selecao.R

cris

2023-10-16

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

## [1] "x1" "x2" "x3" "x4" "x5" "x6" "x7" "x8" "x9" "y"

#A)  
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

# 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

# Contribuição de cada variável  
library(relaimpo)

## Carregando pacotes exigidos: MASS

## Carregando pacotes exigidos: boot

## Carregando pacotes exigidos: survey

## Carregando pacotes exigidos: grid

## Carregando pacotes exigidos: Matrix

## Carregando pacotes exigidos: survival

##   
## Attaching package: 'survival'

## The following object is masked from 'package:boot':  
##   
## aml

##   
## Attaching package: 'survey'

## The following object is masked from 'package:graphics':  
##   
## dotchart

## Carregando pacotes exigidos: mitools

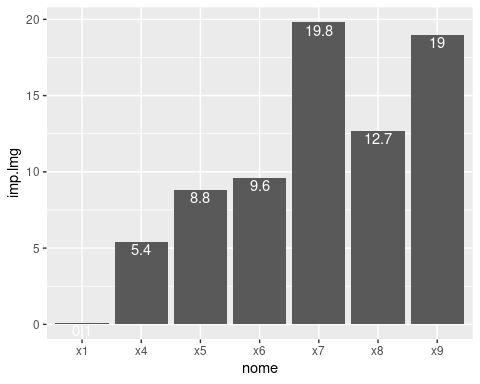
## This is the global version of package relaimpo.

## If you are a non-US user, a version with the interesting additional metric pmvd is available

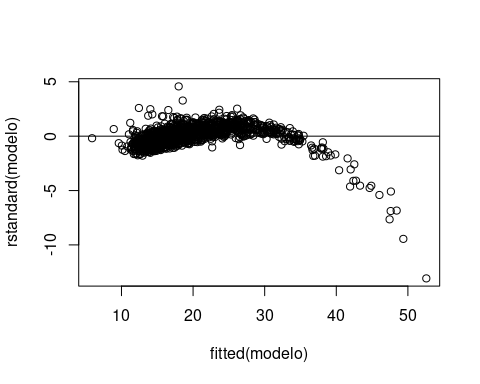
## from Ulrike Groempings web site at prof.beuth-hochschule.de/groemping.

imp<-calc.relimp(modelo)  
var.exp<-data.frame(round(imp$lmg\*100,1))  
colnames(var.exp)<-"imp.lmg"  
nome<-rownames(var.exp)  
var.exp<-data.frame(nome,var.exp)  
library(ggplot2)  
ggplot(var.exp, aes(nome,imp.lmg)) +  
 geom\_bar(stat = "identity")+  
 geom\_text(aes(label = imp.lmg), vjust = 1.2, lwd=5, colour = "white")

## Warning in geom\_text(aes(label = imp.lmg), vjust = 1.2, lwd = 5, colour =  
## "white"): Ignoring unknown parameters: `linewidth`



#B)  
plot(fitted(modelo), rstandard(modelo))  
abline(0,0)



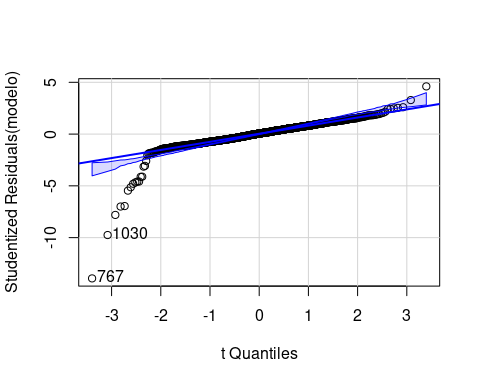
library(car)

## Carregando pacotes exigidos: carData

##   
## Attaching package: 'car'

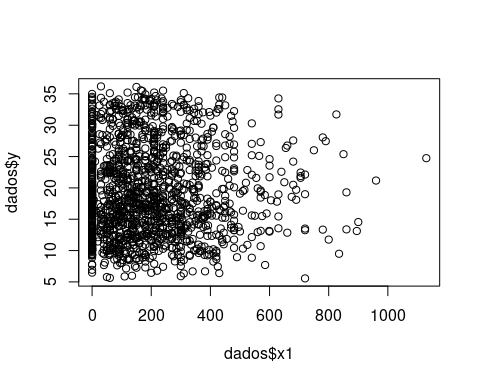
## The following object is masked from 'package:boot':  
##   
## logit

qqPlot(modelo)

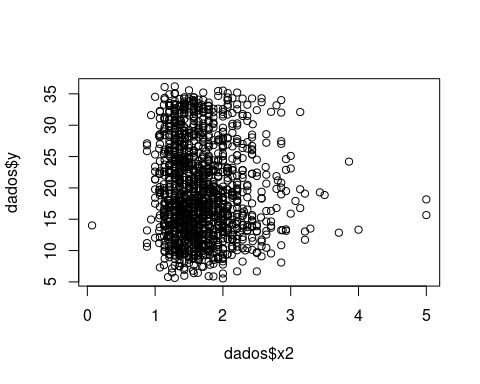


## [1] 767 1030

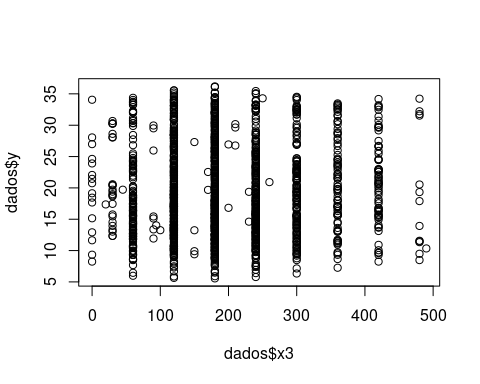
#Não está aleatório  
  
plot(dados$x1,dados$y)



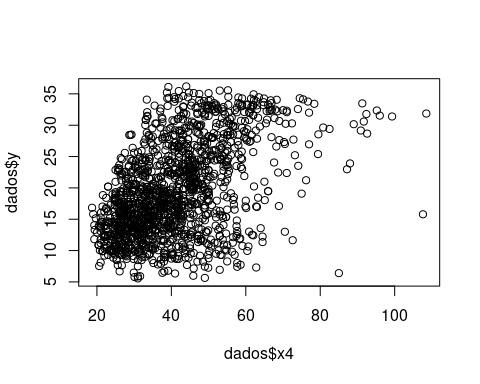
plot(dados$x2,dados$y)



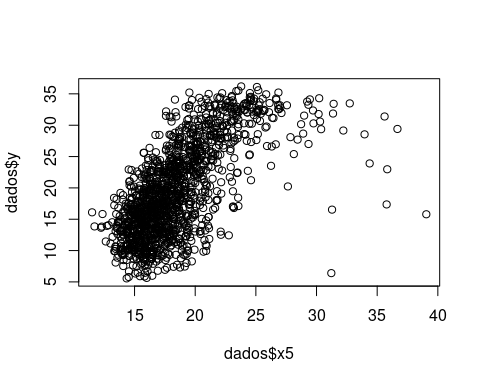
plot(dados$x3,dados$y)



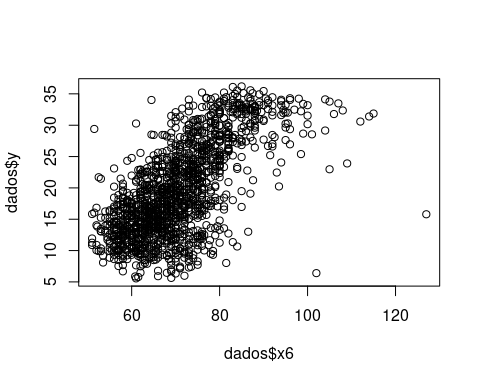
plot(dados$x4,dados$y)



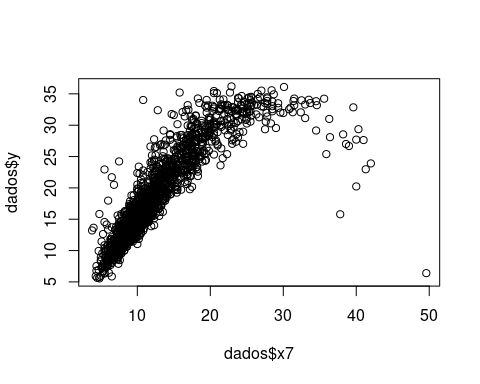
plot(dados$x5,dados$y)



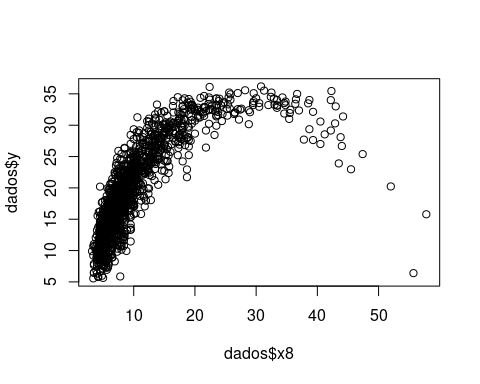
plot(dados$x6,dados$y)



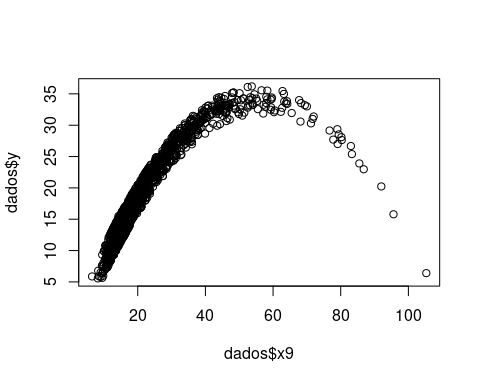
plot(dados$x7,dados$y) # Logaritmo



plot(dados$x8,dados$y) # Parabola



plot(dados$x9,dados$y) # Parabola



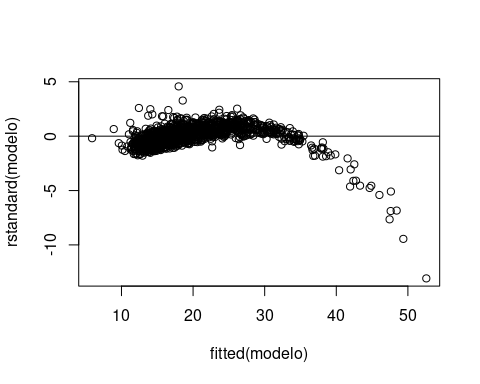
dados$x7\_2 <- log(dados$x7)  
dados$x8\_2 <- dados$x8^2  
dados$x9\_2 <- dados$x9^2  
  
#C)  
m0=lm(y ~ 1, data=dados)  
m2=step(m0,list(lower = ~ 1,  
 upper = ~ x1+x2+x3+x4+x5+x6+x7\_2+x8\_2+x9\_2),  
 direction="forward")

## Start: AIC=5655.48  
## y ~ 1  
##   
## Df Sum of Sq RSS AIC  
## + x7\_2 1 61816 13488 3206.9  
## + 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=3206.87  
## y ~ x7\_2  
##   
## Df Sum of Sq RSS AIC  
## + x6 1 505.75 12982 3154.4  
## + x5 1 328.23 13160 3173.8  
## + x9\_2 1 161.81 13326 3191.7  
## + x1 1 128.41 13360 3195.2  
## + x4 1 78.83 13409 3200.5  
## + x2 1 25.67 13462 3206.2  
## <none> 13488 3206.9  
## + x3 1 5.89 13482 3208.3  
## + x8\_2 1 0.84 13487 3208.8  
##   
## Step: AIC=3154.41  
## y ~ x7\_2 + x6  
##   
## Df Sum of Sq RSS AIC  
## + x9\_2 1 627.46 12355 3085.8  
## + x4 1 502.34 12480 3100.2  
## + x1 1 175.60 12807 3137.0  
## + x8\_2 1 115.17 12867 3143.7  
## + x2 1 19.54 12963 3154.3  
## <none> 12982 3154.4  
## + x5 1 4.74 12978 3155.9  
## + x3 1 0.12 12982 3156.4  
##   
## Step: AIC=3085.82  
## y ~ x7\_2 + x6 + x9\_2  
##   
## Df Sum of Sq RSS AIC  
## + x8\_2 1 1484.22 10871 2905.4  
## + x4 1 512.66 11842 3027.4  
## + x1 1 226.32 12128 3061.5  
## + x5 1 98.26 12257 3076.4  
## + x2 1 32.37 12322 3084.1  
## <none> 12355 3085.8  
## + x3 1 0.61 12354 3087.8  
##   
## Step: AIC=2905.44  
## y ~ x7\_2 + x6 + x9\_2 + x8\_2  
##   
## Df Sum of Sq RSS AIC  
## + x4 1 417.22 10453 2851.7  
## + x1 1 169.18 10701 2885.1  
## + x5 1 80.50 10790 2896.8  
## + x2 1 46.76 10824 2901.3  
## <none> 10871 2905.4  
## + x3 1 1.36 10869 2907.3  
##   
## Step: AIC=2851.67  
## y ~ x7\_2 + x6 + x9\_2 + x8\_2 + x4  
##   
## Df Sum of Sq RSS AIC  
## + x5 1 586.85 9866.5 2771.3  
## + x1 1 150.16 10303.2 2833.1  
## + x2 1 46.40 10407.0 2847.3  
## <none> 10453.4 2851.7  
## + x3 1 0.00 10453.4 2853.7  
##   
## Step: AIC=2771.34  
## y ~ x7\_2 + x6 + x9\_2 + x8\_2 + x4 + x5  
##   
## Df Sum of Sq RSS AIC  
## + x1 1 118.449 9748.1 2756.1  
## + x2 1 49.972 9816.5 2766.1  
## <none> 9866.5 2771.3  
## + x3 1 3.839 9862.7 2772.8  
##   
## Step: AIC=2756.13  
## y ~ x7\_2 + x6 + x9\_2 + x8\_2 + x4 + x5 + x1  
##   
## Df Sum of Sq RSS AIC  
## <none> 9748.1 2756.1  
## + x2 1 8.2254 9739.8 2756.9  
## + x3 1 4.0633 9744.0 2757.5

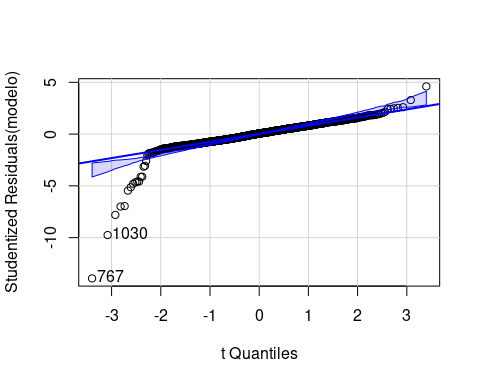
# Modelo  
modelo2 <- lm(y ~ x7\_2 + x6 + x9\_2 + x8\_2 + x4 + x5 + x1, data=dados)  
summary(modelo2)

##   
## 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

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



library(car)  
qqPlot(modelo)



## [1] 767 1030