TP 1 DM en C y T

Jairo Jiménez, Sergio De Raco, Diego Acosta

8 de septiembre de 2015

```
library(knitr)
library(ggplot2)
library(reshape)
```

```
setwd("/run/media/ahriman/Stuff/MDMKD/Segundo Cuatrimestre/CYT/TP 1")
glx = read.csv("COMBO17.csv")
```

Punto 1

Para saber cual es la variable que presenta problemas, se usa la función str:

```
str(glx)
```

```
## 'data.frame':    3462 obs. of  65 variables:
##  $ Nr      : int  6 9 16 21 26 29 45 49 50 51 ...
##  $ Rmag    : num  25 25 24.2 25.2 25.5 ...
##  $ e.Rmag  : num  0.097 0.181 0.054 0.128 0.112 0.056 0.257 0.217 0.098 0.097 ...
##  $ ApDRmag : num  0.935 -0.135 0.821 0.639 -1.588 ...
##  $ mumax   : num  24.2 25.3 23.5 24.9 24.9 ...
##  $ Mcz     : num  0.832 0.927 1.202 0.912 0.848 ...
##  $ e.Mcz   : num  0.036 0.122 0.037 0.177 0.067 0.183 0.174 0.147 0.052 0.057 ...
##  $ MCzml   : num  1.4 0.864 1.217 0.776 1.33 ...
##  $ chi2red : num  0.64 0.41 0.92 0.39 1.45 0.52 1.31 1.84 1.03 0.55 ...
##  $ UjMAG   : num  -17.7 -18.3 -19.8 -17.8 -17.7 ...
##  $ e.UjMAG : num  0.14 0.22 0.14 0.17 0.42 0.16 0.3 0.44 0.15 0.16 ...
##  $ BjMAG   : num  -17.5 17.9 -19.9 -17.4 -18.4 ...
##  $ e.BjMAG : num  0.25 0.55 0.14 0.31 0.83 1.37 1.94 1.81 0.15 0.19 ...
##  $ VjMAG   : num  -17.8 -18.2 -20.4 -17.7 -19.4 ...
##  $ e.VjMAG : num  0.25 0.55 0.14 0.31 0.83 1.37 1.94 1.81 0.32 0.14 ...
##  $ usMAG   : num  -17.8 -18.4 -19.9 -18 -17.8 ...
##  $ e.usMAG : num  0.14 0.22 0.14 0.17 0.42 0.16 0.3 0.44 0.15 0.16 ...
##  $ gsMAG   : num  -17.6 -18 -20.1 -17.5 -18.7 ...
##  $ e.gsMAG : num  0.25 0.55 0.14 0.31 0.83 1.37 1.94 1.81 0.32 0.14 ...
##  $ rsMAG   : num  -18 -18.4 -20.7 -17.9 -19.9 ...
##  $ e.rsMAG : num  0.25 0.55 0.14 0.31 0.83 1.37 1.94 1.81 0.32 0.14 ...
##  $ UbMAG   : num  -17.8 -18.4 -19.8 -17.9 -17.8 ...
##  $ e.UbMAG : num  0.14 0.22 0.14 0.17 0.42 0.16 0.3 0.44 0.15 0.16 ...
##  $ BbMAG   : num  -17.5 -17.9 -19.9 -17.4 -18.4 ...
##  $ e.BbMAG : num  0.25 0.55 0.14 0.31 0.83 1.37 1.94 1.81 0.15 0.19 ...
##  $ VnMAG   : num  -17.8 -18.2 -20.4 -17.7 -19.4 ...
##  $ e.VbMAG : num  0.25 0.55 0.14 0.31 0.83 1.37 1.94 1.81 0.32 0.14 ...
##  $ S280MAG : num  -18.2 -18 -19.8 -18.1 -13.9 ...
##  $ e.S280MA: num  0.17 0.54 0.12 0.28 45.11 ...
##  $ W420FE  : num  0.00066 0.000324 0.013 0.0119 0.00135 0.00324 0.00898 0.00436 0.0144 0.02 ...
```

```
##  $ e.W420FE: Factor w/ 486 levels "1.01E-02","1.06E-02",..: 240 174 297 43 226 157 61 312 239 56 ...
##  $ W462FE  : num  0.0127 0.00514 0.0197 0.0159 0.00509 0.00332 0.00406 0.00116 0.0128 0.0212 ...
##  $ e.W462FE: num  0.00372 0.00323 0.00432 0.00314 0.00268 0.00275 0.00265 0.00365 0.00492 0.00275 ..
##  $ W485FD  : num  0.0189 0.00273 0.0255 0.00156 0.00185 0.00401 0.00486 0.000102 0.00437 0.015 ...
##  $ e.W485FD: num  0.00448 0.00485 0.00428 0.00493 0.00401 0.00497 0.00363 0.00389 0.00483 0.00375 ..
##  $ W518FE  : num  0.0182 0.000785 0.0159 0.00261 0.00996 0.00166 0.00178 0.00622 0.0165 0.0098 ...
##  $ e.W518FE: num  0.00355 0.00485 0.00464 0.00476 0.00432 0.00342 0.00357 0.00553 0.00461 0.00351 ..
##  $ W571FS  : num  0.0147 0.00991 0.0229 0.00176 0.00344 0.00446 0.00537 0.00216 0.00745 0.00941 ...
##  $ e.W571FS: num  0.00301 0.00284 0.00455 0.0031 0.00448 0.00311 0.00301 0.00357 0.00459 0.00297 ...
##  $ W604FE  : num  0.0166 0.00905 0.0234 0.00916 0.00632 0.00451 0.00262 0.00807 0.0107 0.0135 ...
##  $ e.W604FE: num  0.00409 0.00445 0.00374 0.00332 0.00366 0.00429 0.00368 0.00296 0.00433 0.00382 ..
##  $ W646FD  : num  0.0188 0.00298 0.0231 0.00633 -0.000184 -0.000551 0.0132 0.00628 -0.004 0.0139 ...
##  $ e.W646FD: num  0.00563 0.00892 0.00667 0.00596 0.0124 0.00966 0.00644 0.0147 0.00795 0.0112 ...
##  $ W696FE  : num  0.0246 0.00983 0.0272 0.0123 0.00554 0.00283 0.00776 0.014 0.0175 0.0168 ...
##  $ e.W696FE: num  0.00351 0.00343 0.00405 0.00248 0.00293 0.00272 0.00308 0.0116 0.00284 0.00266 ...
##  $ W753FE  : num  0.0245 0.0142 0.0354 0.00225 0.0162 0.0174 0.0119 0.0154 0.0193 0.00767 ...
##  $ e.W753FE: num  0.00524 0.00527 0.00456 0.00692 0.00497 0.0044 0.00443 0.00608 0.00468 0.00577 ...
##  $ W815FS  : num  0.0216 0.0147 0.0453 0.0169 0.00676 0.00829 0.00561 0.00687 0.0207 0.0128 ...
##  $ e.W815FS: num  0.00266 0.00308 0.0036 0.00276 0.00314 0.00371 0.00275 0.00357 0.00285 0.00255 ...
##  $ W856FD  : num  0.0244 0.0114 0.0781 0.00875 0.0102 0.0039 0.00684 0.0115 0.0205 0.00587 ...
##  $ e.W856FD: num  0.00546 0.00627 0.00658 0.00672 0.0061 0.00696 0.00557 0.0102 0.00524 0.00617 ...
##  $ W914FD  : num  0.0377 0.0103 0.0711 0.007 0.0133 0.00485 0.0144 0.0169 0.0276 0.013 ...
##  $ e.W914FD: num  0.0061 0.00646 0.00613 0.00557 0.00682 0.00563 0.00615 0.00761 0.00663 0.00664 ...
##  $ W914FE  : num  0.0117 0.0263 0.0641 0.00587 0.0199 0.0264 0.0185 0.0106 0.0449 0.00219 ...
##  $ e.W914FE: num  0.0101 0.0148 0.0127 0.0114 0.0103 0.0097 0.00876 0.00909 0.0139 0.0115 ...
##  $ UFS     : num  0.0187 0.00706 0.0126 0.0141 0.00514 0.00292 0.0123 0.00691 0.00677 0.0149 ...
##  $ e.UFS   : num  0.00239 0.00238 0.00184 0.00186 0.0017 0.00198 0.0021 0.00181 0.00187 0.00224 ...
##  $ BFS     : num  0.0163 0.0042 0.0183 0.0118 0.00102 0.00329 0.00622 0.00266 0.0076 0.017 ...
##  $ e.BFS   : num  0.00129 0.00115 0.00115 0.0011 0.00127 0.00104 0.00124 0.00137 0.00125 0.00109 ...
##  $ VFD     : num  1.73e-02 3.93e-03 1.88e-02 9.67e-03 3.85e-05 3.55e-03 5.04e-03 1.20e-04 8.59e-03 1
##  $ e.VFD   : num  0.00141 0.00182 0.00167 0.00204 0.0016 0.0013 0.00129 0.00158 0.00172 0.0017 ...
##  $ RFS     : num  0.0165 0.00723 0.0288 0.0105 0.00139 0.00474 0.00398 0.00162 0.0116 0.0122 ...
##  $ e.RFS   : num  0.000434 0.0005 0.000655 0.000416 0.000499 0.000489 0.000429 0.000552 0.000495 0.0
##  $ IFD     : num  0.0247 0.00973 0.057 0.0134 0.0059 0.00356 0.00271 0.00232 0.0164 0.0113 ...
##  $ e.IFD   : num  0.00483 0.0046 0.00465 0.0033 0.00444 0.00446 0.0048 0.00385 0.00444 0.00316 ...
```

La variable que está causando el problema es la variable ´Combo_data$e.W420FE´, la cual aparece como variable categórica, sin embargo, esta parece ser una variable numérica. Esto se debe a que dentro de los datos de la variable se encuentra uno de tipo caracter, lo que causa que toda la variable sea tomada como factor En este caso particular, el problema son los espacios en algunos de los registros.

```
head(levels(glx$e.W420FE), 20)
```

```
##  [1] "1.01E-02"     "1.06E-02"     "1.07E-02"     "1.09E-02"
##  [5] "1.15E-02"     "1.22E-02"     "1.23E-02"     "1.31E-02"
##  [9] "1.47E-02"     "1.56E-02"     "1.81E-02"     "2  1.296E-02"
## [13] "2  1.586E-02" "2  1.682E-02" "2  1.743E-02" "2  1.948E-02"
## [17] "2  2.082E-02" "2  2.149E-02" "2  2.309E-02" "2  2.358E-02"
```

Punto 2

En adelante se trabaja con el siguiente conjunto de datos restringido

```
variables_de_interes = c("Nr", "Rmag", "e.Rmag", "ApDRmag", "Mcz",
                         "UjMAG", "BjMAG", "VjMAG", "usMAG", "gsMAG",
                         "rsMAG", "UbMAG", "BbMAG", "VnMAG", "S280MAG")
glx_vars_interes = glx[, variables_de_interes]
kable(head(glx_vars_interes[, 1:10]))
```

Nr	Rmag	e.Rmag	ApDRmag	Mcz	UjMAG	BjMAG	VjMAG	usMAG	gsMAG
6	24.995	0.097	0.935	0.832	-17.67	-17.54	-17.76	-17.83	-17.60
9	25.013	0.181	-0.135	0.927	-18.28	17.86	-18.20	-18.42	-17.96
16	24.246	0.054	0.821	1.202	-19.75	-19.91	-20.41	-19.87	-20.05
21	25.203	0.128	0.639	0.912	-17.83	-17.39	-17.67	-17.98	-17.47
26	25.504	0.112	-1.588	0.848	-17.69	-18.40	-19.37	-17.81	-18.69
29	23.740	0.056	-1.636	0.882	-19.22	-18.11	-18.70	-19.34	-18.27

Los datos atípicos se calculan de forma univariada usando la función boxplot.stats, la cual permite seleccionar el umbran en el cual se va a decidir si un dato es atípico o no. Según lo aprendido en AID, los datos atípicos "fuertes" son aquellos que están a más de 3 rangos intercuartílicos del primer y tercer cuartíl, los cuales se calculan a continuación

```
var_limites = NULL
outliers = NULL
for(i in 2:ncol(glx_vars_interes)){
  Var_stats = boxplot.stats(glx_vars_interes[,i], coef = 3)$stats
  var_limites = cbind(var_limites, Var_stats[c(1,5)])
  outliers_var = which(glx_vars_interes[,i] < Var_stats[1] |
                         glx_vars_interes[,i] > Var_stats[5])
  outliers = union(outliers, outliers_var)
  # print(colnames(glx_vars_interes)[i])
}

colnames(var_limites) =  colnames(glx_vars_interes)[2:ncol(glx_vars_interes)]
rownames(var_limites) = c("Lim_Inf", "Lim_Sup")
```

La cantidad de datos atípicos y el porcentaje que representa se muestra a continuación:

```
atipicos = data.frame(length(outliers), round(length(outliers)/nrow(glx)*100, 2))
colnames(atipicos) = c("Cantidad", "Porcentaje")
kable(atipicos)
```

Cantidad	Porcentaje
77	2.22

Para eliminar los datos atípicos se usa la siguiente sentencia:

```
glx_vars_interes_no_out = glx_vars_interes[-outliers, ]
```

Los límites para determinar si un dato es atípico según los boxplots son los siguientes:

```
kable(var_limites[,1:7])
```

	Rmag	e.Rmag	ApDRmag	Mcz	UjMAG	BjMAG	VjMAG
Lim_Inf	17.578	0.001	-1.927	0.007	-23.21	-23.15	-23.62
Lim_Sup	27.000	0.311	1.462	1.379	-12.00	-11.22	-11.43

```
kable(var_limites[,8:14])
```

	usMAG	gsMAG	rsMAG	UbMAG	BbMAG	VnMAG	S280MAG
Lim_Inf	-23.33	-23.28	-23.94	-23.28	-23.13	-23.62	-23.59
Lim_Sup	-12.16	-11.31	-11.53	-12.09	-11.28	-11.43	-11.74

Punto 3

Para eliminar los datos faltantes se usa la función na.omit, además de esto se usó la función mostrada a continuación para determinar la cantidad de registros que presentaban datos faltantes

```
reg_faltantes = unique(unlist(apply(glx_vars_interes_no_out, 2, function(x) which(is.na(x))), use.names
numero_faltantes = length(reg_faltantes)
names(numero_faltantes) = "Número de faltantes"
kable(numero_faltantes)
```

Número de faltantes	23

Eliminando los faltantes:

```
glx_vars_interes_no_out_no_missing = na.omit(glx_vars_interes_no_out)
```

Punto 4

Las correlaciones de las variables no normalizadas son

```
vars_normalizar = c("UjMAG", "BjMAG", "VjMAG", "usMAG", "gsMAG", "rsMAG",
                    "UbMAG", "BbMAG", "VnMAG")

cor_matrix = (cor(glx_vars_interes_no_out_no_missing[, vars_normalizar]))
kable(cor_matrix[,1:4])
```

	UjMAG	BjMAG	VjMAG	usMAG
UjMAG	1.0000000	0.9710243	0.9544010	0.9998331
BjMAG	0.9710243	1.0000000	0.9913592	0.9697681
VjMAG	0.9544010	0.9913592	1.0000000	0.9523379
usMAG	0.9998331	0.9697681	0.9523379	1.0000000

	UjMAG	BjMAG	VjMAG	usMAG
gsMAG	0.9653332	0.9954905	0.9975596	0.9638778
rsMAG	0.9470592	0.9874218	0.9992085	0.9447930
UbMAG	0.9999832	0.9704484	0.9533846	0.9998795
BbMAG	0.9714537	0.9999890	0.9909583	0.9702330
VnMAG	0.9542722	0.9913028	0.9999966	0.9522028

```
kable(cor_matrix[,5:9])
```

	gsMAG	rsMAG	UbMAG	BbMAG	VnMAG
UjMAG	0.9653332	0.9470592	0.9999832	0.9714537	0.9542722
BjMAG	0.9954905	0.9874218	0.9704484	0.9999890	0.9913028
VjMAG	0.9975596	0.9992085	0.9533846	0.9909583	0.9999966
usMAG	0.9638778	0.9447930	0.9998795	0.9702330	0.9522028
gsMAG	1.0000000	0.9945943	0.9646317	0.9953610	0.9975206
rsMAG	0.9945943	1.0000000	0.9458997	0.9868826	0.9992291
UbMAG	0.9646317	0.9458997	1.0000000	0.9708965	0.9532532
BbMAG	0.9953610	0.9868826	0.9708965	1.0000000	0.9908994
VnMAG	0.9975206	0.9992291	0.9532532	0.9908994	1.0000000

Para normalizar las variables se usa el siguiente código:

```
glx_normalizado = NULL
for(i in vars_normalizar){
  var_normalizada = glx_vars_interes_no_out_no_missing$S280MAG - glx_vars_interes_no_out_no_missing[,i]
  glx_normalizado = cbind(glx_normalizado, var_normalizada)
}
colnames(glx_normalizado) = paste(vars_normalizar, "_normalizada", sep = "")
```

La matriz de correlación para las variables normalizadas se presenta a continuación

```
cor_matrix_normal = cor(glx_normalizado)
kable(cor_matrix_normal[,1:3])
```

	UjMAG_normalizada	BjMAG_normalizada	VjMAG_normalizada
UjMAG_normalizada	1.0000000	0.8360888	0.8397515
BjMAG_normalizada	0.8360888	1.0000000	0.9724434
VjMAG_normalizada	0.8397515	0.9724434	1.0000000
usMAG_normalizada	0.9981834	0.8295592	0.8315121
gsMAG_normalizada	0.8323133	0.9772886	0.9947701
rsMAG_normalizada	0.8391912	0.9668592	0.9982269
UbMAG_normalizada	0.9998571	0.8332277	0.8357050
BbMAG_normalizada	0.8350811	0.9999629	0.9714787
VnMAG_normalizada	0.8397170	0.9723778	0.9999862

```
kable(cor_matrix_normal[,4:6])
```

	usMAG_normalizada	gsMAG_normalizada	rsMAG_normalizada
UjMAG_normalizada	0.9981834	0.8323133	0.8391912
BjMAG_normalizada	0.8295592	0.9772886	0.9668592
VjMAG_normalizada	0.8315121	0.9947701	0.9982269
usMAG_normalizada	1.0000000	0.8256540	0.8309646
gsMAG_normalizada	0.8256540	1.0000000	0.9895272
rsMAG_normalizada	0.8309646	0.9895272	1.0000000
UbMAG_normalizada	0.9985462	0.8291575	0.8348480
BbMAG_normalizada	0.8286433	0.9769064	0.9656219
VnMAG_normalizada	0.8314577	0.9947095	0.9982601

```
kable(cor_matrix_normal[,7:9])
```

	UbMAG_normalizada	BbMAG_normalizada	VnMAG_normalizada
UjMAG_normalizada	0.9998571	0.8350811	0.8397170
BjMAG_normalizada	0.8332277	0.9999629	0.9723778
VjMAG_normalizada	0.8357050	0.9714787	0.9999862
usMAG_normalizada	0.9985462	0.8286433	0.8314577
gsMAG_normalizada	0.8291575	0.9769064	0.9947095
rsMAG_normalizada	0.8348480	0.9656219	0.9982601
UbMAG_normalizada	1.0000000	0.8322780	0.8356648
BbMAG_normalizada	0.8322780	1.0000000	0.9714085
VnMAG_normalizada	0.8356648	0.9714085	1.0000000

El efecto que tiene la normalización con la variable “S280MAG” es el de reducir la correlación entre las variables.