fac-simile1

Tommaso Farneti

2023-05-05

EXERCISE 1

Consider the data from the following shell.Rdata (from uci public archive) concerning a sample of marine abalones. We aim to explain age (eta in years) considering the following covariates: length (lun), diameter (dia), height (alt), total weight (pesot), weight without shell (pesom), viscera weight (pesov) and the weight of the shell after drying (pesog).

.1 Illustrate the sample data through descriptive statistics commenting on the results concerning the applicative context.

.2 Depict the empirical cumulative distribution function of the observed age and add the theoretical cumulative distribution function of the univariate Gaussian distribution with mean and variance equal to the same quantities. Comment on the figure.

.3 Fit a multiple linear regression and comment on each quantity of the output the summary function provides. How do you judge these results? What do you suggest for improving them?

.4 Describe the problem of multicollinearity. Do you think it can be present in this case? (respond without using the code)

1.1: First, we load the data and we illustrate the sample through descriptive statistics:

```
load('shell.RData')
require(skimr)

## Loading required package: skimr

skimr::skim_without_charts(shell)
```

Data summary

Name	shell
Number of rows	300
Number of columns	8

_____

Column type frequency:	
numeric	8

_____

Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
eta	0	1	11.33	3.24	4.50	9.50	11.50	12.50	24.50
lun	0	1	0.52	0.12	0.13	0.45	0.54	0.60	0.72
dia	0	1	0.40	0.10	0.10	0.35	0.42	0.48	0.60
alt	0	1	0.14	0.04	0.01	0.11	0.14	0.17	0.23
pesot	0	1	0.82	0.49	0.01	0.45	0.81	1.14	2.27
pesom	0	1	0.36	0.22	0.00	0.18	0.33	0.50	1.09
pesov	0	1	0.18	0.11	0.00	0.09	0.17	0.25	0.45
pesog	0	1	0.24	0.14	0.00	0.13	0.24	0.32	0.68

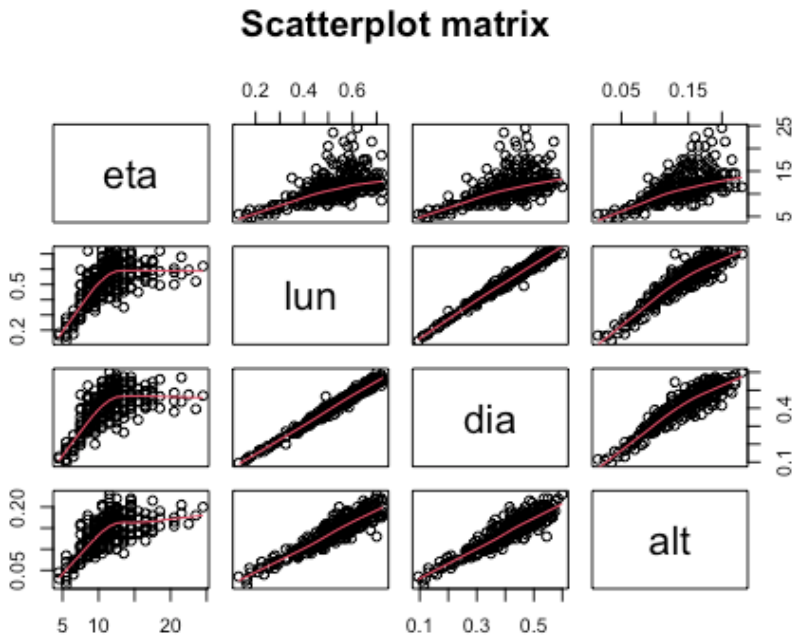
The variable age, which is the most variable (highest standard deviation), has an average value of 11.3 years, very close to the median value. From the quartiles seems that the variable age has a symmetric distribution, and goes from a minimum of 4.5 years old to a maximum of 24.5.

The three “dimension” variables are length (lun), diameter (dia) and height (alt). The most variable of the three is the length one, however all the three variables present low standard deviation. Length goes from a minimum of 0.13 to a maximum of 0.725, diameter goes from a minimum of 0.095 to a maximum of 0.6 and height goes from a minimum of 0.04 to a maximum of 0.23. So, between these three variables, the one which present the greatest range is the diameter. All these three variables have a mean value similar to the median one, so they seem to be pretty symmetrical.

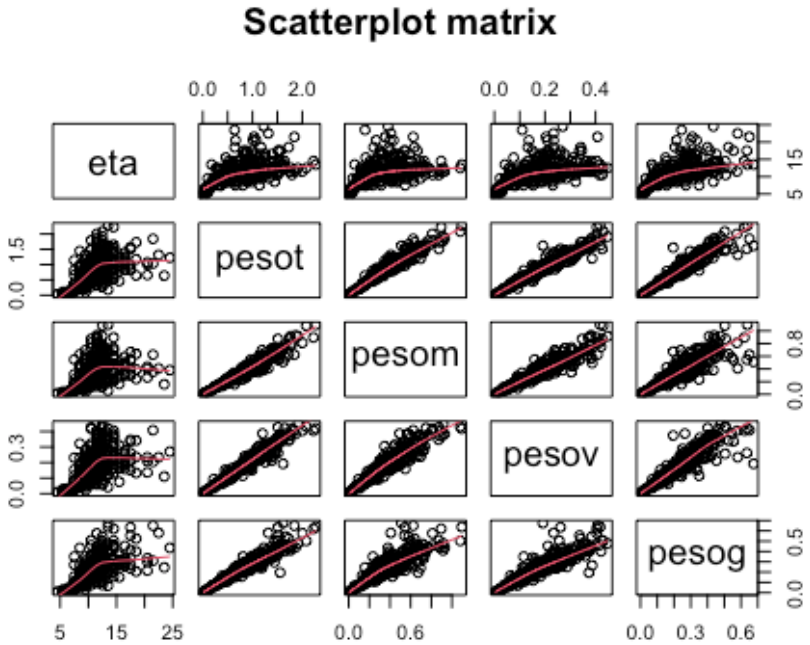
Talking about the four “weight” variables, they are: total weight (pesot), weight without shell (pesom), viscera weight (pesov) and the weight of the shell after drying (pesog). Between these four, the most variable is the total weight variable. Also these four variables have a mean value similar to the median one and they seem to have a symmetrical distribution. We can notice how the total weight variable goes from a minimum of 0.11 to a maximum of 2.27.

Since some columns are referred to a similar characteristic of the abolone, we plot the scatterplot matrices, to check the correlations between the variables. We divide the scatterplot between the “dimension” variables and the “weight” variables.

```
pairs(shell[,1:4],
panel = panel.smooth,
      main = "Scatterplot matrix ")
```



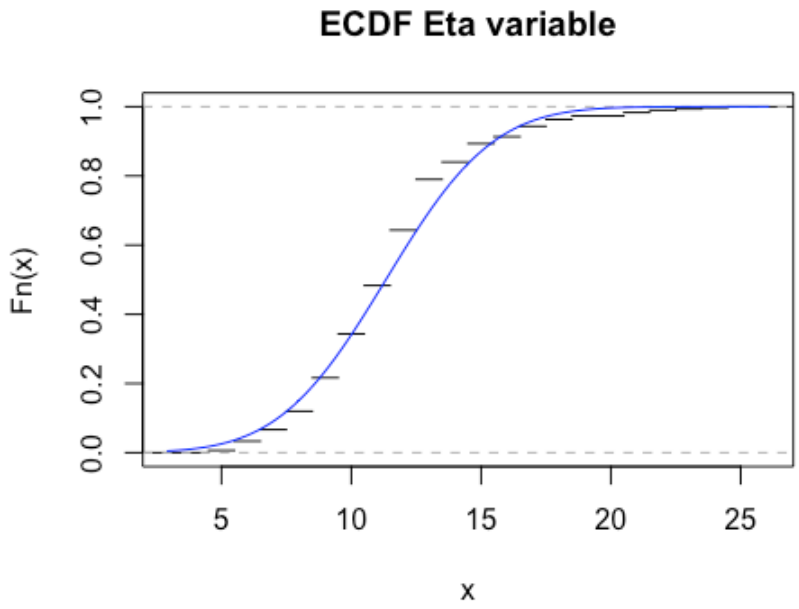

```
pairs(shell[,c(1,5:8)],
panel = panel.smooth,
      main = "Scatterplot matrix ")
```



We can notice a high correlation between these variables. It is very likely we will have a multicollinearity problem.

1.2: Now we depict the empirical cumulative distribution function of the observed age, adding the theoretical cumulative distribution function of the univariate Gaussian distribution with mean and variance equal to the same quantities.

```
plot(ecdf(shell$eta),
     do.points = FALSE,
     main = "ECDF Eta variable")
curve(pnorm(x, mean(shell$eta), sd(shell$eta)),
      col = "blue",
      add = TRUE)
```



Obviously the variable age can't take negative values. The graphical representation of the empirical cumulative distribution function provides information on the possible form of the distribution of the population. In this case, there are no significant deviations observed between the empirical and theoretical functions, which could confirm that the random variable in the reference population is normally distributed. However, the graphical analysis of the empirical distribution function is generally not very sensitive to subtle deviations from the reference distribution (normal in this case), and therefore a further examination of the data with other tools (like QQ-plot) is appropriate.

1.3: Fit a multiple linear regression and comment on each quantity of the output the summary function provides. How do you judge these results? What do you suggest for improving them?

```
lm <- lm(eta ~ ., shell)
summary(lm)

##
## Call:
## lm(formula = eta ~ ., data = shell)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5848 -1.1779 -0.2789  0.8193  8.2534
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.7049     0.9036   2.994  0.00299 **
## lun           9.0275     7.0593   1.279  0.20198
## dia           1.9965     8.2398   0.242  0.80872
## alt          26.5671     8.9494   2.969  0.00324 **
## pesot        11.5984     2.8257   4.105 5.26e-05 ***
## pesom       -22.8850     3.1683  -7.223 4.43e-12 ***
## pesov       -12.2573     5.0720  -2.417  0.01628 *
## pesog         1.1417     4.3483   0.263  0.79307
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.164 on 292 degrees of freedom
## Multiple R-squared:  0.5657, Adjusted R-squared:  0.5553
## F-statistic: 54.34 on 7 and 292 DF,  p-value: < 2.2e-16
```

Be observe that the range of the residuals is not too wide, but still it indicates the for some statistical units, the deviation between the observed value and the value predicted by the model is quite high. The median value is slightly different from 0 (as expected based on theoretical assumptions). The quartiles are equidistant from the center of the distribution, which may suggest symmetry. Further analysis of the residuals will be necessary to verify the validity of the assumptions underlying the model.

The estimated coefficient $\widehat{\beta_0}$ for the intercept represents the expected value of the response variable when all explanatory variables have a value of zero. Frequently (including in this case), this value may not have practical interpretability; it would be non-sense to consider the age of a marine abalone with null height and weight. However, the null hypothesis H_0: $\beta_0 = 0$ is rejected at any level of significance.

The estimated coefficient $\widehat{\beta_1}$ for the "lun" variable represents the expected increase of the response variable (eta) for a unit increase in the length of the abalone, while holding the remaining covariates constant. This means that, on average, increasing of 1 unit the length of an abalone correspond to an increment of 9 years of age. A similar interpretation can be given to the "dia" variable. The estimated coefficient $\widehat{\beta_2}$ tells us that, on average, increasing of 1 unit the

diameter of an abalone correspond to an increment of 1.99 years of age. However, these two variables are not statistically significant, so they are not significantly different from 0.

The estimated coefficient $\widehat{\beta_3}$ tells us that, on average, increasing of 1 unit the height of an abalone correspond to an increment of 26.6 years of age. This seems strange, and tells us that probably we have a multicollinearity problem. In fact, multicollinearity causes many problems, including the fact that the coefficient estimates can swing wildly based on which other independent variables are in the model. The coefficients become very sensitive to small changes in the model.

…. all the other coefficients

The residual standard error (RSE), which is 2.164, along with its degrees of freedom: 292, (n-p-1 = 300-7-1 = 292), where p represents the number of covariates and one is subtracted if the model contains an intercept. This is the square root of the ratio of the sum of squares of residuals and the number of degrees of freedom. The RSE can be interpreted as the average deviation around the mean of residuals, which is assumed to be zero, and thus as the average deviation between observed and corresponding interpolated values. In other words, this value states that, on average, interpolated values deviate from observed ones by 2.2. Additionally, the percentage error can be obtained: it is sufficient to take the ratio with the sample mean value of the score: 2.164/11.3 = 0.19, hence an error of 19%.

The multiple linear determination coefficient represents the ratio of the variance explained by the interpolating plane and the total variance of the response variable. In this case, a value of 0.57 is observed, indicating that the interpolating plane explains approximately 57% of the variability of the initial salary. Note that this is a goodness-of-fit index, which cannot detect whether the model has been correctly specified.

The results of the F-test, i.e., the test statistic value, which is 54.34, and the corresponding p-value (of the order of 10−16). The proximity to 0 of the p-value allows us to reject the null hypothesis that all regression coefficients, except for the intercept, are equal to 0.

1.4: As we saw previously, the explanatory variables are strongly linearly linked to each other. This phenomena is called collinearity, and it can leads to additional difficulty in separating the individual effect of the explanatory variables on the response. In fact, in order for the model to provide a reasonable estimated X′X must be invertible (so the matrix of covariates must be full-column rank). But when there is perfect multicollinearity, then the X′X matrix is non-invertible and it implies that there are no unique least-squares estimates. When the matrix is invertible, but there is not enough variation in the data, that is, the columns are almost linearly dependent, it implies that standard errors are relatively large (they are inflated) and the statistics like the t-test are inappropriate. The signs of the coefficients can be the opposite of what intuition about the effect of the predictor might suggest. Collinearity can be detected in several ways: examination of the correlation matrix of the predictors may reveal large pairwise collinearities.

EXERCISE 2

Consider the data from the estimated velocity veloc.Rdata (km/sec, simulated data) of a sample of galaxies in a certain region of the space. The interest lies in the identification of subgroups of galaxies with similar structures of pattern expansion.

.1 Depict appropriate explanatory plots for the data and describe the figures.

.2 Fit a finite mixture model performing model selection for the number of clusters and the clustering structure. Comment on the results with reference to the applied context.

.3 Print the results of the best model selected in the previous step and comment on them. What can we conclude about galaxies' expansion?

.4 Describe the bootstrap procedure and explain its purpose (respond without using the code).

2.1:

```
load("veloc.RData")
skimr::skim_without_charts(veloc)
```

Data summary

Name	veloc
Number of rows	32
Number of columns	1

_____

Column type frequency:	
numeric	1

_____

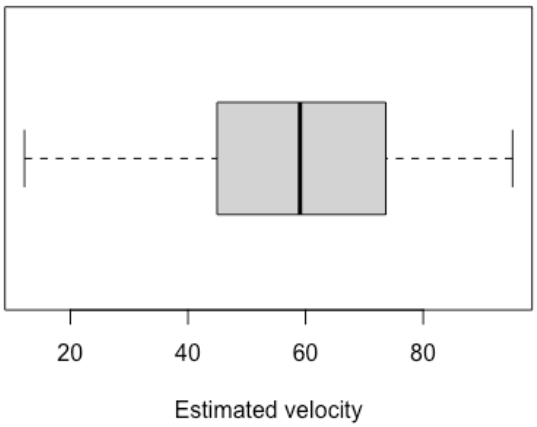
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
data	0	1	58.77	21.14	12.2	45.37	59.05	73.58	95.2

First of all we take a look at the boxplot:

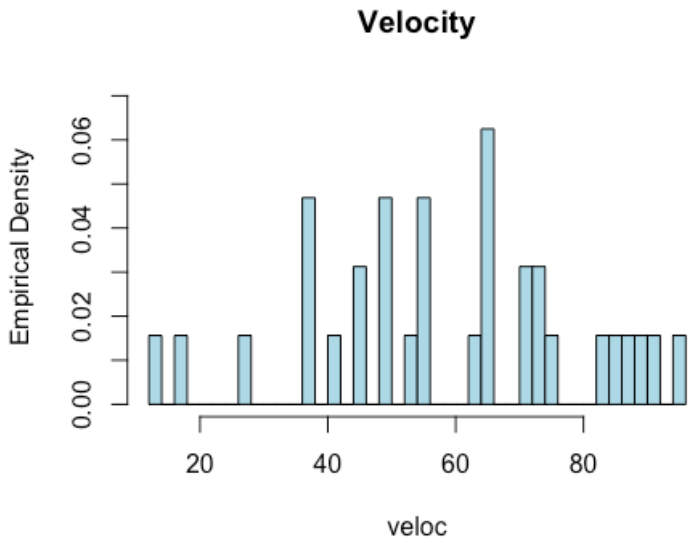
```
boxplot(veloc, xlab = "Estimated velocity", horizontal = TRUE)
```



Median and mean values are very close to each other and by looking also at the first and third quartile the distribution should not be skewed. The galaxy with the minimum estimated velocity present a value of 12.2 km/s, while the one with the maximum value has a velocity of 95.2 km/s. The value of the standard deviation is 21.14, which indicates the average distance between the sample values from the arithmetic mean.

We can also take a look at the histogram:

```
hist(veloc, breaks = 50,
     ylim = c(0, 0.07),
     main = "Velocity",
     ylab=" Empirical Density", freq =FALSE,
     col = "lightblue")
```



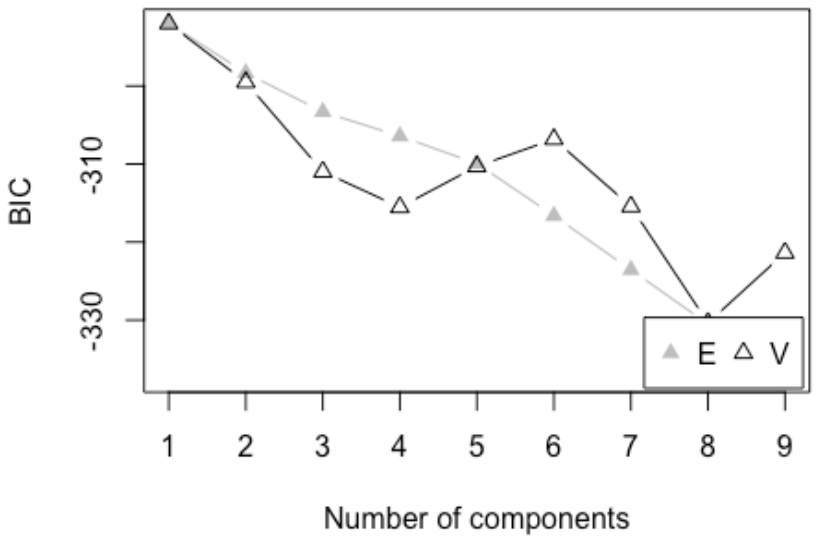
2.2: Fit a finite mixture model performing model selection for the number of clusters and the clustering structure. Comment on the results with reference to the applied context.

The choice of the number of components and model specification is performed using the mclustBIC() function. In the univariate case, among the possible model specifications, we focus on the assumption of spherical components (assuming that there is no correlation between variables). The encoding for this situation is E (equal variance) or V (variable variance).

```
require(mclust)

## Loading required package: mclust

## Package 'mclust' version 6.0.0
## Type 'citation("mclust")' for citing this R package in publications.

sel <- mclustBIC(veloc)
sel

## Bayesian Information Criterion (BIC):
##           E         V
## 1 -291.9992 -291.9992
## 2 -298.3368 -299.5150
## 3 -303.3167 -310.9852
## 4 -306.4282 -315.5288
## 5 -309.9439 -310.3042
## 6 -316.6261 -306.8127
## 7 -323.5604 -315.4417
## 8 -330.4907 -330.3440
## 9 -337.4117 -321.3681
##
## Top 3 models based on the BIC criterion:
##       E,1       V,1       E,2
## -291.9992 -291.9992 -298.3368
```

The output shows the results in matrix form, with the number of components and the model specification on the columns. Considering the top three models from the perspective of the BIC index, it is found that the best choice is to select one component, but it does not make any difference assuming common or variable variability between the sample observations. It is also possible to graphically represent the results obtained for the BIC of the different models by simply using the plot() function:

```
plot(sel)
```



We then estimate the selected model and comment on its parameters. The function used is mclust::Mclust(), which requires as input the data, the number of components, and the specification for the variance of the model.

```
mod <- Mclust(veloc,
              G = 1,
              modelNames = "E")
summary(mod)

## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust X (univariate normal) model with 1 component:
##
##  log-likelihood  n df       BIC       ICL
##       -142.5339 32  2 -291.9992 -291.9992
##
## Clustering table:
##  1
## 32
```

The summary of the model only shows information related to the value of the log-likelihood at convergence (equal to -142.53), BIC and ICL indices, number of observations and degrees of freedom; the latter are computed as the number of free parameters estimated by the model. The output also provides the number of observations classified in the group. We observe that the group contains all the observations.

To obtain more information (on the parameters of the model), it is necessary to specify the option parameters=TRUE.

```
summary(mod, parameters = TRUE)

## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust X (univariate normal) model with 1 component:
##
##  log-likelihood  n df       BIC       ICL
##       -142.5339 32  2 -291.9992 -291.9992
##
## Clustering table:
##  1
## 32
##
## Mixing probabilities:
## 1
## 1
##
## Means:
## [1] 58.76563
##
## Variances:
## [1] 432.8929
```

s.86 - interpretation of the results (but with at least two components). What if there is only one selected?

2.4: Describe the bootstrap procedure and explain its purpose (respond without using the code).

Bootstrap is introduced as a basic tool for inference since it may assess estimation accuracy of any estimator (and algorithm) no matter how complicated. We know that in statistics a measure of accuracy of an estimate is provided by the associated standard error. Even if the statistic has an approximately normal sampling distribution, without the standard error we cannot use the confidence interval formula of a point estimate. Since there are many estimators (and algorithms) not having a direct mathematical formula to calculate standard errors the bootstrap method was proposed. The advantage of the bootstrap over the maximum likelihood is that it allows us to compute maximum likelihood estimates of standard errors and other quantities in settings where no mathematical formulas are available…