Selecao.R

Cristiane Rodrigues Maragno

16/10/2023

```
library(data.table)
dados <- fread(input = paste0("selecao.csv"), header = T, na.strings =
"NA", data.table = FALSE, dec=",")
names(dados)
```

A) seleção de variáveis “forward”

```
m0=lm(y ~ 1, data=dados)
m1=step(m0,list(lower = ~ 1,
                upper = ~ x1+x2+x3+x4+x5+x6+x7+x8+x9),
                direction="forward")
```

```
## Start:  AIC=5655.48
## y ~ 1
##
##        Df Sum of Sq   RSS    AIC
## + x7    1     53916 21387 3863.8
## + x9    1     53030 22274 3921.7
## + x8    1     42439 32865 4476.0
## + x5    1     33450 41854 4820.5
## + x6    1     33024 42279 4834.9
## + x4    1     17408 57896 5282.9
## + x2    1       257 75047 5652.6
## + x3    1       253 75050 5652.7
## <none>              75304 5655.5
## + x1    1        32 75272 5656.9
##
## Step:  AIC=3863.78
## y ~ x7
##
##        Df Sum of Sq   RSS    AIC
## + x9    1   1283.44 20104 3777.6
## + x8    1   1260.79 20126 3779.2
## + x6    1    783.20 20604 3812.6
## + x5    1    420.78 20966 3837.5
## + x1    1     89.07 21298 3859.8
## + x4    1     53.18 21334 3862.2
## <none>              21387 3863.8
## + x3    1      8.26 21379 3865.2
## + x2    1      3.51 21384 3865.6
##
```

```
## Step:  AIC=3777.6
## y ~ x7 + x9
##
##        Df Sum of Sq   RSS    AIC
## + x6    1   123.072 19981 3770.8
## + x4    1    81.652 20022 3773.8
## + x1    1    54.956 20049 3775.7
## + x8    1    36.266 20068 3777.0
## <none>              20104 3777.6
## + x5    1     4.217 20100 3779.3
## + x2    1     0.713 20103 3779.5
## + x3    1     0.536 20103 3779.6
##
## Step:  AIC=3770.85
## y ~ x7 + x9 + x6
##
##        Df Sum of Sq   RSS    AIC
## + x4    1   1080.45 18900 3693.6
## + x1    1     81.08 19900 3767.1
## + x8    1     36.90 19944 3770.2
## + x5    1     34.74 19946 3770.4
## <none>              19981 3770.8
## + x2    1      0.78 19980 3772.8
## + x3    1      0.23 19980 3772.8
##
## Step:  AIC=3693.63
## y ~ x7 + x9 + x6 + x4
##
##        Df Sum of Sq   RSS    AIC
## + x5    1   236.064 18664 3677.7
## + x1    1    62.304 18838 3690.9
## + x8    1    36.381 18864 3692.9
## <none>              18900 3693.6
## + x3    1     1.815 18898 3695.5
## + x2    1     1.213 18899 3695.5
##
## Step:  AIC=3677.72
## y ~ x7 + x9 + x6 + x4 + x5
##
##        Df Sum of Sq   RSS    AIC
## + x1    1    52.902 18611 3675.7
## + x8    1    40.046 18624 3676.7
## <none>              18664 3677.7
## + x3    1     6.807 18657 3679.2
## + x2    1     1.911 18662 3679.6
##
## Step:  AIC=3675.67
## y ~ x7 + x9 + x6 + x4 + x5 + x1
##
##        Df Sum of Sq   RSS    AIC
```

```
## + x8    1    41.362 18570 3674.5
## <none>              18611 3675.7
## + x3    1     7.054 18604 3677.1
## + x2    1     3.030 18608 3677.4
##
## Step:  AIC=3674.5
## y ~ x7 + x9 + x6 + x4 + x5 + x1 + x8
##
##        Df Sum of Sq   RSS    AIC
## <none>              18570 3674.5
## + x3    1    6.5176 18563 3676.0
## + x2    1    3.3522 18567 3676.2
```

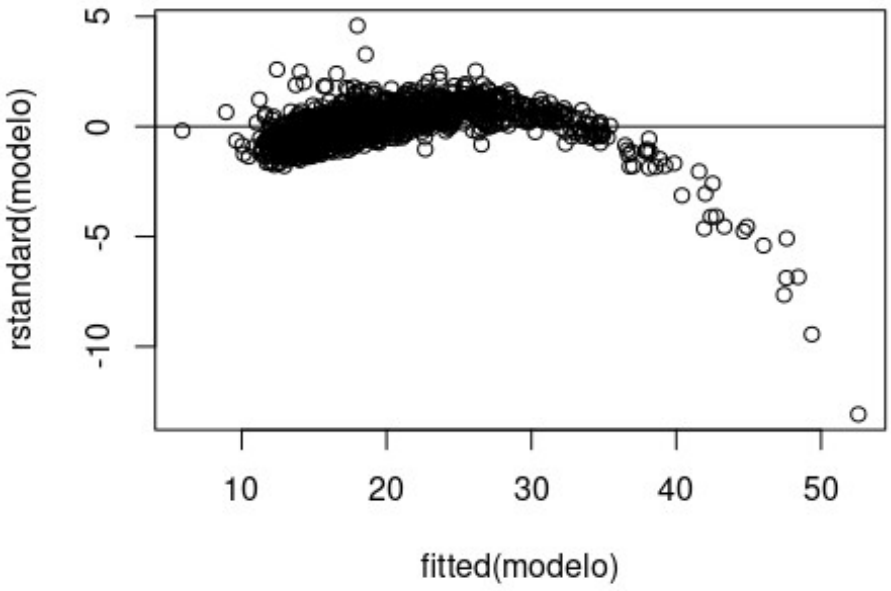
Primeira versão do modelo:

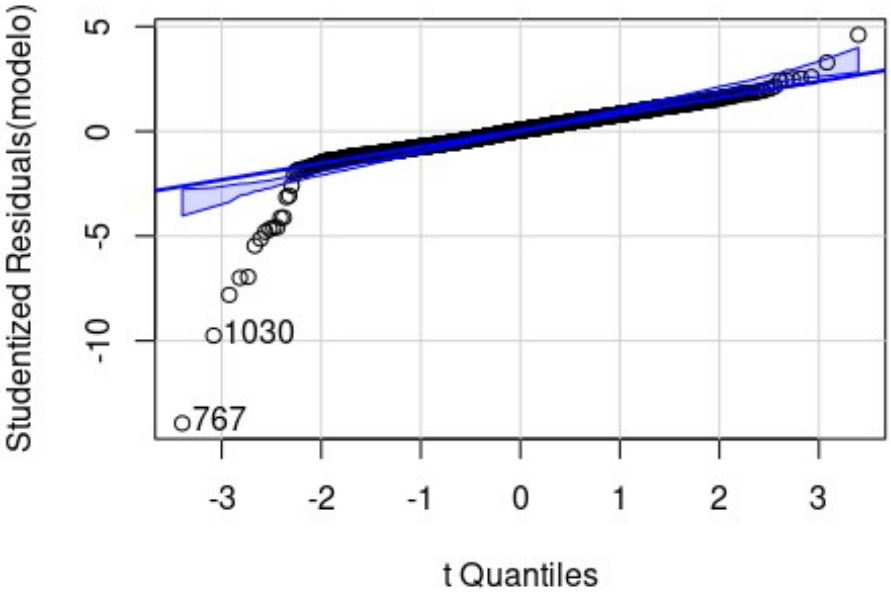
```
modelo <- lm(y ~ x7 + x9 + x6 + x4 + x5 + x1 + x8, data=dados)
summary(modelo)
```

```
##
## Call:
## lm(formula = y ~ x7 + x9 + x6 + x4 + x5 + x1 + x8, data = dados)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.172  -1.925   0.169   2.063  16.032
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.5374685  1.4415446  -4.535 6.24e-06 ***
## x7          -0.2638935  0.4659336  -0.566   0.5712
## x9           0.9462288  0.4643885   2.038   0.0418 *
## x6           0.2664206  0.0278735   9.558  < 2e-16 ***
## x4          -0.2068910  0.0211328  -9.790  < 2e-16 ***
## x5           0.3019813  0.0720512   4.191 2.95e-05 ***
## x1          -0.0011609  0.0005707  -2.034   0.0421 *
## x8          -0.8250781  0.4644232  -1.777   0.0759 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.62 on 1417 degrees of freedom
## Multiple R-squared:  0.7534, Adjusted R-squared:  0.7522
## F-statistic: 618.4 on 7 and 1417 DF,  p-value: < 2.2e-16
```

B)Análise de resíduos

```
plot(fitted(modelo), rstandard(modelo))
abline(0,0)

qqPlot(modelo)
```

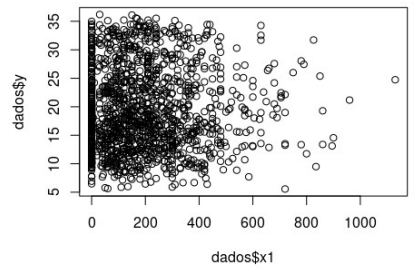




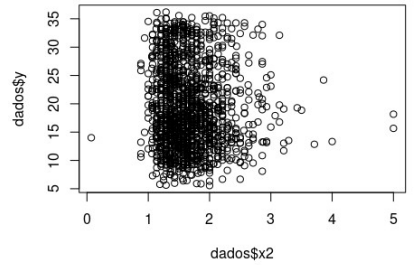
Podemos observar que o resíduo não apresenta um corpotamento aleatório e há muitos devios. Logo, o modelo não está bem ajustado.

Analisando os gráficos de cada variável independente em relação a variável dependende, podemos observar que as últimas precisam passar por uma transformação dos dados.

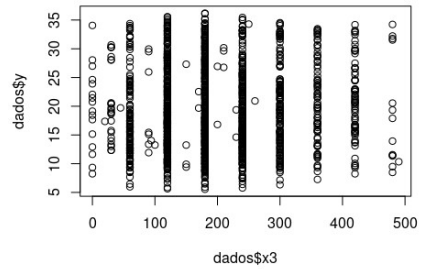
```
plot(dados$x1,dados$y)
```



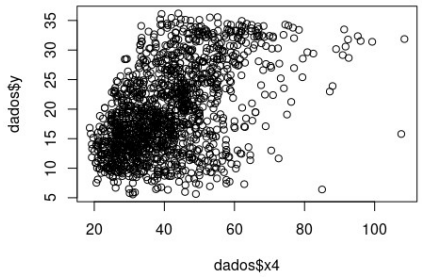

```
plot(dados$x2,dados$y)
```



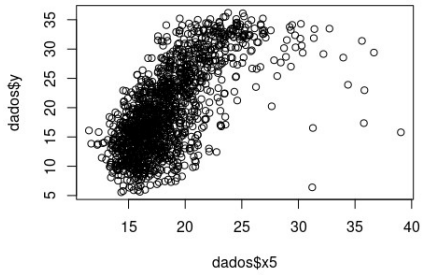

```
plot(dados$x3,dados$y)
```



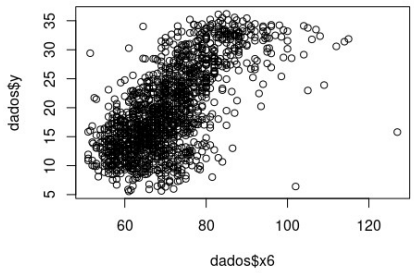

```
plot(dados$x4,dados$y)
```



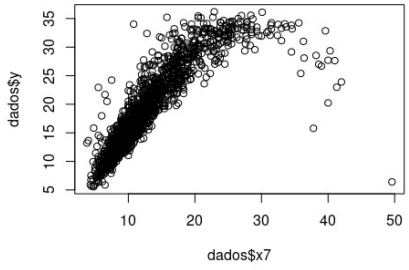

```
plot(dados$x5,dados$y)
```



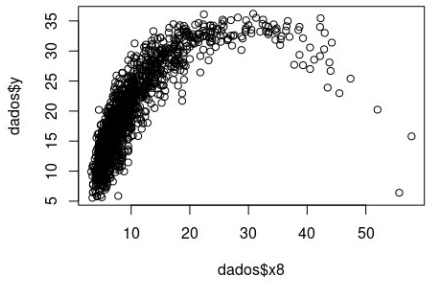

```
plot(dados$x6,dados$y)
```



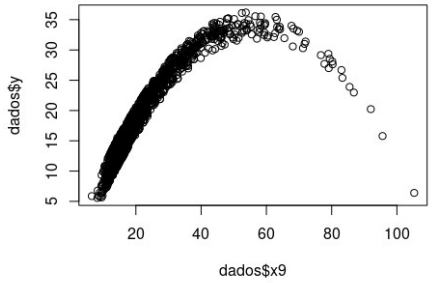

```
plot(dados$x7,dados$y)
```



```r
plot(dados$x8,dados$y)
```




```r
plot(dados$x9,dados$y)
```




```r
dados$x8_2 <- dados$x8^2
dados$x9_2 <- dados$x9^2
```

```r
C) Refazer modelo e análise de resíduos
m0=lm(y ~ 1, data=dados)
m2=step(m0,list(lower = ~ 1,
                upper = ~ x1+x2+x3+x4+x5+x6+x7+x8_2+x9_2),
                direction="forward")

## Start:  AIC=5655.48
y ~ 1


       Df Sum of Sq   RSS    AIC

+ x7    1     53916 21387 3863.8

+ x5    1     33450 41854 4820.5

+ x6    1     33024 42279 4834.9

+ x9_2  1     30948 44356 4903.2
```

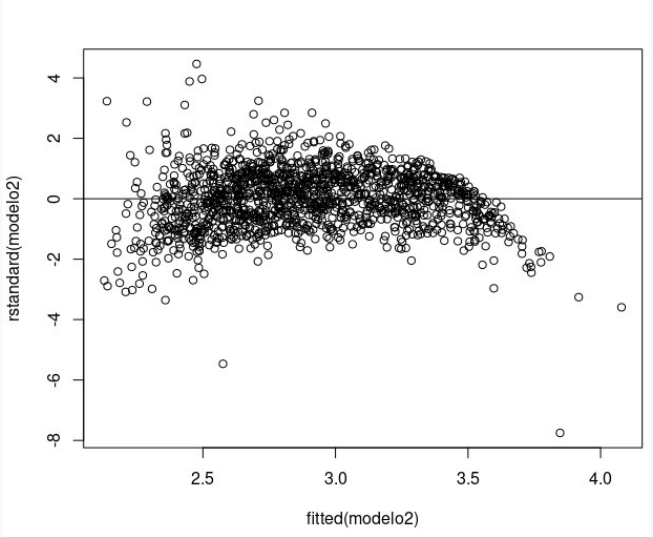
```
+ x8_2  1     21855 53449 5169.0
+ x4    1     17408 57896 5282.9
+ x2    1       257 75047 5652.6
+ x3    1       253 75050 5652.7
<none>              75304 5655.5
+ x1    1        32 75272 5656.9

Step:  AIC=3863.78
y ~ x7

       Df Sum of Sq   RSS    AIC
+ x9_2  1   2311.64 19076 3702.8
+ x6    1    783.20 20604 3812.6
+ x8_2  1    487.63 20900 3832.9
+ x5    1    420.78 20966 3837.5
+ x1    1     89.07 21298 3859.8
+ x4    1     53.18 21334 3862.2
<none>              21387 3863.8
+ x3    1      8.26 21379 3865.2
+ x2    1      3.51 21384 3865.6

Step:  AIC=3702.79
y ~ x7 + x9_2

       Df Sum of Sq   RSS    AIC
+ x8_2  1    8871.6 10204 2813.3
+ x6    1    2004.0 17072 3546.6
+ x5    1    1627.3 17448 3577.7
+ x4    1     491.3 18584 3667.6
```

```
+ x1    1     135.6 18940 3694.6
<none>              19076 3702.8
+ x3    1      20.4 19055 3703.3
+ x2    1      13.5 19062 3703.8

Step:  AIC=2813.26
y ~ x7 + x9_2 + x8_2

       Df Sum of Sq     RSS    AIC
+ x6    1    464.33  9739.7 2748.9
+ x5    1    428.90  9775.1 2754.1
+ x1    1    106.75 10097.2 2800.3
+ x4    1     64.11 10139.9 2806.3
+ x2    1     38.52 10165.5 2809.9
<none>              10204.0 2813.3
+ x3    1      2.49 10201.5 2814.9

Step:  AIC=2748.9
y ~ x7 + x9_2 + x8_2 + x6

       Df Sum of Sq    RSS    AIC
+ x4    1    387.55 9352.1 2693.0
+ x1    1    165.25 9574.4 2726.5
+ x5    1     74.02 9665.6 2740.0
+ x2    1     36.59 9703.1 2745.5
<none>              9739.7 2748.9
+ x3    1      0.82 9738.8 2750.8

Step:  AIC=2693.04
```

```
y ~ x7 + x9_2 + x8_2 + x6 + x4


       Df Sum of Sq    RSS    AIC
+ x5    1    542.00 8810.1 2610.0
+ x1    1    147.22 9204.9 2672.4
+ x2    1     36.62 9315.5 2689.4
<none>              9352.1 2693.0
+ x3    1      0.04 9352.1 2695.0


Step:  AIC=2609.96
y ~ x7 + x9_2 + x8_2 + x6 + x4 + x5


       Df Sum of Sq    RSS    AIC
+ x1    1   117.148 8692.9 2592.9
+ x2    1    40.185 8769.9 2605.4
<none>              8810.1 2610.0
+ x3    1     4.342 8805.8 2611.3


Step:  AIC=2592.88
y ~ x7 + x9_2 + x8_2 + x6 + x4 + x5 + x1


       Df Sum of Sq    RSS    AIC
<none>              8692.9 2592.9
+ x3    1    4.5806 8688.4 2594.1
+ x2    1    4.3853 8688.6 2594.2


Segunda versão do modelo


modelo2 <- lm(y ~ x7 + x9_2 + x8_2 + x6 + x4 + x5 + x1, data=dados)
summary(modelo2)
```

```
##
## Call:
## lm(formula = y ~ x7_2 + x6 + x9_2 + x8_2 + x4 + x5 + x1, data =
dados)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.8421  -1.6048  -0.0842   1.5021  15.2194
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.730e+01  9.910e-01 -37.635  < 2e-16 ***
## x7_2         1.674e+01  3.379e-01  49.535  < 2e-16 ***
## x6           2.294e-01  1.981e-02  11.580  < 2e-16 ***
## x9_2        -5.886e-03  3.461e-04 -17.006  < 2e-16 ***
## x8_2         1.283e-02  9.615e-04  13.340  < 2e-16 ***
## x4          -1.728e-01  1.529e-02 -11.304  < 2e-16 ***
## x5           4.589e-01  5.108e-02   8.983  < 2e-16 ***
## x1          -1.712e-03  4.125e-04  -4.149 3.53e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.623 on 1417 degrees of freedom
## Multiple R-squared:  0.8705, Adjusted R-squared:  0.8699
## F-statistic:  1361 on 7 and 1417 DF,  p-value: < 2.2e-16
```

```r
plot(fitted(modelo2), rstandard(modelo2))
abline(0,0)
```



Agora a análise de resíduos demonstra um corpotamento mais próximo de aleatório, porém com valores fora do limite de 3 e -3 no eixo y. Provavelmente outras trasformações dos dados seriam mais relevantes, porém, após diversos experimentos, não encontrei um que tivesse melhores resultados.