Trabajo RLM Parte I

Alejandro Salazar Mejía

10/5/2021

Análisis Descriptivo

```
library(plotrix)
```

```
## Warning: package 'plotrix' was built under R version 4.0.3
```

```
library(knitr)
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 4.0.5

## Loading required package: ggplot2

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(ggplot2)
```

```
datos <- read.table("APC1modifm3.csv", header = T, sep = ";", dec = ",",
                    colClasses = c(rep("numeric",7),"factor",rep("numeric",3),"factor"))
```

```
str(datos)
```

```
## 'data.frame':    90 obs. of  12 variables:
##  $ ID     : num  5 10 13 18 27 28 29 31 33 34 ...
##  $ DPERM  : num  11.2 8.84 12.78 11.62 9.31 ...
##  $ EDAD   : num  56.5 56.3 56.8 53.9 47.2 52.1 54.5 49.9 54.1 54 ...
##  $ RINF   : num  5.7 6.3 7.7 6.4 4.5 3.2 4.4 5 5.3 6.1 ...
##  $ RRC    : num  34.5 29.6 46 25.5 30.2 10.8 18.6 19.7 17.3 24.2 ...
##  $ RRX    : num  88.9 82.6 116.9 99.2 101.3 ...
##  $ NCAMAS : num  180 85 322 133 170 176 248 318 196 312 ...
##  $ AEM    : Factor w/ 2 levels "1","2": 2 2 1 2 2 2 2 2 2 2 ...
##  $ PDP    : num  134 59 252 113 124 156 217 270 164 258 ...
##  $ NENFERM: num  151 66 349 101 173 88 189 335 165 169 ...
##  $ FSD    : num  40 40 57.1 37.1 37.1 37.1 37.1 57.1 34.3 54.3 ...
##  $ REGION : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 1 ...
```

```
summary(datos)
```

```
##        ID              DPERM             EDAD            RINF
##  Min.   :  1.00   Min.   : 7.080   Min.   :38.80   Min.   :1.300
##  1st Qu.: 33.25   1st Qu.: 8.390   1st Qu.:51.00   1st Qu.:3.700
##  Median : 62.50   Median : 9.385   Median :53.20   Median :4.400
##  Mean   : 60.17   Mean   : 9.719   Mean   :53.25   Mean   :4.399
##  3rd Qu.: 88.75   3rd Qu.:10.658   3rd Qu.:56.08   3rd Qu.:5.300
##  Max.   :113.00   Max.   :19.560   Max.   :65.90   Max.   :7.800
##       RRC              RRX             NCAMAS      AEM         PDP
##  Min.   : 1.60   Min.   : 39.60   Min.   : 29.0   1:13   Min.   : 20.00
##  1st Qu.: 8.40   1st Qu.: 69.20   1st Qu.:102.0   2:77   1st Qu.: 66.25
##  Median :14.05   Median : 85.40   Median :184.0          Median :136.50
##  Mean   :16.13   Mean   : 82.24   Mean   :246.9          Mean   :186.56
##  3rd Qu.:20.75   3rd Qu.: 96.05   3rd Qu.:305.8          3rd Qu.:247.00
##  Max.   :60.50   Max.   :133.50   Max.   :835.0          Max.   :791.00
##     NENFERM            FSD        REGION
##  Min.   : 19.00   Min.   :14.30   1:23
##  1st Qu.: 66.25   1st Qu.:31.40   2:25
##  Median :124.50   Median :41.45   3:29
##  Mean   :165.97   Mean   :42.16   4:13
##  3rd Qu.:208.75   3rd Qu.:51.40
##  Max.   :629.00   Max.   :74.30
```

```
attach(datos)
```

Análisis de variables numéricas

A continuación se muestran una tabla de estadísticos descriptivos por cada variable numérica de la base de datos, donde:

- Min : Mínimo valor registrado de dicha variable.
- $1st\ Qu$: Primer cuartil de los datos (cuantil 0.25), i.e., el 25% de los datos es menor o igual a dicho valor.
- $Median$: Mediana de los datos, i.e., valor que divide los datos en dos partes iguales.
- $Mean$: Media aritmética de los datos.
- $3st\ Qu$: Tercer cuartil de los datos (cuantil 0.75), i.e., el 25% de los datos es mayor o igual a dicho valor.
- Max : Máximo valor registrado de dicha variable
- sd : Desviación estándar muestral de los datos

```
cond <- names(datos) != c("ID", "AEM", "REGION")
numvar <- names(datos)[cond]
tableList <- list()

for (variable in numvar) {
    currVar <- datos[variable][[1]]
```

```
    Table <- data.frame(round(matrix(c(summary(currVar),sd(currVar)),ncol=7),2))
    names(Table) <- c(names(summary(currVar)),"sd")
    tableList[[variable]] <- list(Table,
                                  paste("Estadísticos de resumen para variable",
                                        variable))
    # print(Table)
}

kable( tableList[[1]][[1]] , caption = tableList[[1]][[2]])
```

Table 1: Estadísticos de resumen para variable DPERM

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd
7.08	8.39	9.38	9.72	10.66	19.56	2.03

```
kable( tableList[[2]][[1]] , caption = tableList[[2]][[2]])
```

Table 2: Estadísticos de resumen para variable EDAD

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd
38.8	51	53.2	53.25	56.08	65.9	4.59

```
kable( tableList[[3]][[1]] , caption = tableList[[3]][[2]])
```

Table 3: Estadísticos de resumen para variable RINF

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd
1.3	3.7	4.4	4.4	5.3	7.8	1.37

```
kable( tableList[[4]][[1]] , caption = tableList[[4]][[2]])
```

Table 4: Estadísticos de resumen para variable RRC

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd
1.6	8.4	14.05	16.13	20.75	60.5	10.73

```
kable( tableList[[5]][[1]] , caption = tableList[[5]][[2]])
```

Table 5: Estadísticos de resumen para variable RRX

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd
39.6	69.2	85.4	82.24	96.05	133.5	20.1

```
kable( tableList[[6]][[1]] , caption = tableList[[6]][[2]])
```

Table 6: Estadísticos de resumen para variable NCAMAS

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd
29	102	184	246.89	305.75	835	185.8

```
kable( tableList[[7]][[1]] , caption = tableList[[7]][[2]])
```

Table 7: Estadísticos de resumen para variable PDP

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd
20	66.25	136.5	186.56	247	791	152.75

```
kable( tableList[[8]][[1]] , caption = tableList[[8]][[2]])
```

Table 8: Estadísticos de resumen para variable NENFERM

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd
19	66.25	124.5	165.97	208.75	629	129.55

```
kable( tableList[[9]][[1]] , caption = tableList[[9]][[2]])
```

Table 9: Estadísticos de resumen para variable FSD

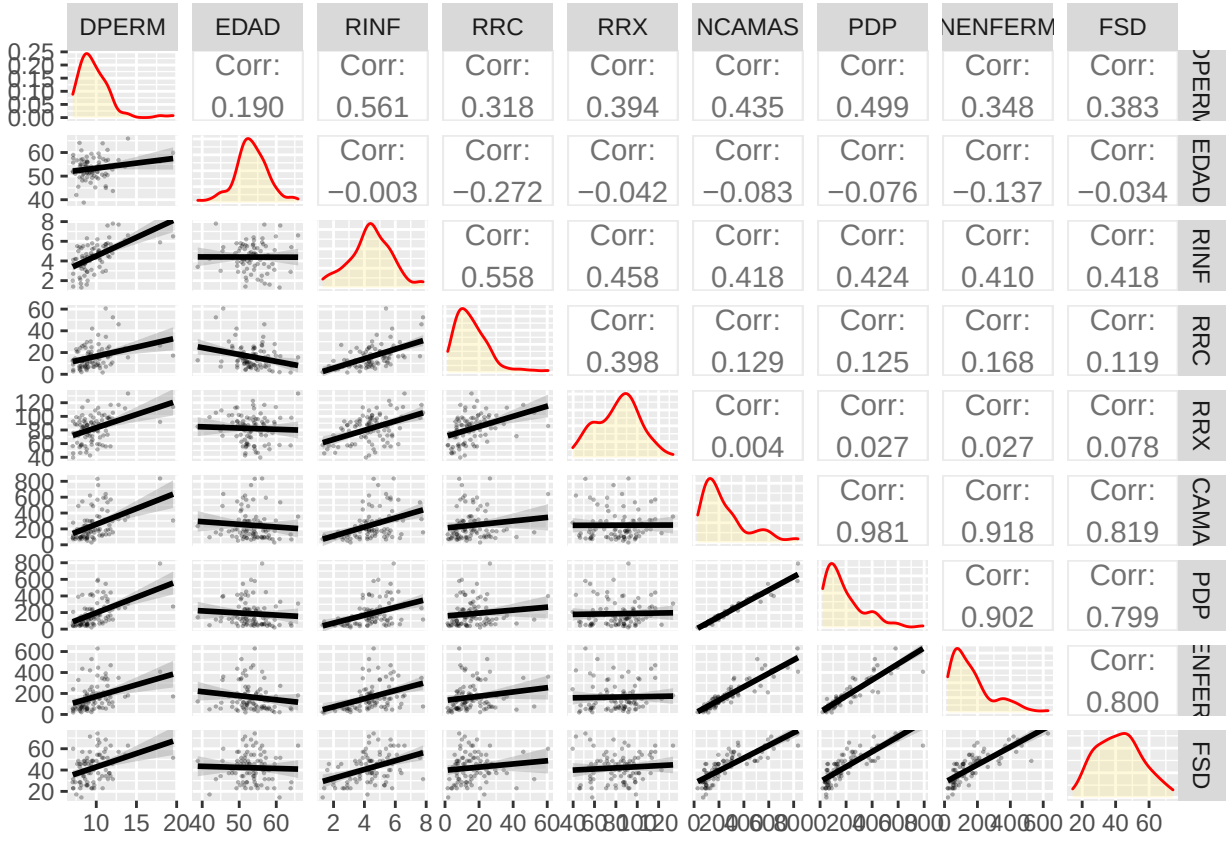
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd
14.3	31.4	41.45	42.16	51.4	74.3	13.47

La figura *numéricas1* es una matriz de gráficos que pretende presentar la relación entre parejas de variables. En el margen superior y lateral derecho se indican el nombre de las variables que se ponen a interactuar. En la diagonal principal de esta matriz, se encuentra la gráfica de densidad (no paramétrica) de cada variable numérica. Por encima de esta diagonal, se presenta la correlación de cada par diferente de variables. Por debajo de la diagonal se encuentran los gráficos de disperción de cada par diferente de variables, junto con una línea recta ajustada y su respectivo intervalo de confianza alrededor de esta.

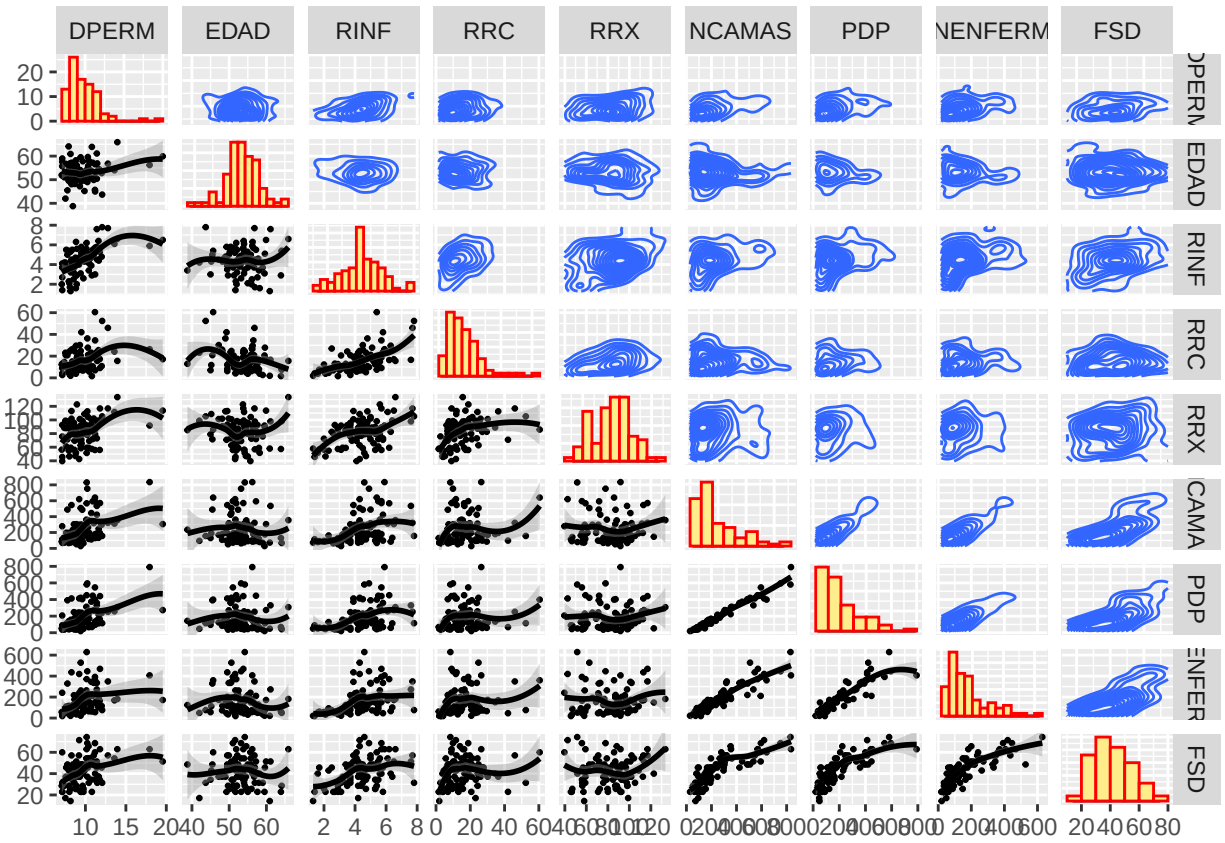
Similarmente, la figura *numérica2* también es una matriz de gráficos. En este caso, la diagonal de la matriz presenta el histograma correspondiente a cada variable numérica. Por encima de esta diagonal, se muestran gráficadas las curvas de nivel de la densidad bivariada (no paramétrica) de las dos variables en cuestión. Por debajo de la diagonal se encuentran los gráficos de disperción de cada par diferente de variables, junto con una curva LOESS ajustada (no paramétrica) y su respectivo intervalo de confianza alrededor de esta.

```
datosNumericos <- datos[numvar]

# numérica1:
win.graph()
```

```
ggpairs(datosNumericos, diag=list(continuous = wrap("densityDiag",color="red",
                                    fill="lightgoldenrod1",alpha=0.3)),
        lower = list(continuous = wrap("smooth", alpha = 0.3, size=0.1,method = "lm")),
        upper = list(continuous = wrap("cor", stars = F)))
```




```
# numérica2:
gg2 <- ggpairs(datosNumericos,
               lower = list(continuous = wrap("smooth_loess", cex = 0.5) ),
               upper = list(continuous = "density"))
for(i in 1:ncol(datosNumericos)){
    gg2[i,i] <- gg2[i,i] +
        geom_histogram(breaks=hist(datosNumericos[,i],breaks ="FD",plot=F)$breaks,
                       colour ="red",fill="lightgoldenrod1")

}
win.graph()
gg2
```



De los gráficos podemos resaltar algunas observaciones interesantes:

i) La varible 'edad' parece tener el efecto más leve sobre la variable respuesta 'Longitud de permancencia' (DPERM). En cambio, podría creerse que el resto de variables son significativas para explicar la variabilidad de DPERM, siguiendo una tendencia positiva.

ii) Del mismo modo, 'edad' no aparenta estar relacionada con el resto de variables explicatorias.

iii) Se observa que las variables 'Número de camas', 'Censo promedio diario', 'Número de enfermeras' y 'Facilidades y servicios disponibles' están altamente relacionadas positivamente entre sí.

iv) La variable 'riesgo de infección' aparenta estar relacionada positivamente con el resto de covariables, excepto 'edad'.

v) Las variables 'Razón de rutina de rayos X del pecho' y 'Razón de rutina de cultivos' no muestran estra relacionadas estadísticamente con las variables 'Número de camas', 'Censo promedio diario', 'Número de enfermeras' y 'Facilidades y servicios disponibles'.

Análisis de variables categóricas

A continuación se muestran un par de tablas de frecuencia correspondientes a las variables categóricas de la base de datos.

```
Table1 <-data.frame(t(summary(AEM)))
colnames(Table1) <- c("Afiliados", "No Afiliados")
kable(Table1, caption = "Hospitales afiliados a la escuela de medicina")
```

Table 10: Hospitales afiliados a la escuela de medicina

Afiliados	No Afiliados
13	77

```
Table2 <-data.frame(t(summary(REGION)))
colnames(Table2) <- c("NE", "NC", "S", "W")
kable(Table2, caption = "Hospitales en regiones geográficas")
```

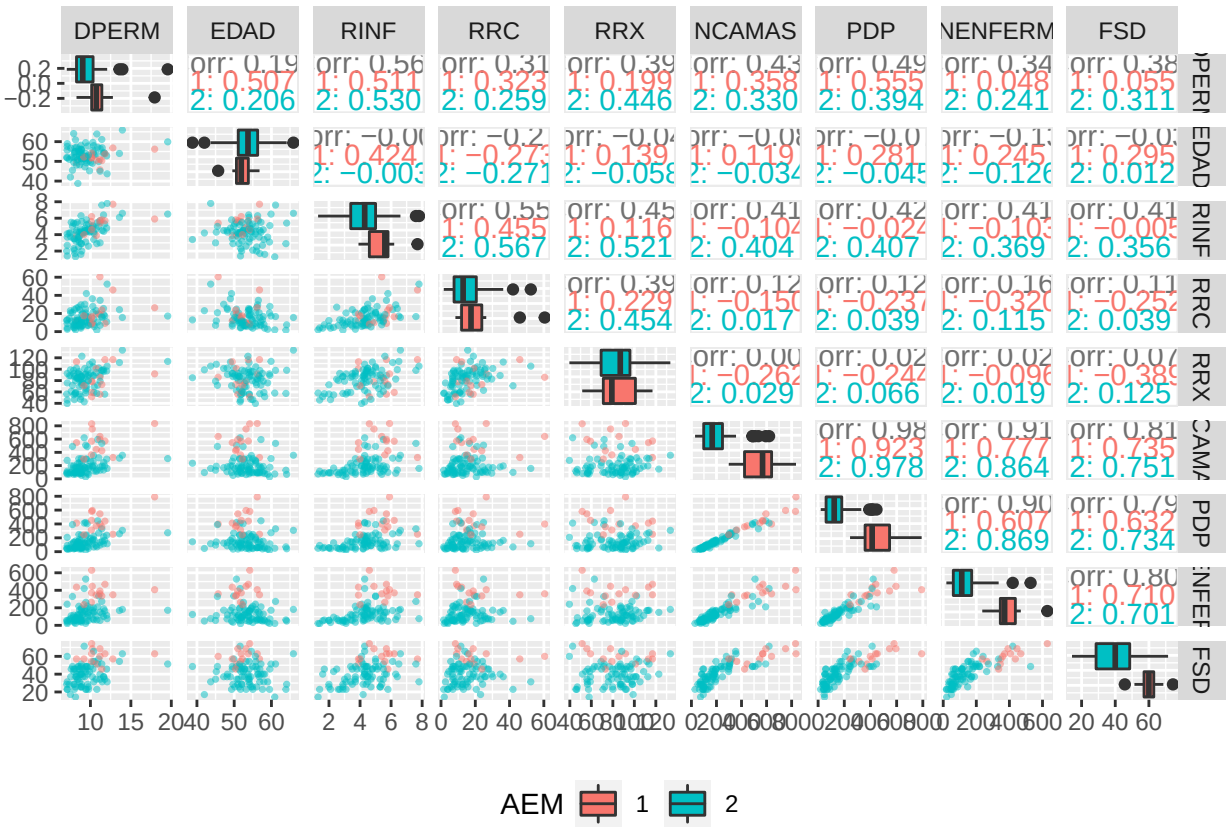
Table 11: Hospitales en regiones geográficas

NE	NC	S	W
23	25	29	13

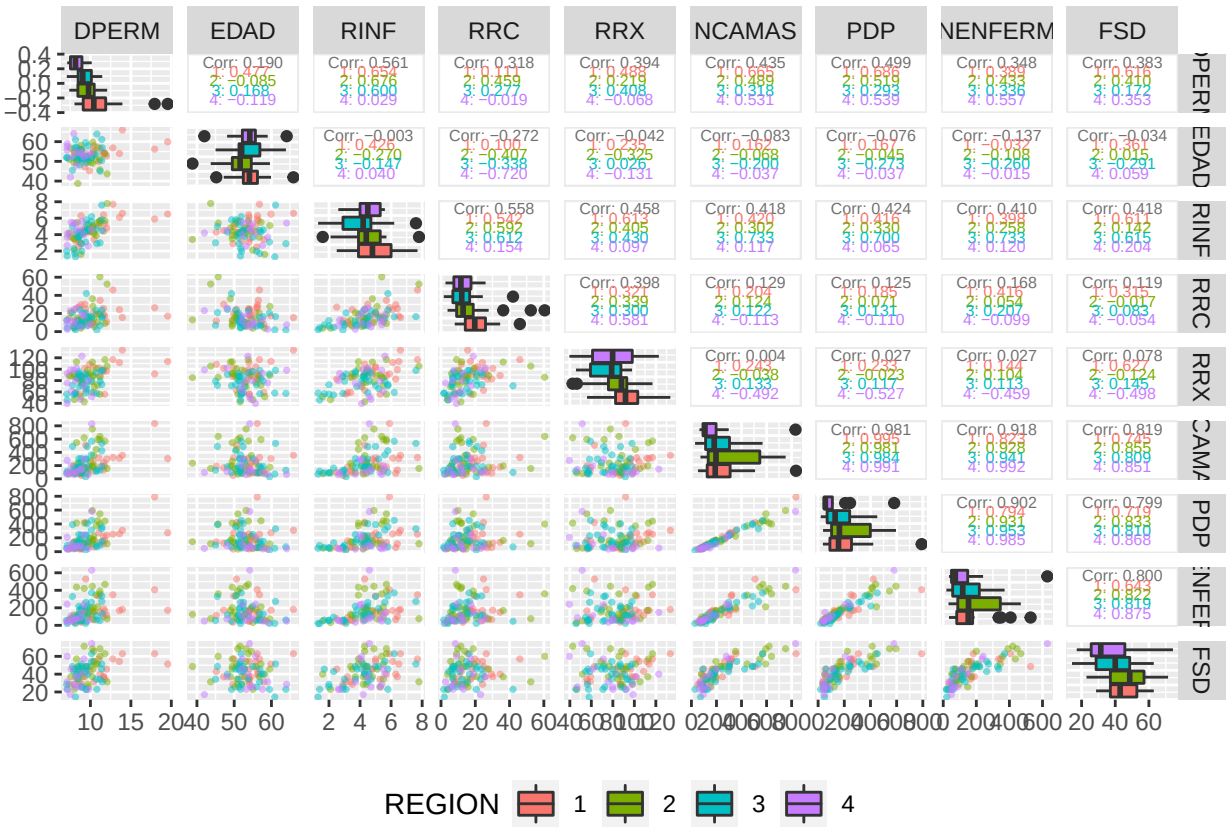
Análisis de variables numéricas agrupadas por variables categóricas

Las figuras *combByAEM* y *combByRegion* son matrices de gráficos que pretenden presentar la relación entre parejas de variables, agrupadas por las variables categóricas 'Afiliación a escuela de medicina' y 'Región', respectivamente. En el margen superior y lateral derecho se indican el nombre de las variables que se ponen a interactuar, y en la parte inferior se muestra la leyenda que explica la agrupación de las variables por colores. En la diagonal principal se presenta el diagrama de Cajas y Bigotes de cada variable, discriminado por la variable categórica correspondiente. Por encima de esta diagonal, se presenta la correlación de cada par diferente de variables, de acuerdo a la categoría correspondiente. Por debajo de la diagonal se encuentran los gráficos de disperción de cada par diferente de variables, donde los colores de cada punto dependen del grupo al que pertenecen.

```
# combByAEM
win.graph()
ggpairs(datos,columns = c(2:7,9:11),legend = c(1,1),mapping =aes(colour=AEM),
        diag = list(continuous = wrap("box_no_facet")),
        upper = list(continuous = wrap("cor", stars = F)),
        lower = list(continuous = wrap("points", cex = 0.6, alpha = 0.5))) +
    theme(legend.position="bottom")
```




```
# combByRegion
win.graph()
ggpairs(datos,columns = c(2:7,9:11),legend = c(1,1),mapping =aes(colour=REGION),
        diag = list(continuous = wrap("box_no_facet")),
        upper = list(continuous = wrap("cor", stars = F, size = 2)),
        lower = list(continuous = wrap("points", cex = 0.6, alpha = 0.5))) +
    theme(legend.position="bottom")
```




```
detach(datos)
```