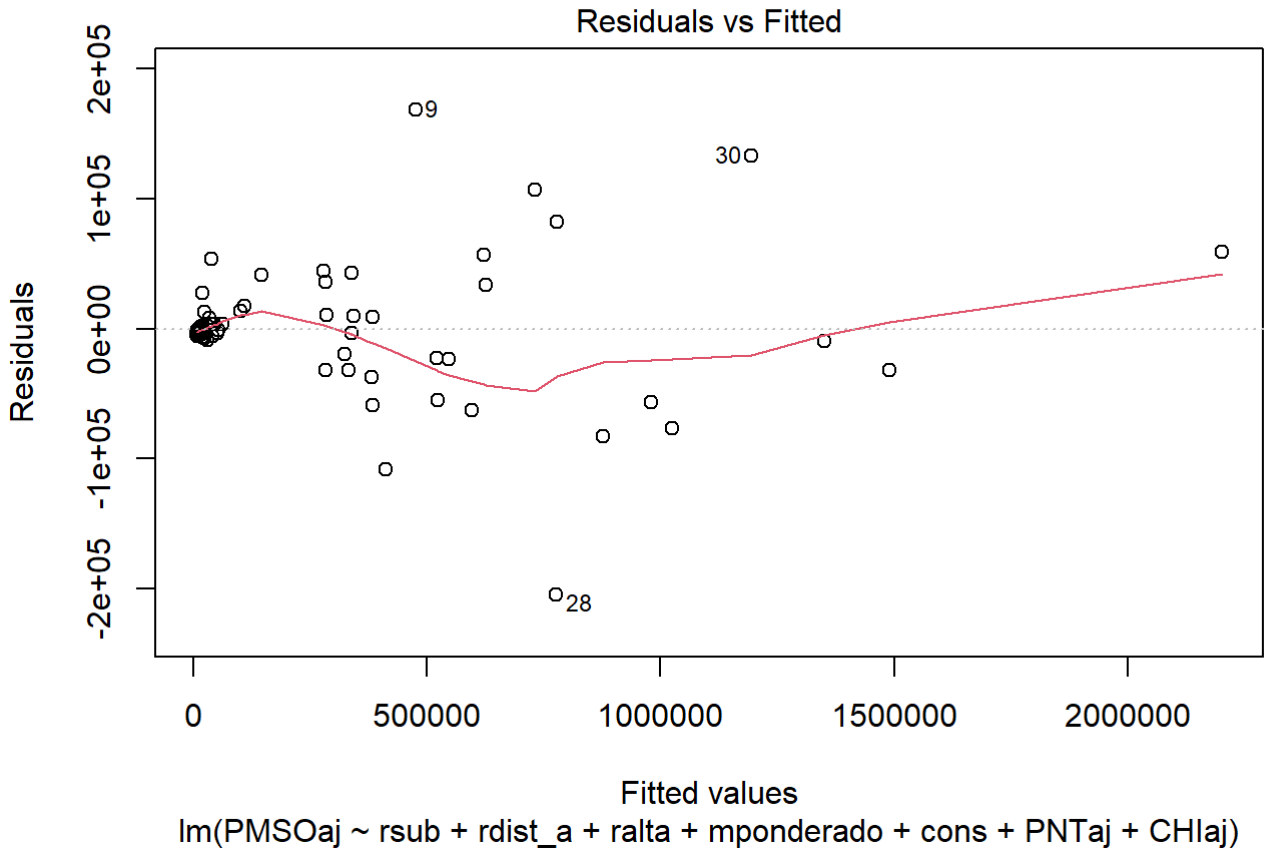
```
## - CHIaj          1 3.1657e+10 1.9024e+11 1357.5
## - cons           1 3.7188e+10 1.9577e+11 1359.2
## - PNTaj          1 7.5315e+10 2.3390e+11 1370.1
## - rdist_a        1 1.9811e+11 3.5669e+11 1395.8
##
## Step:  AIC=1346.56
## PMSOaj ~ rsub + rdist_a + ralta + mponderado + cons + PNTaj +
##     CHIaj + logrsub + logralta + logmponderado + logPNTaj
##
##                 Df  Sum of Sq        RSS    AIC
## - logPNTaj       1 2.1025e+09 1.6109e+11 1345.4
## - logmponderado  1 2.5059e+09 1.6150e+11 1345.5
## - logralta       1 3.6020e+09 1.6259e+11 1345.9
## <none>                        1.5899e+11 1346.6
## - logrsub        1 5.7024e+09 1.6469e+11 1346.7
## - ralta          1 1.0573e+10 1.6956e+11 1348.5
## - rsub           1 1.1132e+10 1.7012e+11 1348.7
## - mponderado     1 1.7846e+10 1.7684e+11 1351.0
## - CHIaj          1 3.3229e+10 1.9222e+11 1356.1
## - cons           1 3.6807e+10 1.9580e+11 1357.3
## - PNTaj          1 7.4908e+10 2.3390e+11 1368.1
## - rdist_a        1 2.0133e+11 3.6032e+11 1394.5
##
## Step:  AIC=1345.36
## PMSOaj ~ rsub + rdist_a + ralta + mponderado + cons + PNTaj +
##     CHIaj + logrsub + logralta + logmponderado
##
##                 Df  Sum of Sq        RSS    AIC
## - logralta       1 3.7148e+09 1.6481e+11 1344.8
## - logmponderado  1 5.0200e+09 1.6611e+11 1345.2
## <none>                        1.6109e+11 1345.4
## - logrsub        1 6.8963e+09 1.6799e+11 1345.9
## - rsub           1 1.0043e+10 1.7114e+11 1347.0
## - ralta          1 1.1458e+10 1.7255e+11 1347.5
## - mponderado     1 2.0970e+10 1.8206e+11 1350.8
## - CHIaj          1 3.2585e+10 1.9368e+11 1354.6
## - cons           1 3.5598e+10 1.9669e+11 1355.5
## - PNTaj          1 7.9753e+10 2.4085e+11 1367.9
## - rdist_a        1 2.0504e+11 3.6614e+11 1393.4
##
## Step:  AIC=1344.75
## PMSOaj ~ rsub + rdist_a + ralta + mponderado + cons + PNTaj +
##     CHIaj + logrsub + logmponderado
##
##                 Df  Sum of Sq        RSS    AIC
## - logmponderado  1 1.3288e+09 1.6614e+11 1343.2
## - logrsub        1 5.1050e+09 1.6991e+11 1344.6
## <none>                        1.6481e+11 1344.8
## - ralta          1 8.4528e+09 1.7326e+11 1345.8
## - rsub           1 1.2604e+10 1.7741e+11 1347.2
## - mponderado     1 1.8405e+10 1.8321e+11 1349.2
## - cons           1 3.4365e+10 1.9917e+11 1354.3
## - CHIaj          1 3.5170e+10 1.9998e+11 1354.5
## - PNTaj          1 7.6860e+10 2.4167e+11 1366.1
## - rdist_a        1 2.0303e+11 3.6784e+11 1391.7
##
## Step:  AIC=1343.24
## PMSOaj ~ rsub + rdist_a + ralta + mponderado + cons + PNTaj +
```

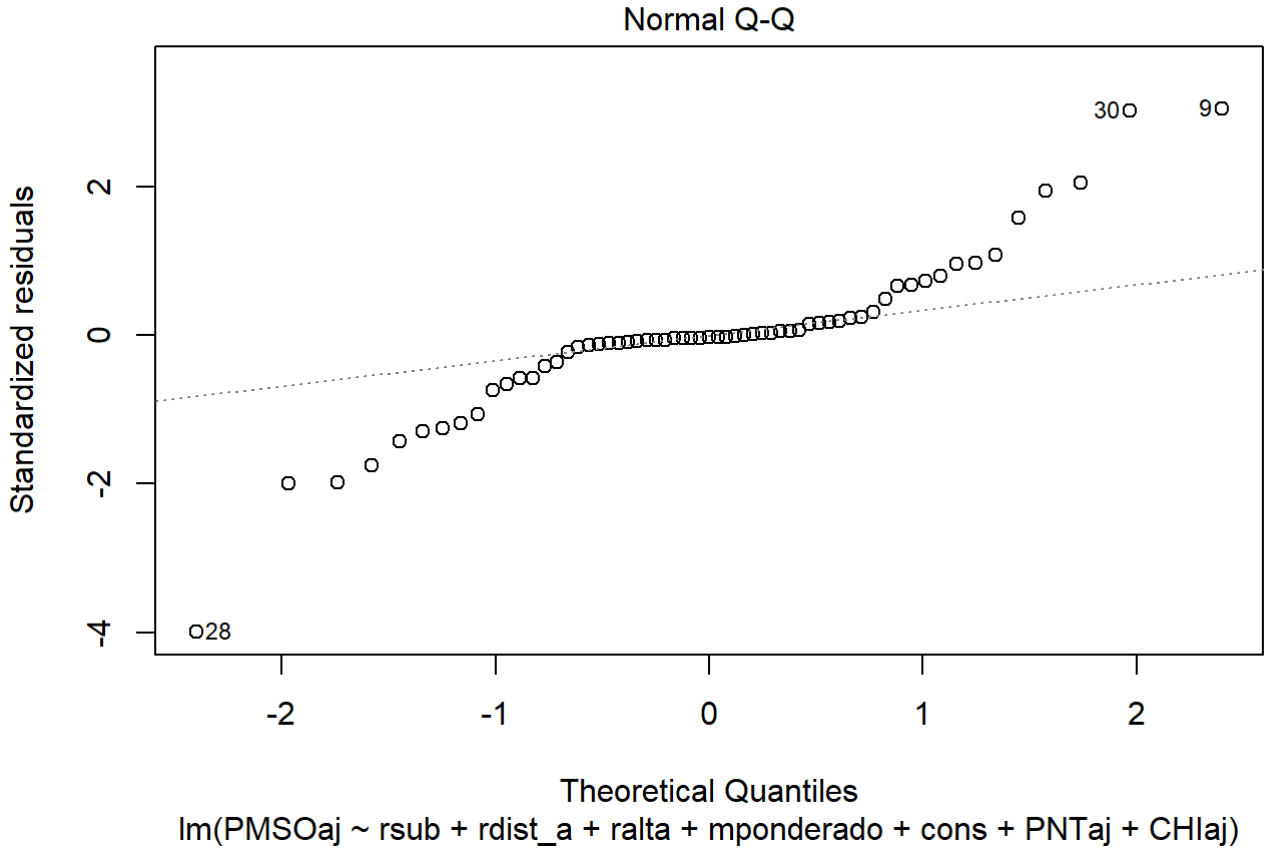
```
##     CHIaj + logrsub
##
##              Df  Sum of Sq        RSS    AIC
## - logrsub     1 3.9686e+09 1.7011e+11 1342.7
## <none>                     1.6614e+11 1343.2
## - ralta       1 9.5182e+09 1.7565e+11 1344.6
## - rsub        1 1.4847e+10 1.8098e+11 1346.5
## - mponderado  1 1.7170e+10 1.8331e+11 1347.2
## - CHIaj       1 3.3878e+10 2.0001e+11 1352.6
## - cons        1 3.5121e+10 2.0126e+11 1352.9
## - PNTaj       1 7.5731e+10 2.4187e+11 1364.2
## - rdist_a     1 2.0492e+11 3.7105e+11 1390.2
##
## Step:  AIC=1342.68
## PMSOaj ~ rsub + rdist_a + ralta + mponderado + cons + PNTaj +
##     CHIaj
##
##              Df  Sum of Sq        RSS    AIC
## <none>                     1.7011e+11 1342.7
## - ralta       1 7.5557e+09 1.7766e+11 1343.3
## - rsub        1 2.2679e+10 1.9278e+11 1348.3
## - mponderado  1 2.3763e+10 1.9387e+11 1348.7
## - cons        1 3.2150e+10 2.0225e+11 1351.2
## - CHIaj       1 3.4872e+10 2.0498e+11 1352.0
## - PNTaj       1 7.5636e+10 2.4574e+11 1363.1
## - rdist_a     1 2.0415e+11 3.7425e+11 1388.8
```

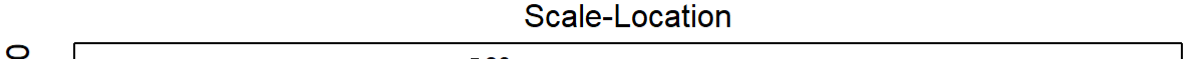
```
summary(modelo)
```

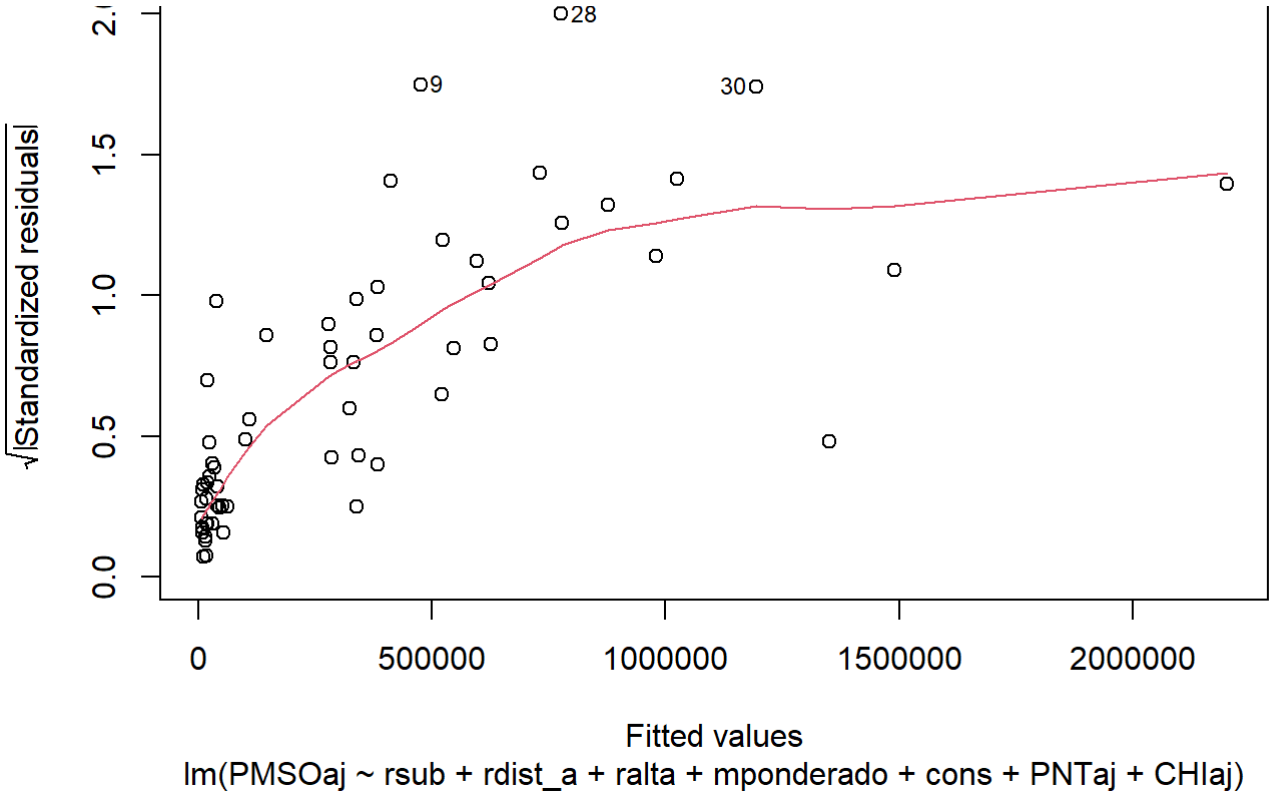
```
##
## Call:
## lm(formula = PMSOaj ~ rsub + rdist_a + ralta + mponderado + cons +
##     PNTaj + CHIaj, data = dados)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -204995   -9505   -1396   12785  168409
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.615e+03  9.285e+03   0.605  0.54790
## rsub         6.736e+01  2.534e+01   2.658  0.01036 *
## rdist_a      2.499e+00  3.134e-01   7.975 1.23e-10 ***
## ralta       -1.215e+01  7.918e+00  -1.534  0.13090
## mponderado   2.186e-02  8.034e-03   2.721  0.00879 **
## cons         7.496e-02  2.368e-02   3.165  0.00257 **
## PNTaj        9.490e-02  1.955e-02   4.854 1.11e-05 ***
## CHIaj        2.178e-03  6.607e-04   3.296  0.00175 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56650 on 53 degrees of freedom
## Multiple R-squared:  0.9855, Adjusted R-squared:  0.9836
## F-statistic: 514.3 on 7 and 53 DF,  p-value: < 2.2e-16
```

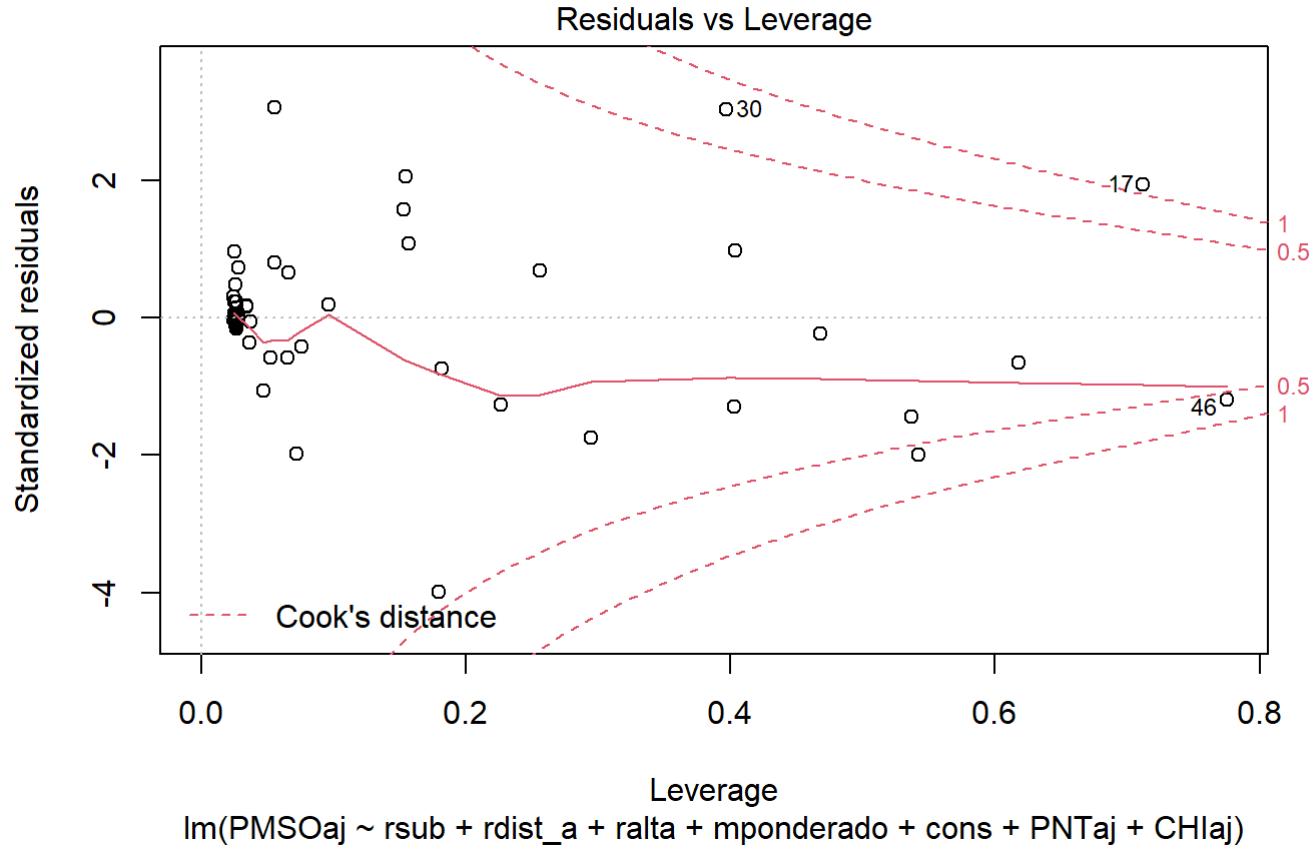
```
plot(modelo)
```


Residuals vs Fitted
Residuals
Fitted values
lm(PMSOaj ~ rsub + rdist_a + ralta + mponderado + cons + PNTaj + CHlaj)
9
30
28
2e+05
1e+05
0e+00
-1e+05
-2e+05
0
500000
1000000
1500000
2000000


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(PMSOaj ~ rsub + rdist_a + ralta + mponderado + cons + PNTaj + CHlaj)
30
9
28
2
0
-2
-4
-2
-1
0
1
2


Scale-Location
0






```
modelo <- lm(logPMSO ~ ., data = dados2)
modelo <- step(modelo)
```

```
## Start:  AIC=-148.12
## logPMSO ~ rsub + rdist_a + ralta + mponderado + cons + PNTaj +
##     CHIaj + logrsub + logrdist_a + logralta + logmponderado +
##     logcons + logCHIaj + logPNTaj
##
##                 Df Sum of Sq    RSS     AIC
## - logPNTaj       1   0.00001 3.2899 -150.12
## - mponderado     1   0.00089 3.2907 -150.10
## - logrdist_a     1   0.01054 3.3004 -149.93
## - logrsub        1   0.02909 3.3189 -149.59
## - ralta          1   0.04039 3.3302 -149.38
## - logCHIaj       1   0.06166 3.3515 -148.99
## - cons           1   0.06538 3.3552 -148.92
## <none>                       3.2899 -148.12
## - rsub           1   0.16310 3.4530 -147.17
## - CHIaj          1   0.22732 3.5172 -146.05
## - rdist_a        1   0.28174 3.5716 -145.11
## - logralta       1   0.28344 3.5733 -145.08
## - logmponderado  1   0.32607 3.6159 -144.36
## - logcons        1   0.37571 3.6656 -143.53
## - PNTaj          1   0.40916 3.6990 -142.97
##
## Step:  AIC=-150.12
## logPMSO ~ rsub + rdist_a + ralta + mponderado + cons + PNTaj +
##     CHIaj + logrsub + logrdist_a + logralta + logmponderado +
##     logcons + logCHIaj
##
##                 Df Sum of Sq    RSS     AIC
## - mponderado     1   0.00099 3.2909 -152.10
## - logrdist_a     1   0.01169 3.3016 -151.91
## - logrsub        1   0.02941 3.3193 -151.58
## - ralta          1   0.04066 3.3305 -151.37
## - logCHIaj       1   0.06525 3.3551 -150.92
## - cons           1   0.06789 3.3578 -150.88
## <none>                       3.2899 -150.12
## - rsub           1   0.16576 3.4556 -149.12
## - CHIaj          1   0.22730 3.5172 -148.05
## - logralta       1   0.28344 3.5733 -147.08
## - rdist_a        1   0.29275 3.5826 -146.92
## - logmponderado  1   0.33520 3.6251 -146.20
## - logcons        1   0.38846 3.6783 -145.31
## - PNTaj          1   0.53085 3.8207 -143.00
##
## Step:  AIC=-152.1
## logPMSO ~ rsub + rdist_a + ralta + cons + PNTaj + CHIaj + logrsub +
##     logrdist_a + logralta + logmponderado + logcons + logCHIaj
##
##                 Df Sum of Sq    RSS     AIC
## - logrdist_a     1   0.01114 3.3020 -153.90
## - logrsub        1   0.03016 3.3210 -153.55
## - ralta          1   0.04002 3.3309 -153.37
## - logCHIaj       1   0.06426 3.3551 -152.92
## <none>                       3.2909 -152.10
## - rsub           1   0.17847 3.4693 -150.88
## - CHIaj          1   0.24514 3.5360 -149.72
## - cons           1   0.26090 3.5518 -149.45
## - logralta       1   0.28442 3.5753 -149.05
```

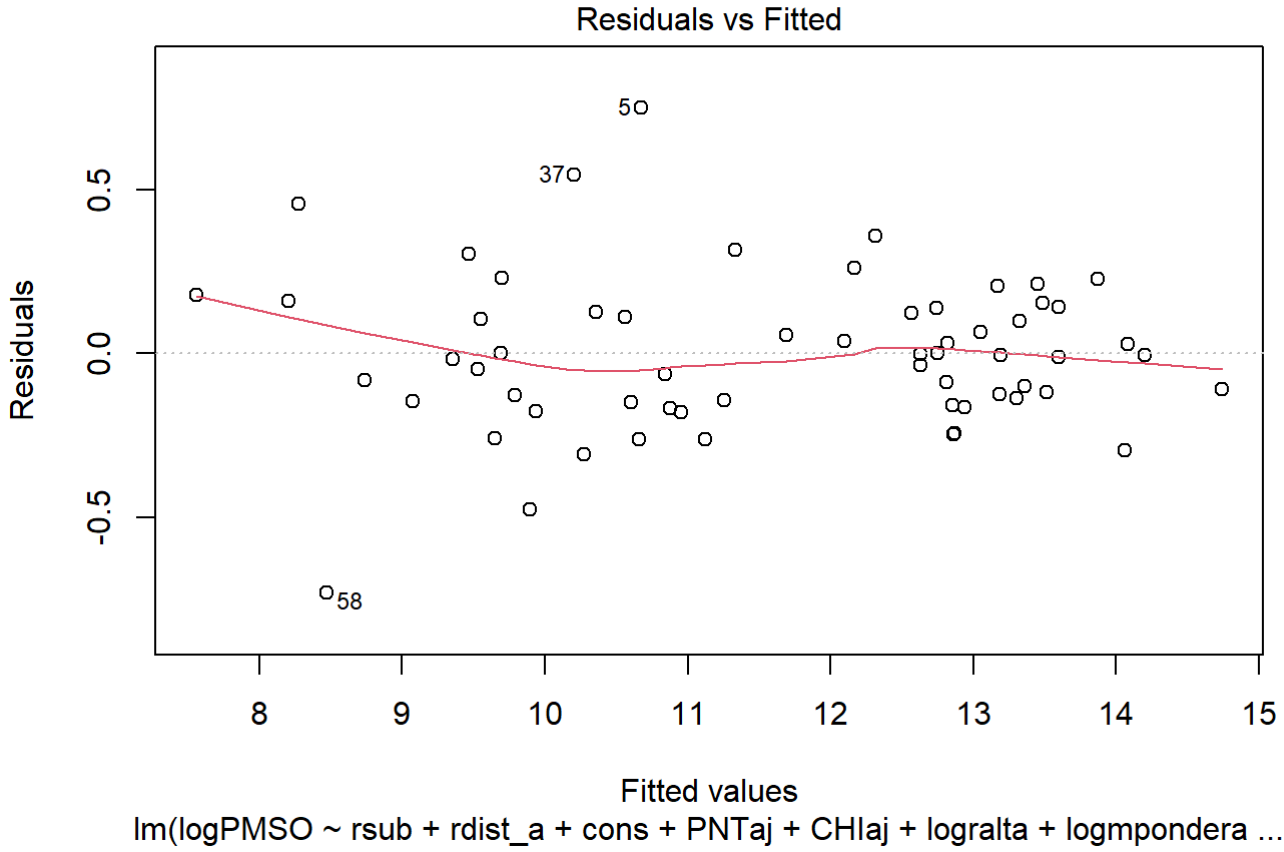
```
## - rdist_a        1   0.31011 3.6010 -148.61
## - logmponderado  1   0.41567 3.7065 -146.85
## - logcons        1   0.49797 3.7888 -145.51
## - PNTaj          1   0.53023 3.8211 -144.99
##
## Step:  AIC=-153.9
## logPMSO ~ rsub + rdist_a + ralta + cons + PNTaj + CHIaj + logrsub +
##     logralta + logmponderado + logcons + logCHIaj
##
##                 Df Sum of Sq    RSS     AIC
## - logrsub        1   0.03868 3.3407 -155.19
## - ralta          1   0.04198 3.3440 -155.13
## - logCHIaj       1   0.06137 3.3634 -154.77
## <none>                       3.3020 -153.90
## - rsub           1   0.18580 3.4878 -152.56
## - CHIaj          1   0.24735 3.5493 -151.49
## - logralta       1   0.28659 3.5886 -150.82
## - cons           1   0.35643 3.6584 -149.64
## - rdist_a        1   0.41427 3.7163 -148.69
## - logmponderado  1   0.42108 3.7231 -148.58
## - PNTaj          1   0.52063 3.8226 -146.97
## - logcons        1   0.68023 3.9822 -144.47
##
## Step:  AIC=-155.19
## logPMSO ~ rsub + rdist_a + ralta + cons + PNTaj + CHIaj + logralta +
##     logmponderado + logcons + logCHIaj
##
##                 Df Sum of Sq    RSS     AIC
## - logCHIaj       1   0.05079 3.3915 -156.27
## - ralta          1   0.05823 3.3989 -156.13
## <none>                       3.3407 -155.19
## - rsub           1   0.14735 3.4880 -154.55
## - CHIaj          1   0.24838 3.5891 -152.81
## - logralta       1   0.32400 3.6647 -151.54
## - cons           1   0.34936 3.6900 -151.12
## - logmponderado  1   0.38398 3.7247 -150.55
## - rdist_a        1   0.41969 3.7604 -149.97
## - PNTaj          1   0.56451 3.9052 -147.66
## - logcons        1   0.79225 4.1329 -144.21
##
## Step:  AIC=-156.27
## logPMSO ~ rsub + rdist_a + ralta + cons + PNTaj + CHIaj + logralta +
##     logmponderado + logcons
##
##                 Df Sum of Sq    RSS     AIC
## - ralta          1   0.05089 3.4424 -157.36
## <none>                       3.3915 -156.27
## - rsub           1   0.15559 3.5471 -155.53
## - CHIaj          1   0.21674 3.6082 -154.49
## - logralta       1   0.29781 3.6893 -153.13
## - cons           1   0.32685 3.7183 -152.65
## - logmponderado  1   0.33507 3.7265 -152.52
## - rdist_a        1   0.39940 3.7909 -151.47
## - PNTaj          1   0.54665 3.9381 -149.15
## - logcons        1   0.87370 4.2652 -144.28
##
## Step:  AIC=-157.36
## logPMSO ~ rsub + rdist_a + cons + PNTaj + CHIaj + logralta +
```

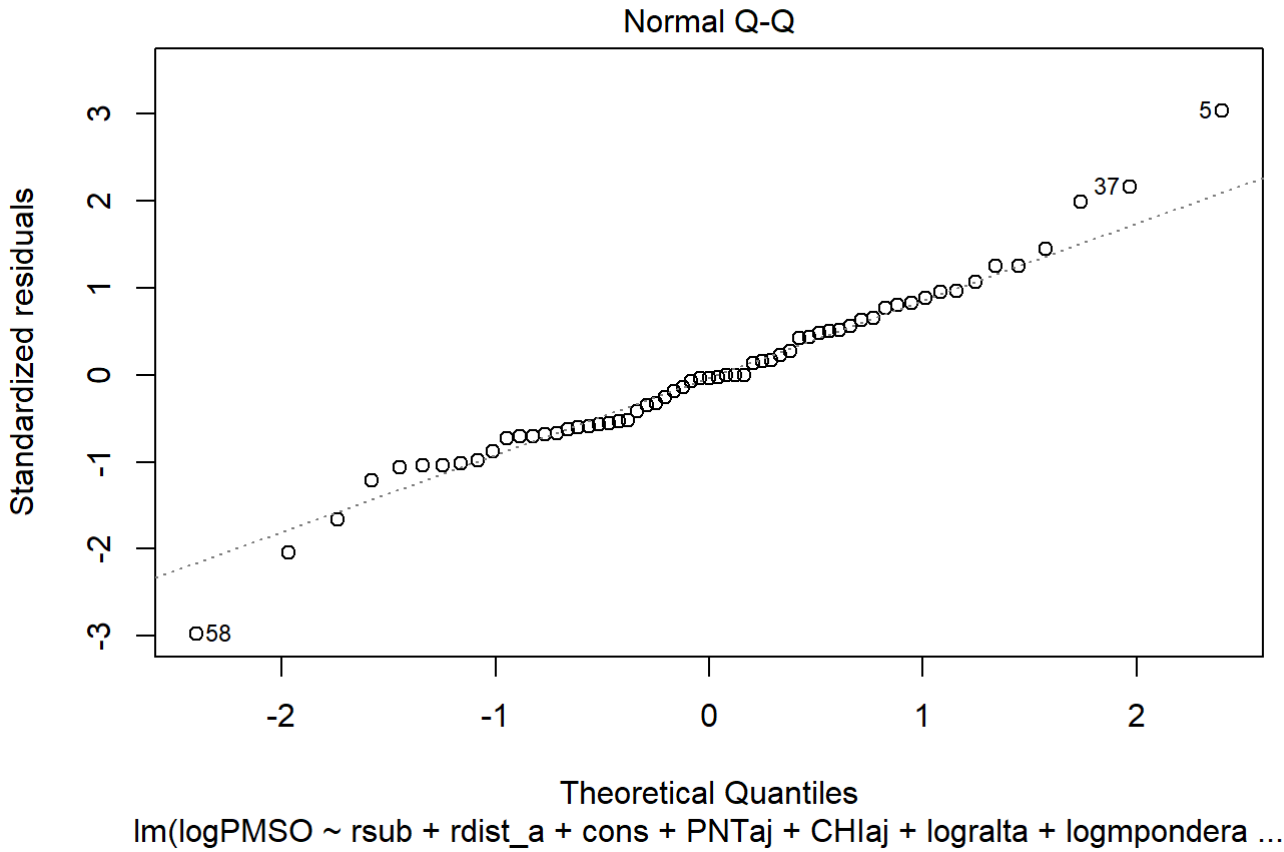
```
##     logmponderado + logcons
##
##                 Df Sum of Sq    RSS     AIC
## <none>                       3.4424 -157.36
## - rsub           1   0.20275 3.6451 -155.87
## - logralta       1   0.24878 3.6911 -155.10
## - logmponderado  1   0.31503 3.7574 -154.02
## - CHIaj          1   0.31598 3.7583 -154.00
## - rdist_a        1   0.45141 3.8938 -151.84
## - cons           1   0.48800 3.9304 -151.27
## - PNTaj          1   0.61626 4.0586 -149.31
## - logcons        1   0.95919 4.4016 -144.36
```

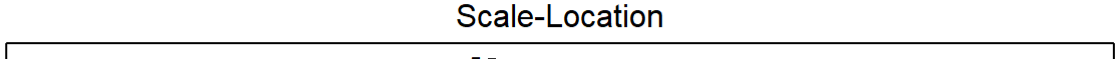
```
summary(modelo)
```

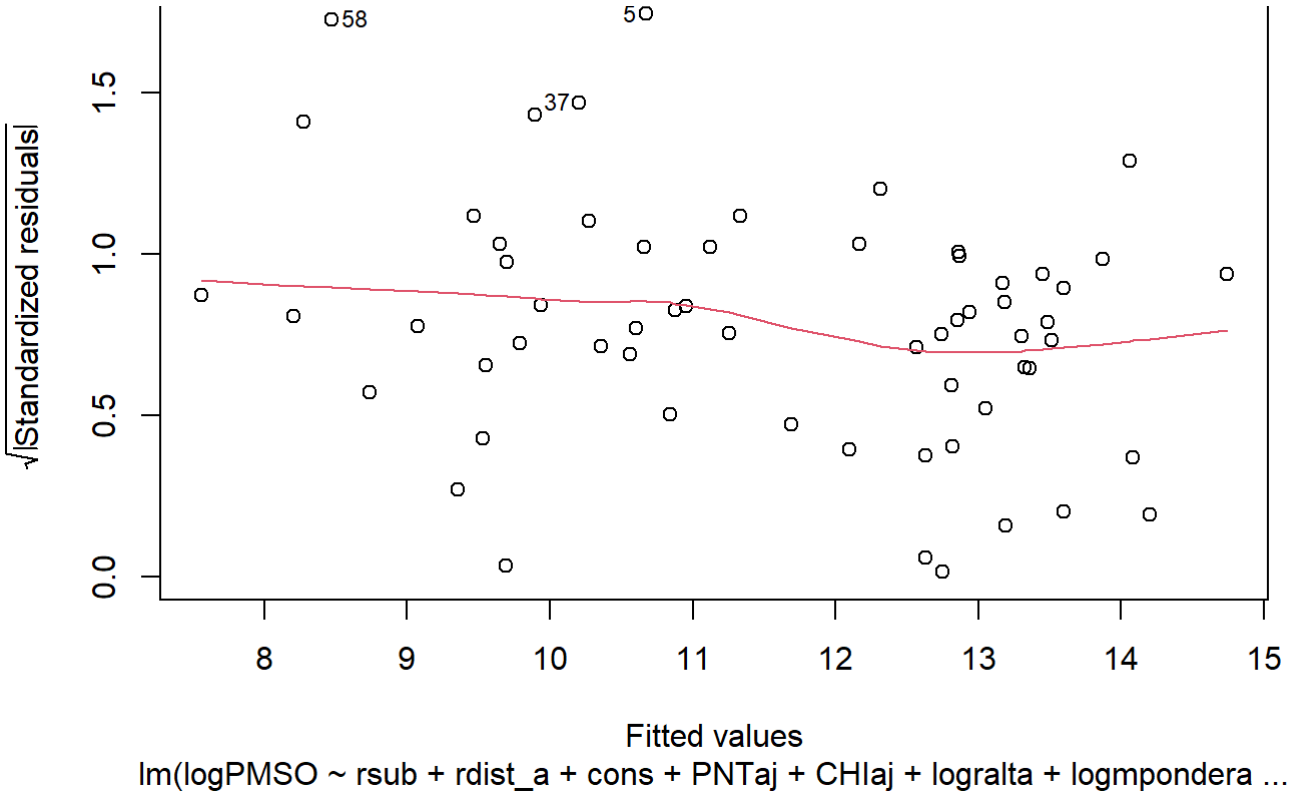
```
##
## Call:
## lm(formula = logPMSO ~ rsub + rdist_a + cons + PNTaj + CHIaj +
##     logralta + logmponderado + logcons, data = dados2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7326 -0.1455 -0.0058  0.1402  0.7516
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.863e-01  5.336e-01   1.099 0.276973
## rsub           1.831e-04  1.046e-04   1.750 0.086009 .
## rdist_a        2.579e-06  9.875e-07   2.611 0.011759 *
## cons          -1.696e-07  6.248e-08  -2.715 0.008971 **
## PNTaj          2.838e-07  9.302e-08   3.051 0.003585 **
## CHIaj          6.258e-09  2.865e-09   2.185 0.033438 *
## logralta       4.978e-02  2.568e-02   1.939 0.057989 .
## logmponderado  2.944e-01  1.349e-01   2.181 0.033693 *
## logcons        5.208e-01  1.368e-01   3.806 0.000373 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2573 on 52 degrees of freedom
## Multiple R-squared:  0.983,  Adjusted R-squared:  0.9804
## F-statistic: 376.9 on 8 and 52 DF,  p-value: < 2.2e-16
```

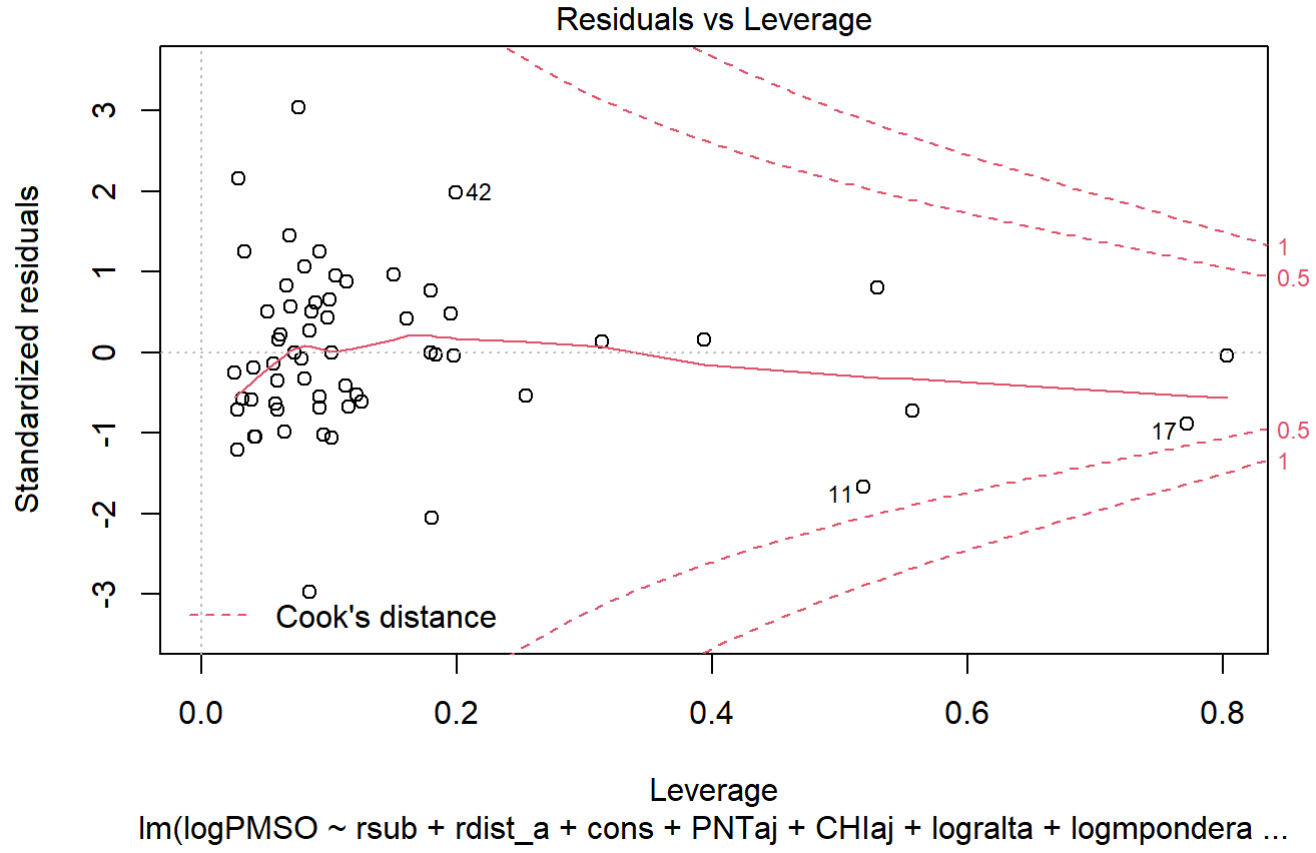
```
plot(modelo)
```


Residuals vs Fitted
Residuals
0.5
0.0
-0.5
5
37
58
8
9
10
11
12
13
14
15
Fitted values
lm(logPMSO ~ rsub + rdist_a + cons + PNTaj + CHlaj + logralta + logmpondera ...


Normal Q-Q
Standardized residuals
3
2
1
0
-1
-2
-3
5
37
58
-2
-1
0
1
2
Theoretical Quantiles
lm(logPMSO ~ rsub + rdist_a + cons + PNTaj + CHlaj + logralta + logmpondera ...


Scale-Location





Predizendo o PMSO utilizando a base de dados original

Utilizando a base de dados original, o R^2 preditivo encontrado foi de 0.97, podendo ser considerada como uma boa base para predizer os valores do ano de 2017.

Pela análise da árvore para essa base de dados, o valor do R^2 preditivo decresce para 0.6276, independente para o valor utilizado como maxdepth, mostrando que o modelo não consegue prever corretamente dados futuros para a variável PMSO.

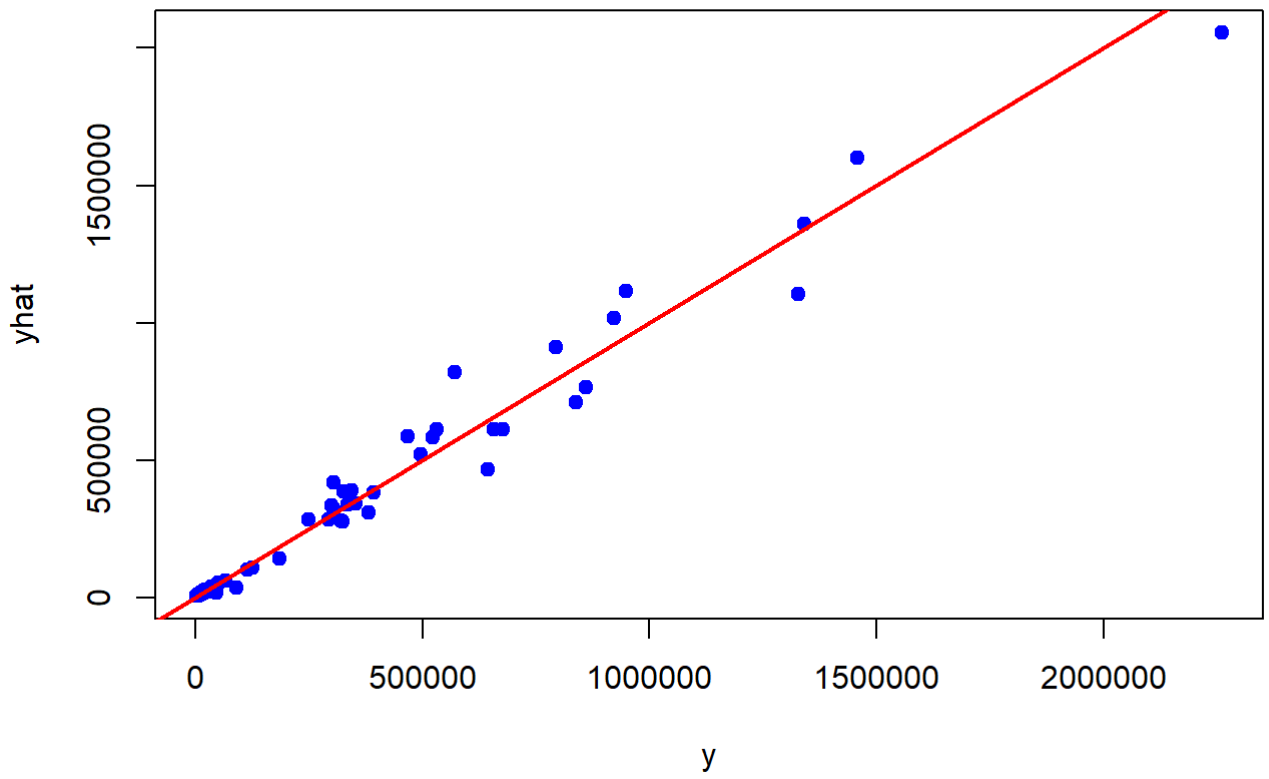
```
require(rpart)
```

```
## Loading required package: rpart
```

```
require(rpart.plot)
```

```
## Loading required package: rpart.plot
```

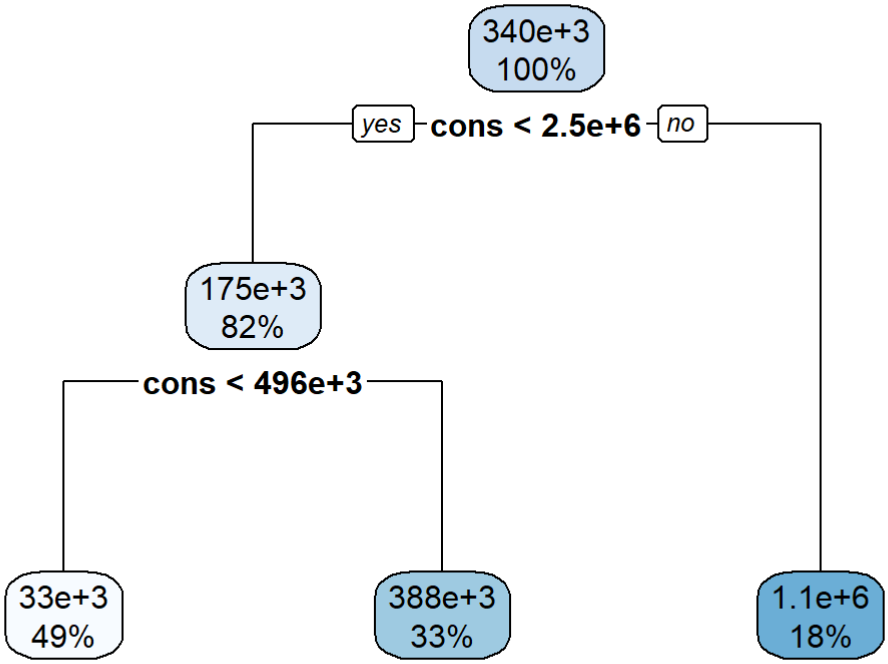
```
dados <- dados[1:8]
y    <- dados$PMSOaj
yhat <- rep(NA, nrow(dados))

for(cont in 1:nrow(dados)){
   modelo <- lm(PMSOaj ~ ., data=dados[-cont,])

   yhat[cont] <- predict(modelo, newdata=dados[cont,])
}

plot(yhat ~ y, pch=19, col="blue")
abline(a=0, b=1, lwd=2, col="red")
```



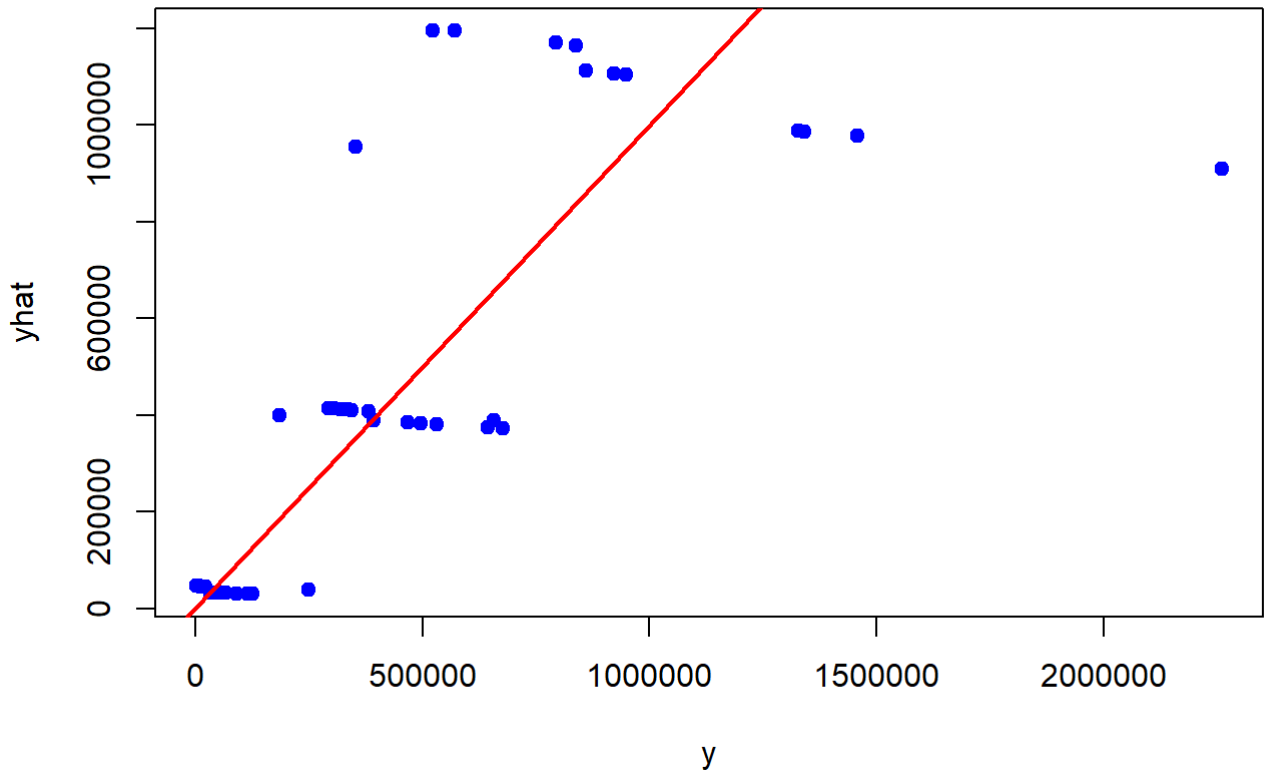
```
R2pred <- 1 - sum( (y-yhat)^2 )/sum( (y-mean(y))^2 )
print(R2pred)
```

```
## [1] 0.9700012
```

```
modelo <- rpart(PMSOaj ~  rsub + rdist_a + ralta + mponderado + cons + PNTaj + CHIaj,  data=d
ados)

rpart.plot(modelo)
```




```
y    <- dados$PMSOaj
yhat <- rep(NA, nrow(dados))

for(cont in 1:nrow(dados)){
 modelo <- rpart(PMSOaj ~  rsub + rdist_a + ralta + mponderado + cons + PNTaj + CHIaj, data=d
ados[-cont,],
            control = rpart.control(maxdepth=2))

 yhat[cont] <- predict(modelo, newdata=dados[cont,])
}

plot(yhat ~ y, pch=19, col="blue")
abline(a=0, b=1, lwd=2, col="red")
```




```
R2pred <- 1 - sum( (y-yhat)^2 )/sum( (y-mean(y))^2 )
print(R2pred)
```

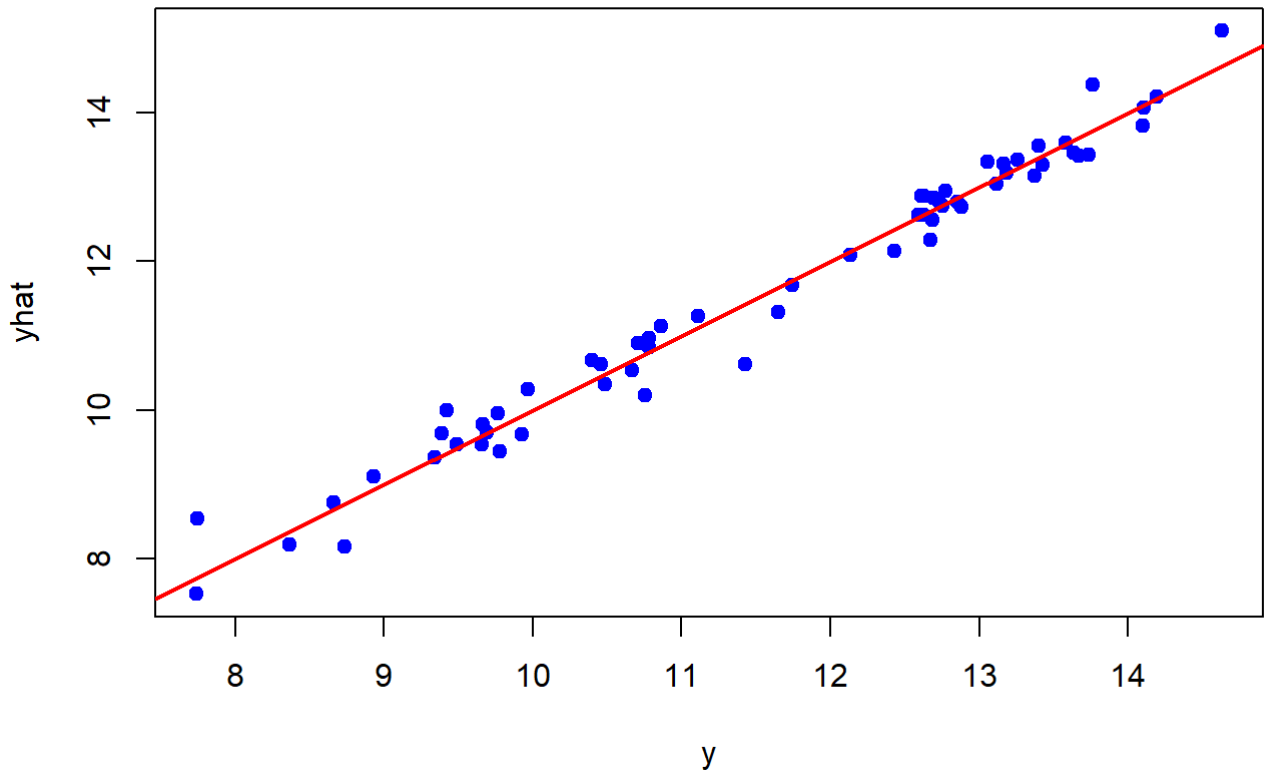
```
## [1] 0.6276422
```

Predizendo o PMSO utilizando a base de dados do modelo mais factível

Pelo fato do R² preditivo da base original já ser extremamente alto, um pequeno acréscimo pode ser considerado como um ganho. Com o modelo encontrado na primeira parte do trabalho, o R² preditivo gerado foi de 0.9763, melhor que quando comparado ao resultado da base original. Aliado a esse melhor R² preditivo, o ganho também se dá com o comportamento mais padronizado dos resíduos desse segundo modelo, já que o comportamento dos resíduos originalmente gera uma solução não factivel para o problema.

O resultado encontrado com a aplicação da árvore também é mais satisfatório que na utilização da base original de dados. O R² preditivo tem um acréscimo para 0.9089, prevendo melhor os valores que a base original.

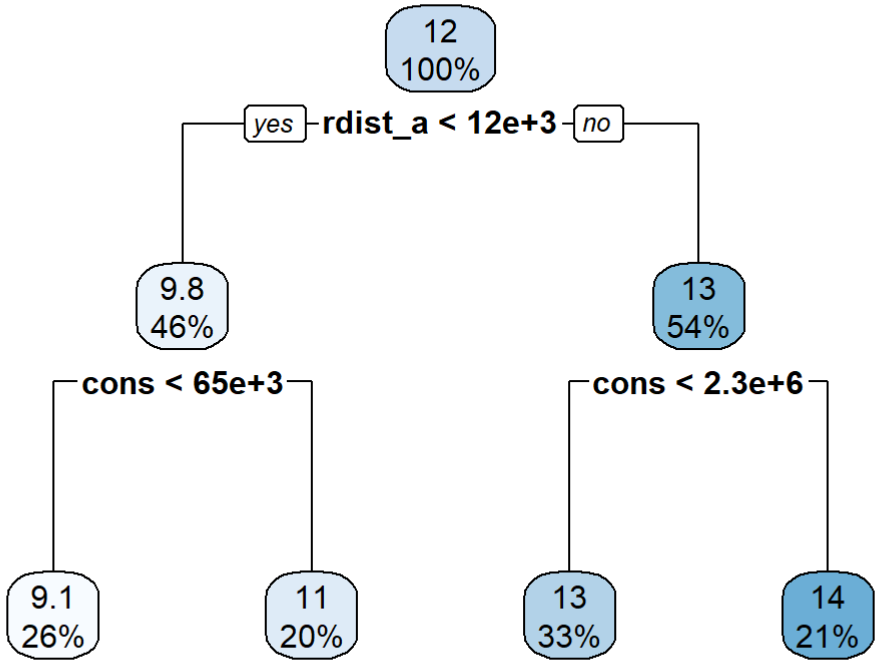
```
y    <- dados2$logPMSO
yhat <- rep(NA, nrow(dados))

for(cont in 1:nrow(dados2)){
   modelo <- lm(logPMSO ~ rsub + rdist_a + cons + PNTaj + CHIaj + logralta + logmponderado +
 logcons, data=dados2[-cont,])

   yhat[cont] <- predict(modelo, newdata=dados2[cont,])
}

plot(yhat ~ y, pch=19, col="blue")
abline(a=0, b=1, lwd=2, col="red")
```



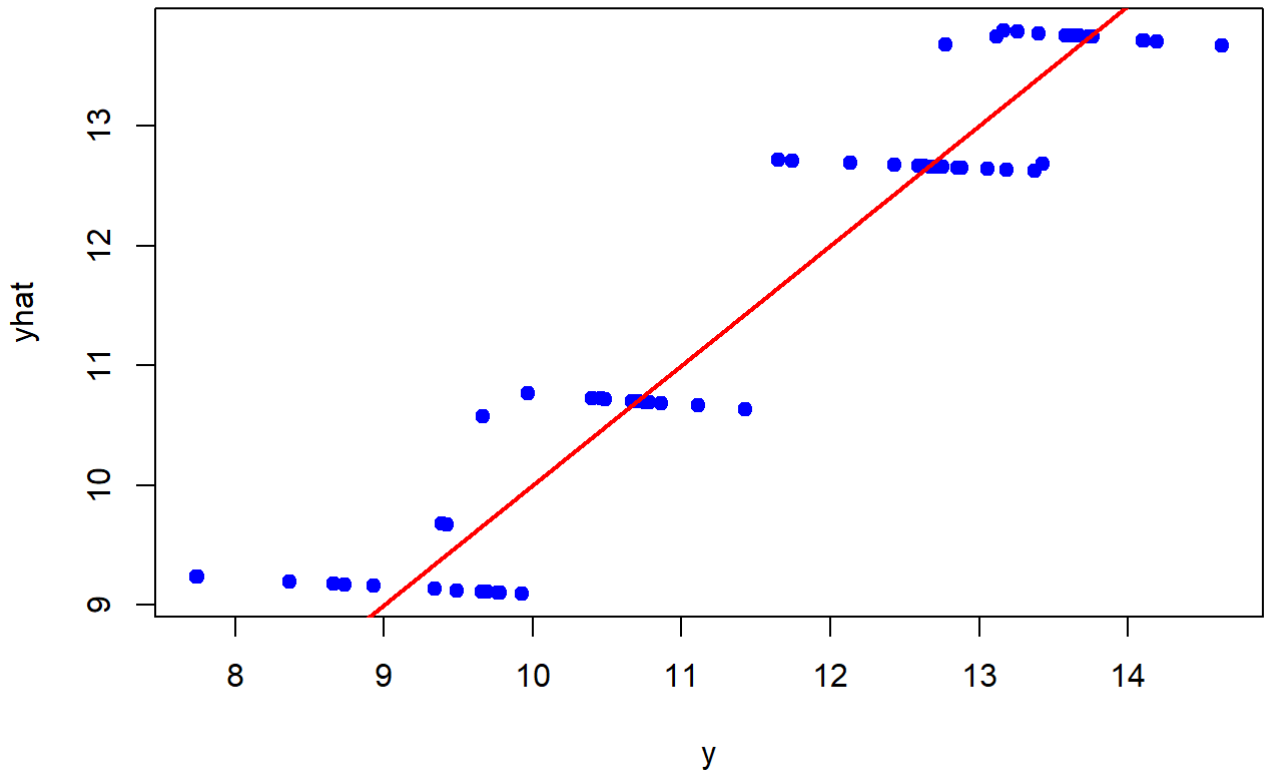

```
R2pred <- 1 - sum( (y-yhat)^2 )/sum( (y-mean(y))^2 )
print(R2pred)
```

```
## [1] 0.9763237
```

```
modelo <- rpart(logPMSO ~ rsub + rdist_a + cons + PNTaj + CHIaj + logralta + logmponderado +
logcons, data=dados2)

rpart.plot(modelo)
```




```
y    <- dados2$logPMSO
yhat <- rep(NA, nrow(dados2))

for(cont in 1:nrow(dados2)){
 modelo <- rpart(logPMSO ~ rsub + rdist_a + cons + PNTaj + CHIaj + logralta + logmponderado +
logcons, data=dados2[-cont,],
            control = rpart.control(maxdepth=3))

 yhat[cont] <- predict(modelo, newdata=dados2[cont,])
}

plot(yhat ~ y, pch=19, col="blue")
abline(a=0, b=1, lwd=2, col="red")
```




```
R2pred <- 1 - sum( (y-yhat)^2 )/sum( (y-mean(y))^2 )
print(R2pred)
```

```
## [1] 0.9089139
```