paraTrabajo3.Rmd

Sylvia Acid

2 de mayo de 2017

*** El dataset Auto***

Description: Gas mileage, horsepower, and other information for cars.

```r
#install.packages("ISLR")
library("ISLR", lib.loc="~/R/x86_64-redhat-linux-gnu-library/3.2")
data("Auto")   # loads the dataset
```

```r
class(Auto)
```

```
## [1] "data.frame"
```

```r
colnames(Auto)
```

```
## [1] "mpg"          "cylinders"    "displacement" "horsepower"
## [5] "weight"       "acceleration" "year"         "origin"
## [9] "name"
```

```r
head(Auto)
```

```
##   mpg cylinders displacement horsepower weight acceleration year origin
## 1  18         8          307        130   3504         12.0   70      1
## 2  15         8          350        165   3693         11.5   70      1
## 3  18         8          318        150   3436         11.0   70      1
## 4  16         8          304        150   3433         12.0   70      1
## 5  17         8          302        140   3449         10.5   70      1
## 6  15         8          429        198   4341         10.0   70      1
##                        name
## 1 chevrolet chevelle malibu
## 2         buick skylark 320
## 3        plymouth satellite
## 4             amc rebel sst
## 5               ford torino
## 6          ford galaxie 500
```

```r
summary(Auto)
```

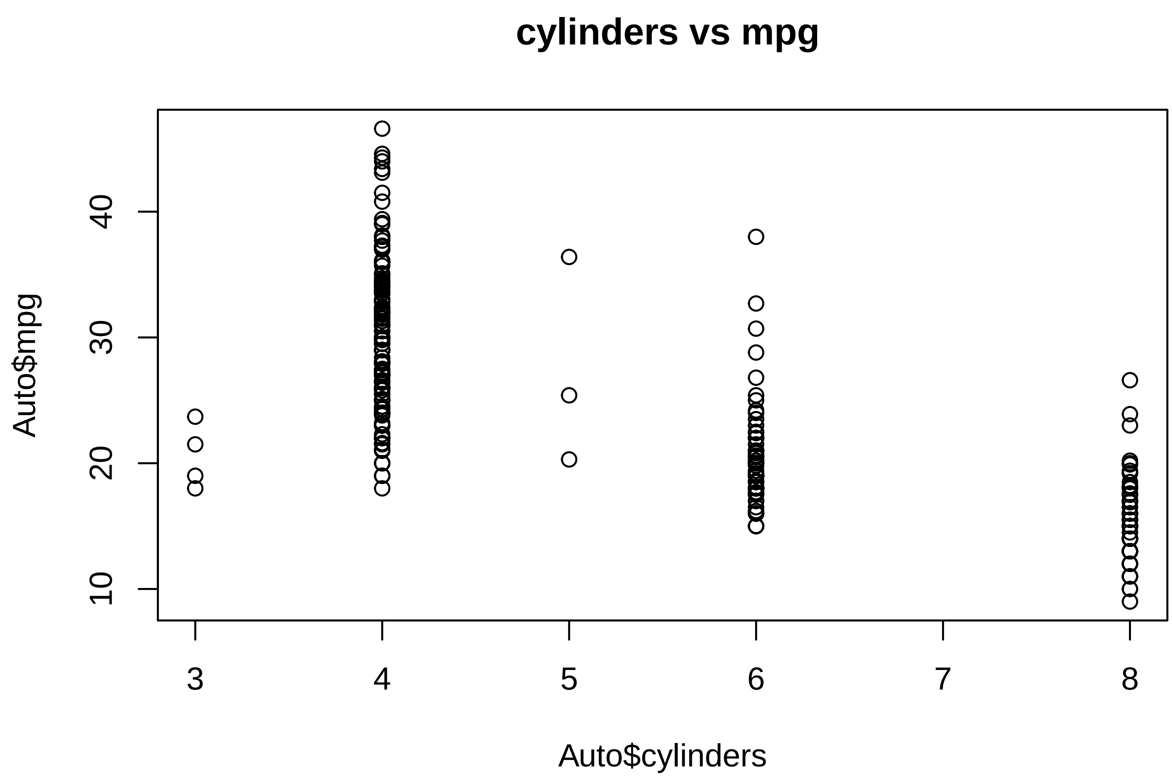
```
##       mpg          cylinders      displacement     horsepower
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0
##  1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0
##  Median :22.75   Median :4.000   Median :151.0   Median : 93.5
##  Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5
##  3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0
##  Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0
##
##      weight      acceleration        year           origin
##  Min.   :1613   Min.   : 8.00   Min.   :70.00   Min.   :1.000
##  1st Qu.:2225   1st Qu.:13.78   1st Qu.:73.00   1st Qu.:1.000
##  Median :2804   Median :15.50   Median :76.00   Median :1.000
##  Mean   :2978   Mean   :15.54   Mean   :75.98   Mean   :1.577
##  3rd Qu.:3615   3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000
```

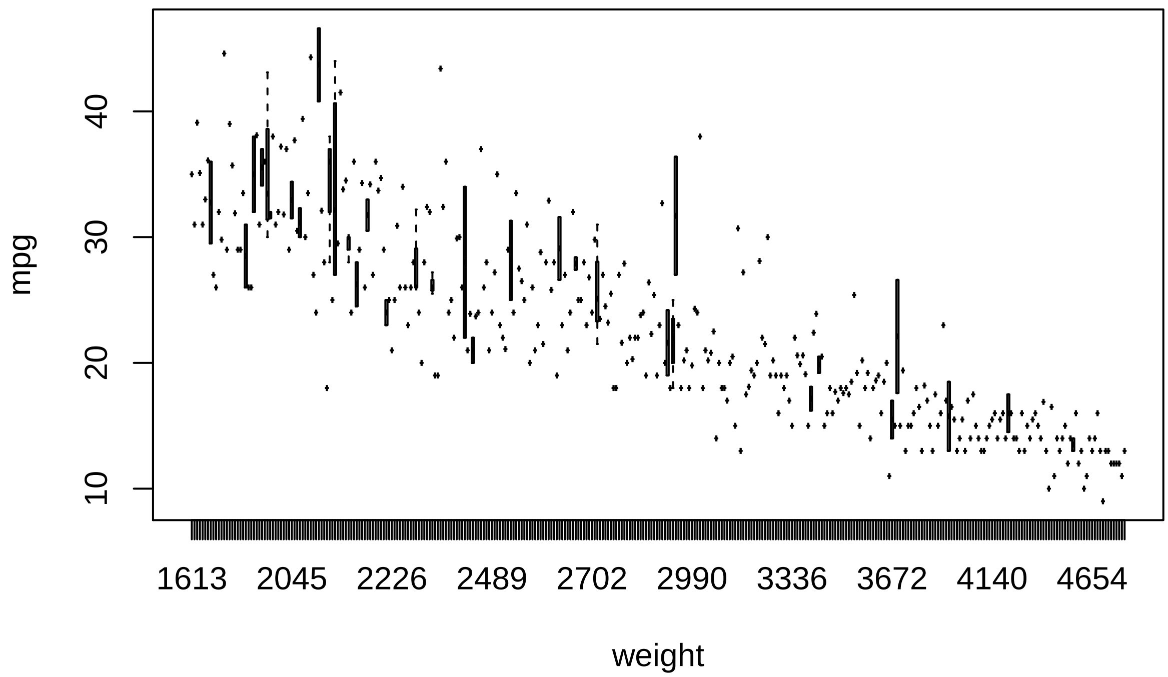
```
##  Max.   :5140   Max.   :24.80   Max.   :82.00   Max.   :3.000
##
##                  name
##  amc matador       :  5
##  ford pinto        :  5
##  toyota corolla    :  5
##  amc gremlin       :  4
##  amc hornet        :  4
##  chevrolet chevette:  4
##  (Other)           :365
```

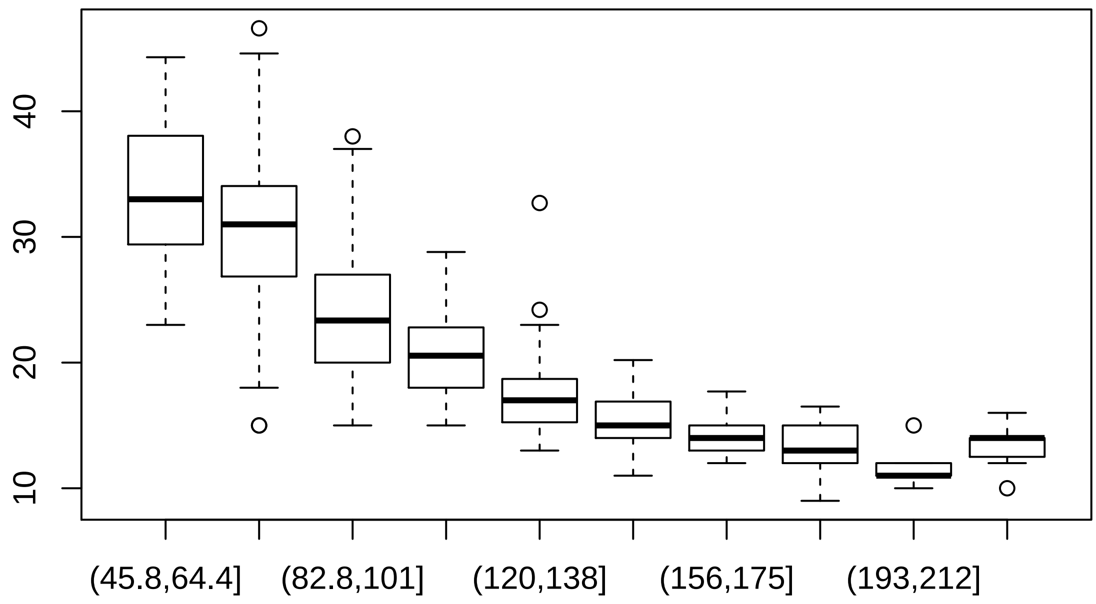
```r
dim(Auto)
```

```
## [1] 392   9
```

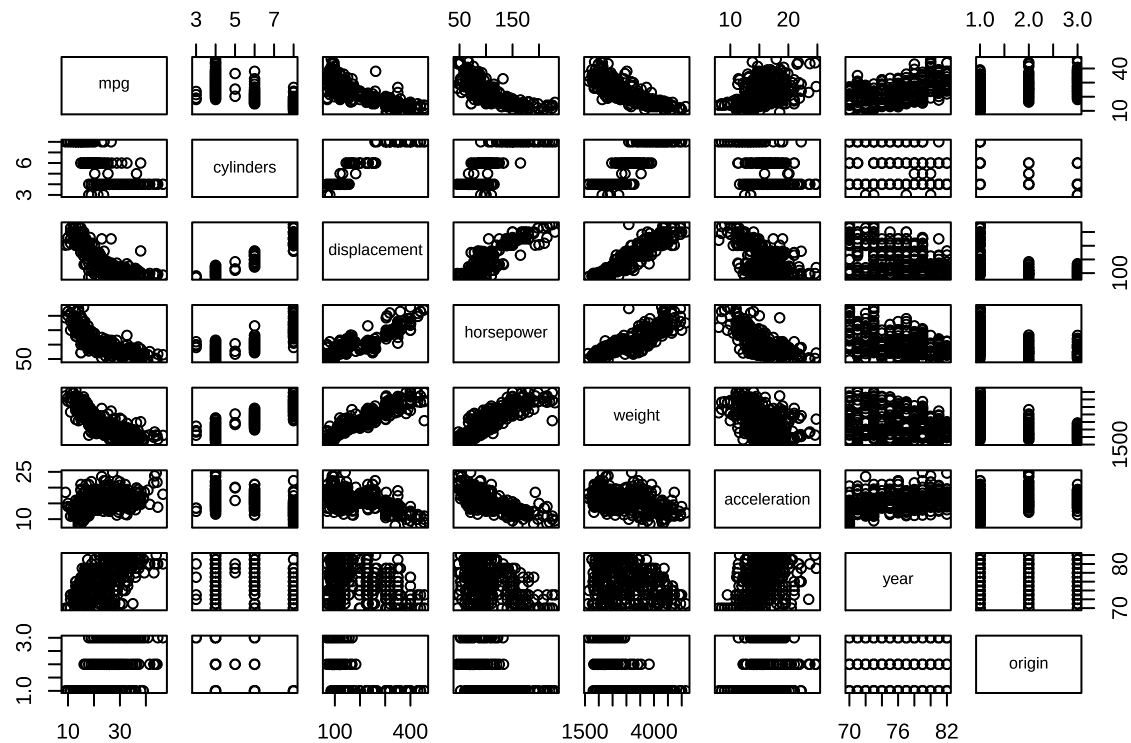
Estamos interesados en el atributo *mpg*, vamos a tratar de visualizar por pares los atributos *mpg* y *cylinders* mediante los comandos *plot* and *boxplot*.




mpg
40
30
20
10
1613
2045
2226
2489
2702
2990
3336
3672
4140
4654
weight




```
attach ( Auto )  # para simplificar y prescindir del prefijo Auto
#pairs(~ .,data = Auto)  # todos con todos
pairs(~ mpg + cylinders + displacement + horsepower + weight + acceleration + year + origin, data= Auto)
```




```
# solo algunas
```

Para evaluar los modelos, partimos el data.frame en training y test

```r
set.seed(1)
train = sample (nrow(Auto), round(nrow(Auto)*0.7)) # nos quedamos con los indices para el training
auto.train = Auto[train,]  # podemos reservarlos aparte ... aunque con el subset no sería necesario :(
auto.test = Auto[-train,]
```

```r
m1 = lm(mpg ~ weight, data=Auto, subset=train)
print(m1)
```

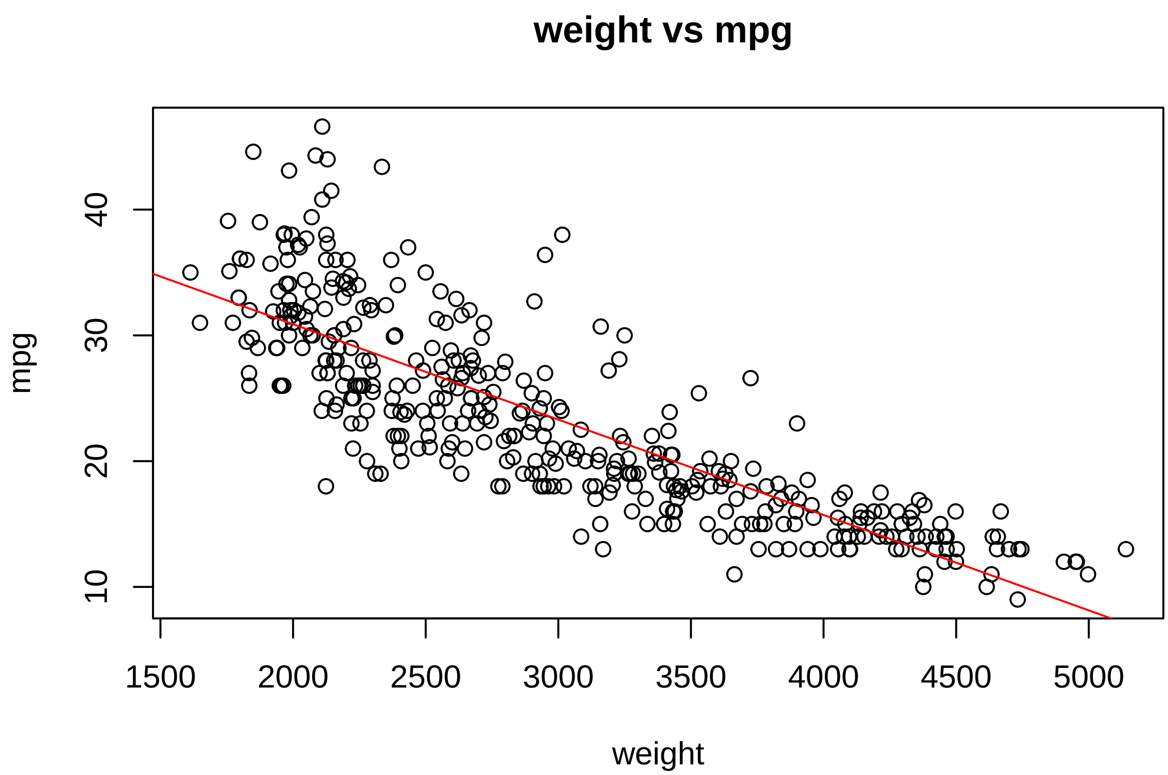
```
##
## Call:
## lm(formula = mpg ~ weight, data = Auto, subset = train)
##
## Coefficients:
## (Intercept)       weight
##   46.058884    -0.007585
```

```r
summary(m1)
```

```
##
## Call:
## lm(formula = mpg ~ weight, data = Auto, subset = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

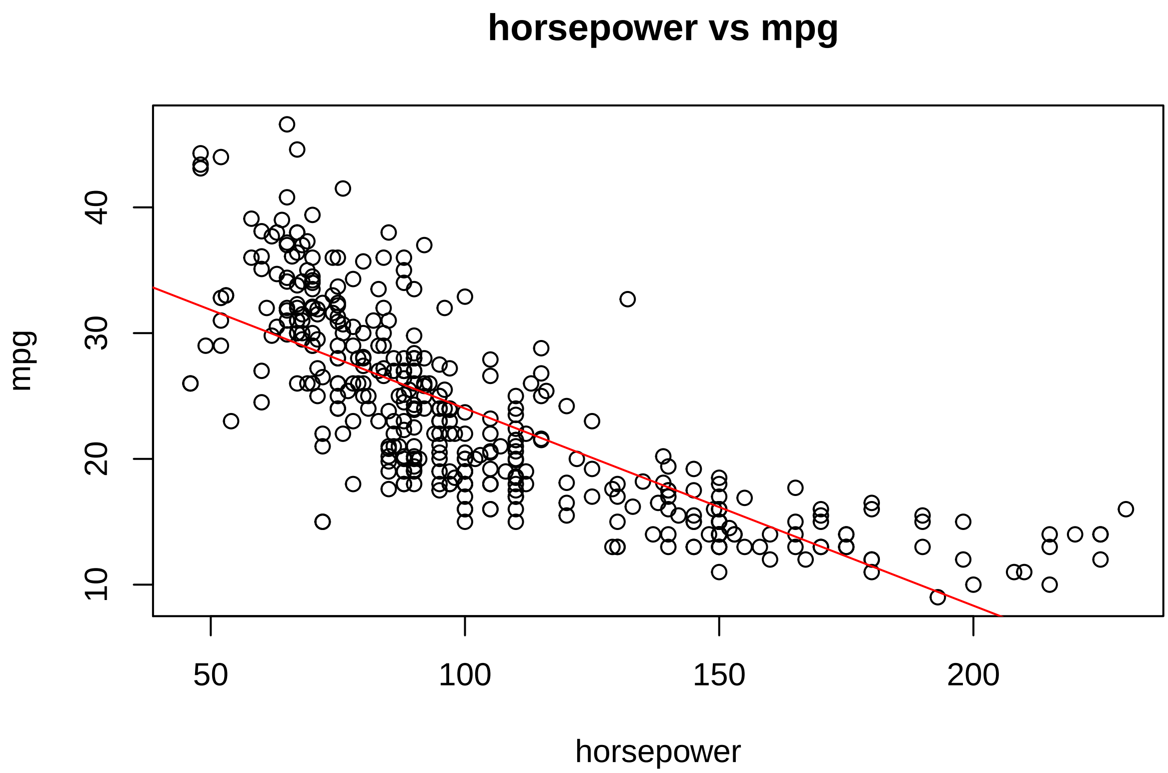
```
## -11.9477  -2.7053  -0.3457   2.2521  16.5461
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 46.0588840  0.9618282   47.89   <2e-16 ***
## weight      -0.0075853  0.0003078  -24.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.33 on 272 degrees of freedom
## Multiple R-squared:  0.6907, Adjusted R-squared:  0.6895
## F-statistic: 607.3 on 1 and 272 DF,  p-value: < 2.2e-16
```

```r
plot(weight, mpg, main=" weight vs mpg")
abline(m1$coefficients, col=2)
```



m1, nuestro primer modelo

```r
m2 = lm(mpg ~ horsepower, data=Auto, subset=train)
plot(horsepower, mpg, main=" horsepower vs mpg")
abline(m2$coefficients, col=2)
```




```r
summary(m2)
```

```
##
## Call:
## lm(formula = mpg ~ horsepower, data = Auto, subset = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.3988  -3.1685  -0.1685   2.9242  17.1036
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.688412   0.849380   46.73   <2e-16 ***
## horsepower  -0.156800   0.007602  -20.63   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.862 on 272 degrees of freedom
## Multiple R-squared:   0.61,  Adjusted R-squared:  0.6086
## F-statistic: 425.4 on 1 and 272 DF,  p-value: < 2.2e-16
```

```r
m3 = lm(mpg ~ ., data=Auto, subset=train) # en función del resto o bien
coef(m3)
```

```
##                             (Intercept)
##                            26.133442017
##                               cylinders
```

```
##                         -0.819274499
##                         displacement
##                         -0.008819364
##                           horsepower
##                         -0.054085458
##                               weight
##                         -0.003469509
##                         acceleration
##                         -0.521388533
##                                 year
##                          0.429590495
##                               origin
##                         -3.145378785
##              nameamc ambassador sst
##                         -0.834471654
##                      nameamc concord
##                         -3.920813986
##                 nameamc concord dl 6
##                         -4.079062879
##                      nameamc gremlin
##                         -6.480595008
##                       nameamc hornet
##                         -3.422799604
##                      nameamc matador
##                         -3.882664930
##                        nameamc pacer
##                         -4.373720635
##                    nameamc pacer d/l
##                         -5.448920714
##                    nameamc rebel sst
##                         -1.543809093
##                    nameamc spirit dl
##                         -3.770217135
##                      nameaudi 100 ls
##                         -0.834516880
##                        nameaudi 4000
##                          4.270542900
##                        nameaudi 5000
##                         -4.119443030
##             nameaudi 5000s (diesel)
##                         11.588001084
##                         nameaudi fox
##                          2.299859284
##                         namebmw 2002
##                          0.810118885
##                         namebmw 320i
##                         -5.433014033
##                    namebuick century
##                         -2.396260173
##                namebuick century 350
##                         -1.239202220
##            namebuick century limited
##                         -1.984655067
##         namebuick century luxus (sw)
```

```
##                                 -0.160610363
##                      namebuick century special
##                                 -3.299348799
##                   namebuick electra 225 custom
##                                  3.300879158
##                    namebuick estate wagon (sw)
##                                  0.245941860
##                       namebuick lesabre custom
##                                 -0.235883926
##                    namebuick opel isuzu deluxe
##                                 -2.290304724
##          namebuick regal sport coupe (turbo)
##                                 -3.980035696
##                              namebuick skyhawk
##                                 -2.940363484
##                              namebuick skylark
##                                 -3.290033626
##                      namebuick skylark limited
##                                 -1.443393101
##                          namecadillac eldorado
##                                  5.079189611
##                           namecadillac seville
##                                  1.744666478
##                                   namecapri ii
##                                 -4.457002990
##                namechevroelt chevelle malibu
##                                 -3.241524107
##                          namechevrolet bel air
##                                 -0.019925296
##                           namechevrolet camaro
##                                 -2.482896920
##                 namechevrolet caprice classic
##                                 -2.298202095
##                         namechevrolet cavalier
##                                 -1.932810076
##                  namechevrolet cavalier 2-door
##                                  2.504371341
##                   namechevrolet cavalier wagon
##                                 -3.332765788
##          namechevrolet chevelle concours (sw)
##                                 -3.108240476
##                 namechevrolet chevelle malibu
##                                 -4.328436236
##        namechevrolet chevelle malibu classic
##                                 -2.092874665
##                         namechevrolet chevette
##                                 -1.932623948
##                         namechevrolet citation
##                                 -2.808047375
##                         namechevrolet concours
##                                 -4.733198691
##                           namechevrolet impala
##                                 -0.616360093
##                           namechevrolet malibu
```

```
##                                   -3.496630872
##        namechevrolet malibu classic (sw)
##                                   -1.727655290
##         namechevrolet monte carlo landau
##                                   -1.444230815
##              namechevrolet monte carlo s
##                                   -0.924217129
##                  namechevrolet monza 2+2
##                                   -2.179046330
##                       namechevrolet nova
##                                   -4.173186853
##                namechevrolet nova custom
##                                   -5.061090863
##                       namechevrolet vega
##                                   -5.329569270
##                  namechevrolet vega (sw)
##                                   -4.822066197
##                      namechevrolet woody
##                                   -4.719713248
##                            namechevy c10
##                                   -4.828053896
##                           namechevy s-10
##                                    1.099128573
##           namechrysler lebaron medallion
##                                   -6.056887931
##               namechrysler lebaron salon
##                                   -8.010726523
## namechrysler lebaron town @ country (sw)
##                                   -0.135809474
##         namechrysler new yorker brougham
##                                    2.878320132
##               namechrysler newport royal
##                                    1.299126039
##                         namedatsun 200sx
##                                    7.930329167
##                           namedatsun 210
##                                   14.609077658
##                           namedatsun 310
##                                    9.242283793
##                        namedatsun 310 gx
##                                    9.147824611
##                           namedatsun 610
##                                    0.113083713
##                           namedatsun 710
##                                    5.013464886
##                           namedatsun 810
##                                    1.000591876
##                    namedatsun 810 maxima
##                                    2.760217015
##                         namedatsun b-210
##                                    6.234471053
##                namedatsun f-10 hatchback
##                                    7.044474938
##                         namedatsun pl510
```

```
##                        4.073644592
##           namedodge aries wagon (sw)
##                       -5.758003468
##                     namedodge aspen
##                       -3.872180166
##                   namedodge aspen 6
##                       -3.163707545
##                  namedodge aspen se
##                       -1.432636080
##               namedodge charger 2.2
##                        1.797194485
##                 namedodge colt (sw)
##                       -2.121500690
##     namedodge colt hatchback custom
##                        1.394624935
##                  namedodge colt m/m
##                        1.553266570
##           namedodge coronet brougham
##                       -0.850073979
##        namedodge coronet custom (sw)
##                       -0.803839768
##                      namedodge d100
##                       -4.837921943
##                      namedodge d200
##                        0.899436509
##               namedodge dart custom
##                       -4.348461329
##                  namedodge diplomat
##                       -0.324458525
##               namedodge monaco (sw)
##                        1.365792630
##           namedodge monaco brougham
##                       -1.858595251
##                 namedodge st. regis
##                       -0.852095872
##                     namefiat 124b
##                        2.544757441
##                      namefiat 128
##                       -1.570510849
##                      namefiat 131
##                        1.010950162
##              namefiat strada custom
##                        6.385689169
##         nameford country squire (sw)
##                       -0.532052096
##                  nameford escort 2h
##                       -1.177768395
##                  nameford escort 4w
##                       -0.186302372
##                     nameford f108
##                       -5.140358839
##                     nameford f250
##                        1.609366995
##                   nameford fairmont
```

```
##                           -2.289349759
##                      nameford fairmont 4
##                           -6.307479906
##                 nameford fairmont futura
##                           -6.235896969
##                          nameford futura
##                           -4.701270744
##                     nameford galaxie 500
##                           -0.501545427
##                     nameford gran torino
##                           -1.475602224
##                nameford gran torino (sw)
##                            0.445648323
##                       nameford granada l
##                           -5.694703105
##                             nameford ltd
##                           -0.356660330
##                      nameford ltd landau
##                           -3.220516269
##                        nameford maverick
##                           -4.466865619
##                         nameford mustang
##                           -5.158057013
##                      nameford mustang gl
##                           -4.237733733
##                      nameford mustang ii
##                           -8.761145344
##                           nameford pinto
##                           -6.134934523
##                  nameford pinto runabout
##                           -7.588130836
##                          nameford ranger
##                           -2.801022637
##                          nameford torino
##                           -1.828873055
##                             namehi 1200d
##                            1.677783508
##                        namehonda accord
##                            6.911506638
##                     namehonda accord lx
##                            3.176294298
##                 namehonda civic (auto)
##                            2.783045069
##                   namehonda civic 1300
##                            5.343149747
##                   namehonda civic cvcc
##                            8.015311800
##                         namemaxda rx3
##                           -7.706558892
##                         namemazda 626
##                            6.849127056
##                         namemazda glc
##                           19.740091928
##                       namemazda glc 4
```

```
##                                 5.592527766
##                         namemazda glc custom
##                                 2.845116286
##                       namemazda glc custom l
##                                 9.348772411
##                         namemazda glc deluxe
##                                 6.374001203
##                               namemazda rx-4
##                                -2.687190594
##                       namemercedes benz 300d
##                                 4.221839764
##                          namemercury marquis
##                                 0.478409479
##                 namemercury marquis brougham
##                                 1.875432101
##                          namemercury monarch
##                                -6.336194665
##                     namemercury monarch ghia
##                                -0.500678237
##                           namemercury zephyr
##                                -5.060756414
##                         namenissan stanza xe
##                                 8.225489297
##      nameoldsmobile cutlass ciera (diesel)
##                                10.933275727
##                  nameoldsmobile cutlass ls
##                                 6.965357004
##      nameoldsmobile cutlass salon brougham
##                                 4.129756367
##             nameoldsmobile delta 88 royale
##                                -1.125054058
##                       nameoldsmobile omega
##                                -5.524258003
##                nameoldsmobile starfire sx
##                                -4.408149043
##              nameoldsmobile vista cruiser
##                                -0.845135030
##                              nameopel 1900
##                                 1.489433313
##                             nameopel manta
##                                -1.741134193
##                           namepeugeot 304
##                                 4.779038665
##                           namepeugeot 504
##                                 1.433750459
##                         namepeugeot 604sl
##                                -3.535208663
##                      nameplymouth arrow gs
##                                -4.959873584
##                         nameplymouth champ
##                                 3.768240946
##                 nameplymouth custom suburb
##                                 0.500672220
##                        nameplymouth duster
```

```
##                                      -2.371465288
##                          nameplymouth fury iii
##                                       0.683881266
##                        nameplymouth grand fury
##                                       1.430208146
##                           nameplymouth horizon
##                                      -0.221350218
##                         nameplymouth horizon 4
##                                      -0.020726271
##                     nameplymouth horizon miser
##                                       2.433149700
##                       nameplymouth horizon tc3
##                                       0.791534828
##                           nameplymouth reliant
##                                      -3.773311236
##                           nameplymouth sapporo
##                                      -4.733238752
##                  nameplymouth satellite custom
##                                      -4.896847035
##             nameplymouth satellite custom (sw)
##                                      -1.002378003
##                           nameplymouth valiant
##                                      -4.337943414
##                    nameplymouth valiant custom
##                                      -4.502533438
##                            nameplymouth volare
##                                      -2.819272868
##                     nameplymouth volare custom
##                                      -2.935086268
##                 nameplymouth volare premier v8
##                                      -4.613173570
##                              namepontiac astro
##                                      -4.838220005
##                           namepontiac catalina
##                                       2.671039169
##                  namepontiac catalina brougham
##                                       1.541765515
##                           namepontiac firebird
##                                      -2.752197435
##                      namepontiac grand prix lj
##                                       0.179534179
##                            namepontiac phoenix
##                                      -0.531337232
##                         namepontiac phoenix lj
##                                      -2.388853862
##                        namepontiac safari (sw)
##                                       3.147847994
##                      namepontiac sunbird coupe
##                                      -4.349517383
##                         namepontiac ventura sj
##                                      -2.974197254
##                            namerenault 12 (sw)
##                                       0.062202657
##                               namerenault 12tl
```

```
##                                 -1.217511511
##                          namerenault 5 gtl
##                                  6.219312729
##                             namesaab 99gle
##                                 -3.303596199
##                              namesaab 99le
##                                 -0.286456943
##                                  namesubaru
##                                  3.149653337
##                               namesubaru dl
##                                  5.368561705
##                          nametoyota carina
##                                 -1.351921625
##                       nametoyota celica gt
##                                  6.309106718
##              nametoyota celica gt liftback
##                                 -3.065987097
##                         nametoyota corolla
##                                  5.125725339
##                    nametoyota corolla 1200
##                                  8.251003375
##               nametoyota corolla 1600 (sw)
##                                  4.153155392
##                nametoyota corolla liftback
##                                  1.760921483
##                  nametoyota corolla tercel
##                                 10.972702105
##                          nametoyota corona
##                                  2.329295118
##                  nametoyota corona hardtop
##                                  1.769047525
##                 nametoyota corona liftback
##                                  5.549400396
##                  nametoyota corona mark ii
##                                  2.693668113
##                        nametoyota cressida
##                                  3.208149675
##                         nametoyota mark ii
##                                  1.610139343
##                         nametoyota starlet
##                                  9.617103383
##                          nametoyota tercel
##                                  9.753699480
##                      nametriumph tr7 coupe
##                                  6.449396475
##                       namevokswagen rabbit
##                                 -2.616115260
##           namevolkswagen 1131 deluxe sedan
##                                 -0.238497445
##                    namevolkswagen 411 (sw)
##                                 -2.221533078
##                      namevolkswagen dasher
##                                  0.879892898
##                       namevolkswagen jetta
```

```
##                             1.567882850
##                 namevolkswagen model 111
##                             0.303556164
##                    namevolkswagen rabbit
##                            -1.755696942
##             namevolkswagen rabbit custom
##                            -1.278928987
##                  namevolkswagen rabbit l
##                             3.983222811
##              namevolkswagen super beetle
##                            -0.867581108
##                    namevolkswagen type 3
##                            -0.647104821
##                          namevolvo 144ea
##                            -3.768903633
##                      namevolvo 145e (sw)
##                            -4.635183574
##                          namevolvo 244dl
##                            -2.639517363
##                            namevolvo 245
##                            -3.436476126
##                          namevolvo 264gl
##                            -5.251714582
##                         namevolvo diesel
##                             7.548299316
##                   namevw dasher (diesel)
##                            16.315230872
##                            namevw pickup
##                            16.092127557
##                            namevw rabbit
##                            -1.510582792
##                     namevw rabbit custom
##                                      NA
```

```r
m4 = lm(mpg ~ weight + horsepower + displacement, data=Auto, subset=train)
summary(m4)
```

```
##
## Call:
## lm(formula = mpg ~ weight + horsepower + displacement, data = Auto,
##     subset = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.2340  -2.7069  -0.3418   2.2375  16.3002
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  44.5516734  1.4316517  31.119  < 2e-16 ***
## weight       -0.0051554  0.0008627  -5.976  7.2e-09 ***
## horsepower   -0.0437096  0.0150185  -2.910  0.00391 **
## displacement -0.0061969  0.0078486  -0.790  0.43048
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 4.233 on 270 degrees of freedom
## Multiple R-squared:  0.7065, Adjusted R-squared:  0.7032
## F-statistic: 216.6 on 3 and 270 DF,  p-value: < 2.2e-16
```

Qué funciones se pueden aplicar sobre un modelo, como m4?

```
methods(class=class(m4))
```

```
##  [1] add1           alias          anova          case.names
##  [5] coerce         confint        cooks.distance deviance
##  [9] dfbeta         dfbetas        drop1          dummy.coef
## [13] effects        extractAIC     family         formula
## [17] hatvalues      influence      initialize     kappa
## [21] labels         logLik         model.frame    model.matrix
## [25] nobs           plot           predict        print
## [29] proj           qr             residuals      rstandard
## [33] rstudent       show           simulate       slotsFromS3
## [37] summary        variable.names vcov
## see '?methods' for accessing help and source code
```

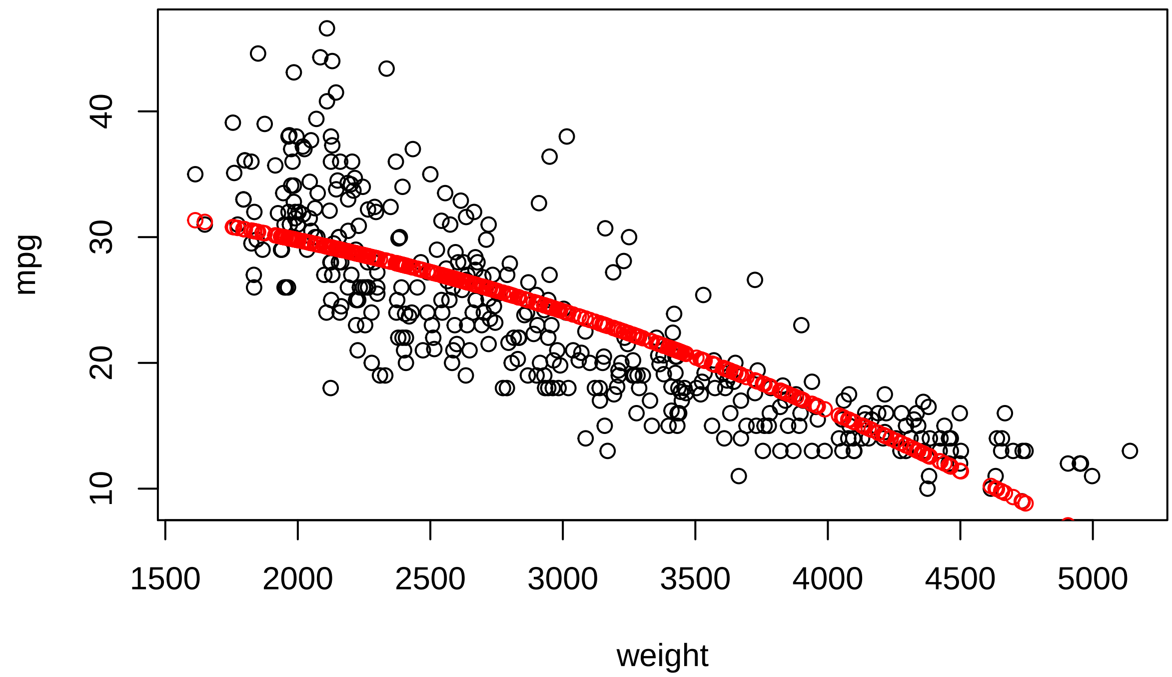
De las gráficas anteriores parece que las relaciones observadas no son lineales ...

Habrá que incorporar algún tipo de transformación no lineal de los atributos ... Por ejemplo, una forma cuadrática

```
m5 = lm(mpg ~ I(weight^2), data=Auto, subset=train)
coef(m5)
```

```
##   (Intercept)   I(weight^2)
##  3.428395e+01 -1.130048e-06
```

```
plot(mpg~weight)
w= m5$coefficients
x = matrix(rep(1, length(weight)),nrow= length(weight))
x= cbind (x, weight^2)
y= apply(x, 1, function(vec) w %*% vec)
points(weight, y, col=2)
```



Con los modelos, podemos obtener predicciones

```
yhatm1Tr = predict(m1) # usa el propio training
yhatm1Tst = predict(m1, auto.test, type= "response")
```

Para ver otras transformaciones p.ej. cúbicas etc.. consultar *poly*(), *log*()

Clasificación

Vamos a convertir el problema en un problema de clasificación binaria Se crea una variable binaria, mpg01

```
Auto2 = data.frame(mpg01 = (ifelse(mpg<median(mpg),0,1)),Auto)
```

Particionar el conjunto en training y test

```
set.seed(1)
train = sample (nrow(Auto), round(nrow(Auto)*0.7)) # nos quedamos con los indices para el training
Auto.train = Auto[train,]  # podemos reservarlos aparte ... aunque con el subset no sería necesario :)
Auto.test = Auto[-train,]
```

Se ajusta un modelo lineal, por ejemplo de regresión logística para predecir *mpg*01. Se puede especificar de forma explícita los atributos a considerar a la hora de construir el modelo, el resto se ignoran.

```
ml1 = glm(mpg01 ~ weight + horsepower + displacement,
  family = binomial(logit), data = Auto2, subset=train)
summary(ml1)
```

```
##
## Call:
## glm(formula = mpg01 ~ weight + horsepower + displacement, family = binomial(logit),
```

```
##     data = Auto2, subset = train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.31258  -0.28359  -0.00467   0.39442   3.13796
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  11.6385137  1.8444083   6.310 2.79e-10 ***
## weight       -0.0020599  0.0008214  -2.508   0.0122 *
## horsepower   -0.0454861  0.0151603  -3.000   0.0027 **
## displacement -0.0081012  0.0060085  -1.348   0.1776
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 379.79  on 273  degrees of freedom
## Residual deviance: 158.16  on 270  degrees of freedom
## AIC: 166.16
##
## Number of Fisher Scoring iterations: 7
```

Una vez aprendido, veamos cómo predice…

```r
#Calculo de probabilidades
probTr.ml1 = predict(ml1, type="response")
probTstml1 = predict(ml1, data.frame(Auto2[-train,-1]), type="response")
```

predicciones con el modelo de regresión logística

```r
predTstml1 = rep(0, length(probTstml1))  # predicciones por defecto 0
predTstml1[probTstml1 >=0.5] = 1          # >= 0.5 clase 1

table(predTstml1, Auto2[-train,1])  # para el calculo del Eval
```

```
##
## predTstml1  0  1
##          0 50  3
##          1  7 58
```

```r
Eval = mean(predTstml1 != Auto2[-train,1])
cat("Eval con el modelo LR "); print(ml1$call)
```

```
## Eval con el modelo LR
```

```
## glm(formula = mpg01 ~ weight + horsepower + displacement, family = binomial(logit),
##     data = Auto2, subset = train)
```

```r
print(Eval)
```

```
## [1] 0.08474576
```

se obtiene el Etest, para obtener el Ein?

Otras familias de funciones …

```r
ml2 = glm(mpg01 ~ weight + horsepower + displacement,
  family = gaussian(identity), data = Auto2, subset=train)
```

```r
summary(ml2)
```

```
##
## Call:
## glm(formula = mpg01 ~ weight + horsepower + displacement, family = gaussian(identity),
##     data = Auto2, subset = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9242  -0.2369   0.0808   0.2052   0.9833
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.567e+00  1.131e-01  13.847  < 2e-16 ***
## weight       -2.641e-04  6.817e-05  -3.875 0.000134 ***
## horsepower    3.488e-04  1.187e-03   0.294 0.769080
## displacement -1.609e-03  6.202e-04  -2.595 0.009977 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1119079)
##
##     Null deviance: 68.485  on 273  degrees of freedom
## Residual deviance: 30.215  on 270  degrees of freedom
## AIC: 183.47
##
## Number of Fisher Scoring iterations: 2
```

Para el preprocesamiento Centrado, escalado, transformación para reducir la asimetria Vamos a trabajar con el dataset segmentationOriginal que trata de
Cell Body Segmentation problema de clasificación , células Pobremente segmentadas o Well segmentadas.

```r
library("AppliedPredictiveModeling", lib.loc="~/R/x86_64-redhat-linux-gnu-library/3.3")
library(help=AppliedPredictiveModeling)
data("segmentationOriginal")
class(segmentationOriginal)
```

```
## [1] "data.frame"
```

```r
names(segmentationOriginal)[1:10]
```

```
##  [1] "Cell"           "Case"           "Class"          "AngleCh1"
##  [5] "AngleStatusCh1" "AreaCh1"        "AreaStatusCh1"  "AvgIntenCh1"
##  [9] "AvgIntenCh2"    "AvgIntenCh3"
```

```r
summary(segmentationOriginal[1:10])
```

```
##       Cell                Case       Class        AngleCh1
##  Min.   :207827637   Test :1010   PS:1300   Min.   :  0.03088
##  1st Qu.:208332462   Train:1009   WS: 719   1st Qu.: 53.89221
##  Median :208384321                          Median : 90.58877
##  Mean   :208402392                          Mean   : 90.49340
##  3rd Qu.:208405230                          3rd Qu.:126.68201
##  Max.   :210964110                          Max.   :179.93932
##  AngleStatusCh1       AreaCh1        AreaStatusCh1       AvgIntenCh1
##  Min.   :0.0000   Min.   : 150.0   Min.   :0.00000   Min.   :  15.16
```

```
##  1st Qu.:0.0000   1st Qu.: 193.0   1st Qu.:0.00000   1st Qu.:  35.36
##  Median :0.0000   Median : 253.0   Median :0.00000   Median :  62.34
##  Mean   :0.5686   Mean   : 320.3   Mean   :0.08024   Mean   : 126.07
##  3rd Qu.:1.0000   3rd Qu.: 362.5   3rd Qu.:0.00000   3rd Qu.: 143.19
##  Max.   :2.0000   Max.   :2186.0   Max.   :1.00000   Max.   :1418.63
##   AvgIntenCh2     AvgIntenCh3
##  Min.   :  0.0   Min.   :   0.12
##  1st Qu.: 44.0   1st Qu.:  33.50
##  Median :172.5   Median :  67.43
##  Mean   :188.1   Mean   :  96.42
##  3rd Qu.:278.3   3rd Qu.: 127.34
##  Max.   :988.5   Max.   :1205.51
```

```
segData = subset(segmentationOriginal, Case == "Train")
head(segData[1:10])
```

```
##         Cell  Case Class  AngleCh1 AngleStatusCh1 AreaCh1 AreaStatusCh1
## 2  207932307 Train    PS 133.75204              0     819             1
## 3  207932463 Train    WS 106.64639              0     431             0
## 4  207932470 Train    PS  69.15032              0     298             0
## 12 207932484 Train    WS 109.41643              0     256             0
## 15 207932459 Train    PS 104.27865              0     258             0
## 16 207827779 Train    PS  77.99194              0     358             0
##    AvgIntenCh1 AvgIntenCh2 AvgIntenCh3
## 2     31.92327    205.8785    69.91688
## 3     28.03883    115.3155    63.94175
## 4     19.45614    101.2947    28.21754
## 12    18.82857    125.9388    13.60000
## 15    17.57085    124.3684    22.46154
## 16    42.28363    217.1316    42.32164
```

```
dim(segData)
```

```
## [1] 1009  119
```

```
dim(segmentationOriginal)
```

```
## [1] 2019  119
```

```
cellID = segData$Cell
#unique(cellID)
cellClass = segData$Class
unique(cellClass)
```

```
## [1] PS WS
## Levels: PS WS
```

```
cellcase = segData$Case
unique(cellcase)
```

```
## [1] Train
## Levels: Test Train
```

```
segData = segData[, -c(1:3)]
```

Se eliminan parte de la información, columnas redundantes … Todas aquellas que contengan status …

```
length(grep("Status", names(segData)))
```

```
## [1] 58
```

```r
b = (grep("Status", names(segData)))
segData = segData[,-b]
dim(segData)
```

```
## [1] 1009   58
```

```r
names(segData)
```

```
##  [1] "AngleCh1"                "AreaCh1"
##  [3] "AvgIntenCh1"             "AvgIntenCh2"
##  [5] "AvgIntenCh3"             "AvgIntenCh4"
##  [7] "ConvexHullAreaRatioCh1"  "ConvexHullPerimRatioCh1"
##  [9] "DiffIntenDensityCh1"     "DiffIntenDensityCh3"
## [11] "DiffIntenDensityCh4"     "EntropyIntenCh1"
## [13] "EntropyIntenCh3"         "EntropyIntenCh4"
## [15] "EqCircDiamCh1"           "EqEllipseLWRCh1"
## [17] "EqEllipseOblateVolCh1"   "EqEllipseProlateVolCh1"
## [19] "EqSphereAreaCh1"         "EqSphereVolCh1"
## [21] "FiberAlign2Ch3"          "FiberAlign2Ch4"
## [23] "FiberLengthCh1"          "FiberWidthCh1"
## [25] "IntenCoocASMCh3"         "IntenCoocASMCh4"
## [27] "IntenCoocContrastCh3"    "IntenCoocContrastCh4"
## [29] "IntenCoocEntropyCh3"     "IntenCoocEntropyCh4"
## [31] "IntenCoocMaxCh3"         "IntenCoocMaxCh4"
## [33] "KurtIntenCh1"            "KurtIntenCh3"
## [35] "KurtIntenCh4"            "LengthCh1"
## [37] "NeighborAvgDistCh1"      "NeighborMinDistCh1"
## [39] "NeighborVarDistCh1"      "PerimCh1"
## [41] "ShapeBFRCh1"             "ShapeLWRCh1"
## [43] "ShapeP2ACh1"             "SkewIntenCh1"
## [45] "SkewIntenCh3"            "SkewIntenCh4"
## [47] "SpotFiberCountCh3"       "SpotFiberCountCh4"
## [49] "TotalIntenCh1"           "TotalIntenCh2"
## [51] "TotalIntenCh3"           "TotalIntenCh4"
## [53] "VarIntenCh1"             "VarIntenCh3"
## [55] "VarIntenCh4"             "WidthCh1"
## [57] "XCentroid"               "YCentroid"
```

Eliminar las variables con varianza 0 o muy próximas, esto es muy desbalanceadas o de valor único.

```r
#nearZeroVar(segData)
```

Transformación de atributos asimétricos, necesarios para la aplicación de algunos métodos de aprendizaje sensibles a distancias. Se consideran asimétricos cuando o bien la ratio entre range >20 o bien el valor skewness se aleja de 0.

$$skewness = \frac{\sum (x_i - mean(x))^3}{(n-1)v^3/2}$$

Una familia de funciones para la transformación son las propuestas por Box y Cox (1964) un parámetro nota como λ:

$\frac{x^\lambda - 1}{\lambda}$ si $\lambda \neq 0$

o bien

$log(x)$ si $\lambda = 0$

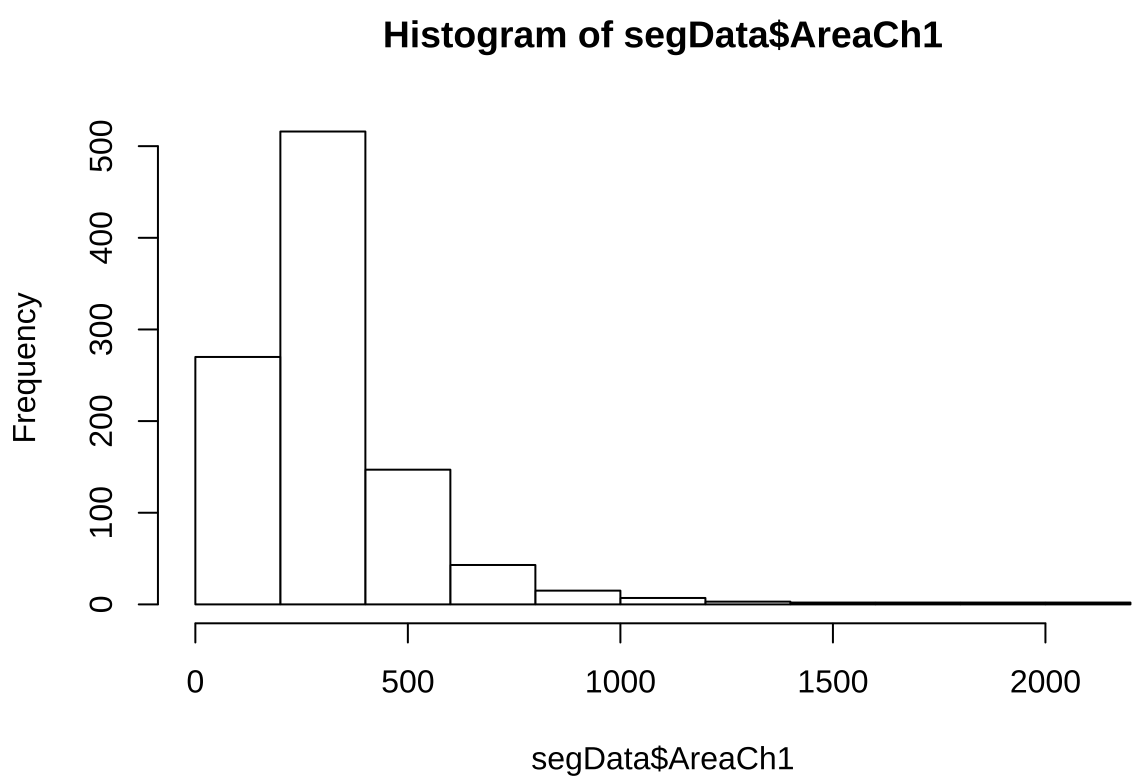
```
hist(segData$AreaCh1)
library("e1071", lib.loc="~/R/x86_64-redhat-linux-gnu-library/3.3")
?skewness
skewness(segData$AreaCh1)
```

```
## [1] 3.525107
```

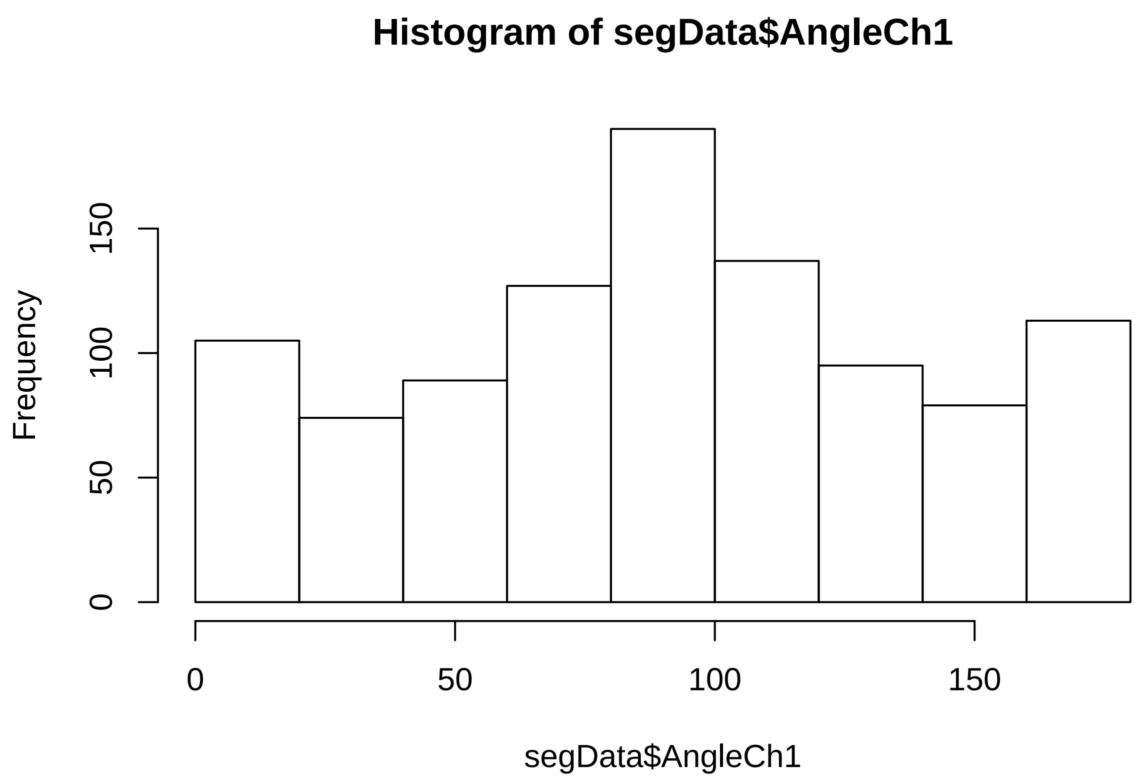
```
skewness(segData$AngleCh1)
```

```
## [1] -0.02426252
```

```
hist (segData$AreaCh1)
```




```
hist(segData$AngleCh1)
```




```r
range(segData$AreaCh1)
```

```
## [1]  150 2186
```

```r
range(segData$AngleCh1)
```

```
## [1]   0.03087639 179.93932283
```

```r
v_asimetria = apply(segData,2,skewness)
v_asimetria[1:15]
```

```
##                AngleCh1                 AreaCh1             AvgIntenCh1
##             -0.02426252              3.52510745              2.95918524
##             AvgIntenCh2             AvgIntenCh3             AvgIntenCh4
##              0.84816033              2.20234214              1.90047128
##  ConvexHullAreaRatioCh1 ConvexHullPerimRatioCh1     DiffIntenDensityCh1
##              2.47658194             -1.30409896              2.76047338
##     DiffIntenDensityCh3     DiffIntenDensityCh4         EntropyIntenCh1
##              2.08518782              1.89923287              0.39789483
##         EntropyIntenCh3         EntropyIntenCh4           EqCircDiamCh1
##             -1.00295336             -0.82790492              1.95553035
```

```r
sort(abs(v_asimetria), decreasing = T)[1:10]
```

```
##          KurtIntenCh1           KurtIntenCh4 EqEllipseProlateVolCh1
##             12.859648               6.918503               6.070834
##         EqSphereVolCh1           KurtIntenCh3  EqEllipseOblateVolCh1
##              5.739502               5.505611               5.489313
##          TotalIntenCh1         EqSphereAreaCh1                AreaCh1
```

```
##               5.399604               3.525140               3.525107
##   IntenCoocContrastCh4
##               3.470305
```

Con *skewness()* se halla pero no se aplica la transformación, para ello:

```r
#install.packages("caret")
library("caret", lib.loc="~/R/x86_64-redhat-linux-gnu-library/3.3")
```

```
## Loading required package: lattice

## Loading required package: ggplot2

##
## Attaching package: 'ggplot2'

## The following object is masked from 'Auto':
##
##     mpg
```

```r
BoxCoxTrans(segData$KurtIntenCh1) # no transformacion
```

```
## Box-Cox Transformation
##
## 1009 data points used to estimate Lambda
##
## Input data summary:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -1.40300 -0.54320 -0.03541  0.90500  0.88470 97.42000
##
## Lambda could not be estimated; no transformation is applied
```

```r
BoxCoxTrans(segData$AngleCh1)  # no transformacion
```

```
## Box-Cox Transformation
##
## 1009 data points used to estimate Lambda
##
## Input data summary:
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##   0.03088  54.66000  90.03000  91.13000 127.90000 179.90000
##
## Largest/Smallest: 5830
## Sample Skewness: -0.0243
##
## Estimated Lambda: 0.8
## With fudge factor, no transformation is applied
```

```r
BoxCoxTrans(segData$AreaCh1)
```

```
## Box-Cox Transformation
##
## 1009 data points used to estimate Lambda
##
## Input data summary:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   150.0   194.0   256.0   325.1   376.0  2186.0
##
## Largest/Smallest: 14.6
```

```
## Sample Skewness: 3.53
##
## Estimated Lambda: -0.9
```

```
Ch1Area_trans = BoxCoxTrans(segData$AreaCh1)

head(segData$AreaCh1)
```

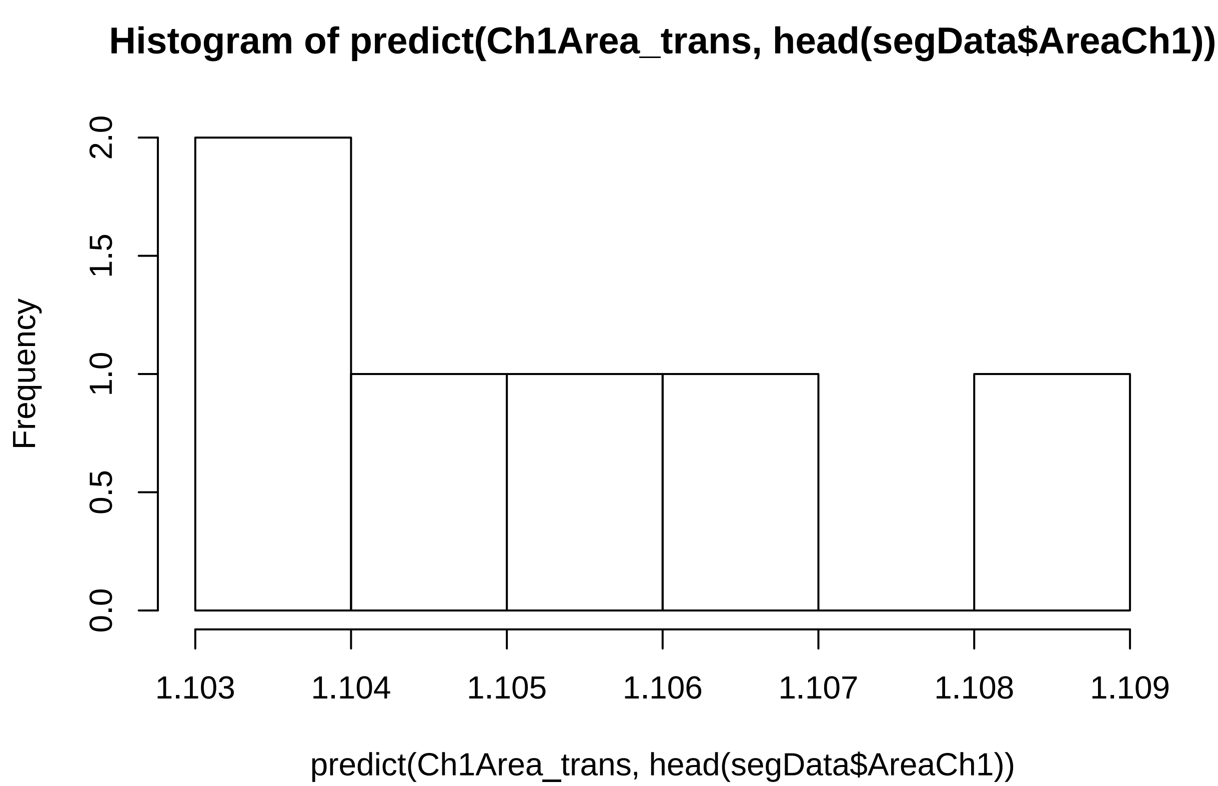
```
## [1] 819 431 298 256 258 358
```

```
# head(Ch1Area_trans) no funciona es necesario aplicar la formula mediante predict
predict(Ch1Area_trans,head(segData$AreaCh1))
```

```
## [1] 1.108458 1.106383 1.104520 1.103554 1.103607 1.105523
```

es justo,la transformación con $lambda = -0.9$ Datos transformados:

```
hist(predict(Ch1Area_trans,head(segData$AreaCh1)))
```



Demasiados atributos Algunos redundantes o irrelevantes, es conveniente reducir dimensionalidad. El algoritmo PCA (principal components analysis) es un filtro no supervisado.

```
pcaObject = prcomp(segData,center = TRUE, scale. = TRUE)
attributes(pcaObject)
```

```
## $names
## [1] "sdev"     "rotation" "center"   "scale"    "x"
##
## $class
## [1] "prcomp"
```

```r
head(pcaObject$center)
```

```
##    AngleCh1     AreaCh1 AvgIntenCh1 AvgIntenCh2 AvgIntenCh3 AvgIntenCh4
##    91.12641   325.12587   127.91503   185.19067    96.12917   140.02605
```

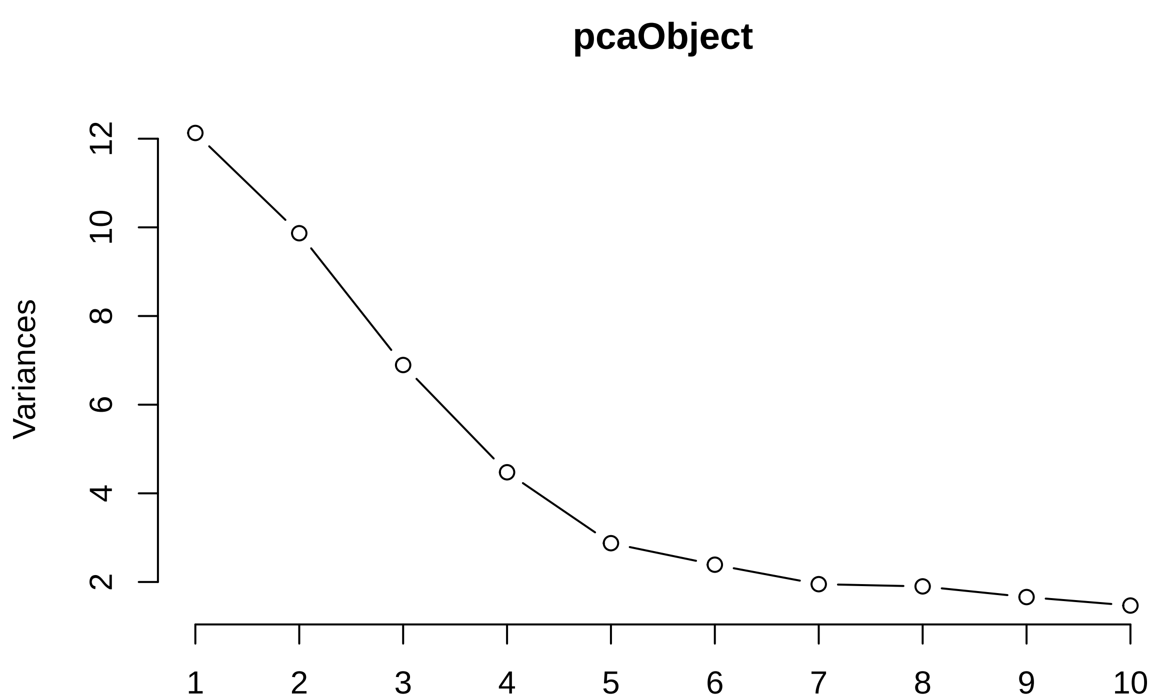
```r
porcentVariance = pcaObject$sd^2/sum(pcaObject$sd^2)*100
porcentVariance[1:4]
```

```
## [1] 20.912359 17.013300 11.886892  7.715243
```

```r
head(pcaObject$x[, 1:5])
```

```
##           PC1        PC2         PC3       PC4        PC5
## 2   5.0985749  4.5513804 -0.03345155 -2.640339  1.2783212
## 3  -0.2546261  1.1980326 -1.02059569 -3.731079  0.9994635
## 4   1.2928941 -1.8639348 -1.25110461 -2.414857 -1.4914838
## 12 -1.4646613 -1.5658327  0.46962088 -3.388716 -0.3302324
## 15 -0.8762771 -1.2790055 -1.33794261 -3.516794  0.3936099
## 16 -0.8615416 -0.3286842 -0.15546723 -2.206636  1.4731658
```

```r
plot(pcaObject,type="l")
```




```r
head(pcaObject$rotation[, 1:5])
```

```
##                      PC1         PC2         PC3         PC4        PC5
## AngleCh1     0.001213758 -0.01284461 0.006816473 -0.02755720 0.02523673
## AreaCh1      0.229171873  0.16061734 0.089811727 -0.05523062 0.05273468
## AvgIntenCh1 -0.102708778  0.17971332 0.067696745  0.18675619 0.02401245
```

```
## AvgIntenCh2 -0.154828672  0.16376018  0.073534399  0.04145772  0.07839174
## AvgIntenCh3 -0.058042158  0.11197704 -0.185473286  0.28291291 -0.07822440
## AvgIntenCh4 -0.117343465  0.21039086 -0.105060977  0.01116373  0.04990515
```

Por filas vemos los atributos que forman parte de cada uno de los componentes y sus coeficientes. Por defecto la función selecciona aquellos componentes que explican hasta el 95% de la variabilidad de los datos… se puede cambiar con argumentos thresh.

A la hora de aplicar las transformaciones, en *caret* existe una función **preProcess()** que realiza todas transformaciones mencionadas de forma ordenada.

```r
ObjetoTrans =  preProcess(segData, method = c("BoxCox", "center", "scale", "pca"),thres=0.8)
ObjetoTrans
```

```
## Created from 1009 samples and 58 variables
##
## Pre-processing:
##   - Box-Cox transformation (47)
##   - centered (58)
##   - ignored (0)
##   - principal component signal extraction (58)
##   - scaled (58)
##
## Lambda estimates for Box-Cox transformation:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -2.00000 -0.50000 -0.10000  0.05106  0.30000  2.00000
##
## PCA needed 10 components to capture 80 percent of the variance
```

Para obtener un nuevo conjunto de datos, se aplican

```r
segTrans = predict(ObjetoTrans,segData)
dim(segTrans)
```

```
## [1] 1009   10
```