Ejemplo de problema tipo ANCOVA

Tema 1: Regresión Lineal Múltiple

Septiembre 2022

En el siguiente ejemplo se buscará explicar el deterioro en la capacidad pulmonar de individuos que han sido expuestos a distintos niveles de cadmio vs individuos que no fueron expuestos.

```
> library(multcomp)
> library(GGally)
> library(ggplot2)
```

```
> rm(list = ls(all.names = TRUE))
> gc()
> #setwd("~/GitHub/Notas 2023-1/ApreEstAut")
> options(digits=4)
```

```
> CADdata <- read.table("C:/Users/ncr/Downloads/cadmium.txt",header=TRUE, sep=" ",
+                       dec=".")
> str(CADdata)
```

```
'data.frame':        84 obs. of  3 variables:
 $ group : int  1 1 1 1 1 1 1 1 1 1 ...
 $ age   : int  39 40 41 41 45 49 52 47 61 65 ...
 $ vitcap: num  4.62 5.29 5.52 3.71 4.02 5.09 2.7 4.31 2.7 3.03 ...
```

```
> print(head(CADdata,10))
```

```
   group age vitcap
1      1  39   4.62
2      1  40   5.29
3      1  41   5.52
4      1  41   3.71
5      1  45   4.02
6      1  49   5.09
7      1  52   2.70
8      1  47   4.31
9      1  61   2.70
10     1  65   3.03
```

Las variables del dataset son:

- group: Indica a que grupo pertenece la observación: alta exposición al cadmio, baja exposición, no tuvo exposición al cadmio.
- age: Edad de las persona
- vitcap: Indicador de capacidad pulmonar, entre más alto más capacidad

La variable grupo debe ser convertida a factor
Esta variable indica si el paciente fue expuesto al cadmio en niveles "Alto", "Bajo" o "No fue expuesto"

```
> hist(CADdata$group)
```

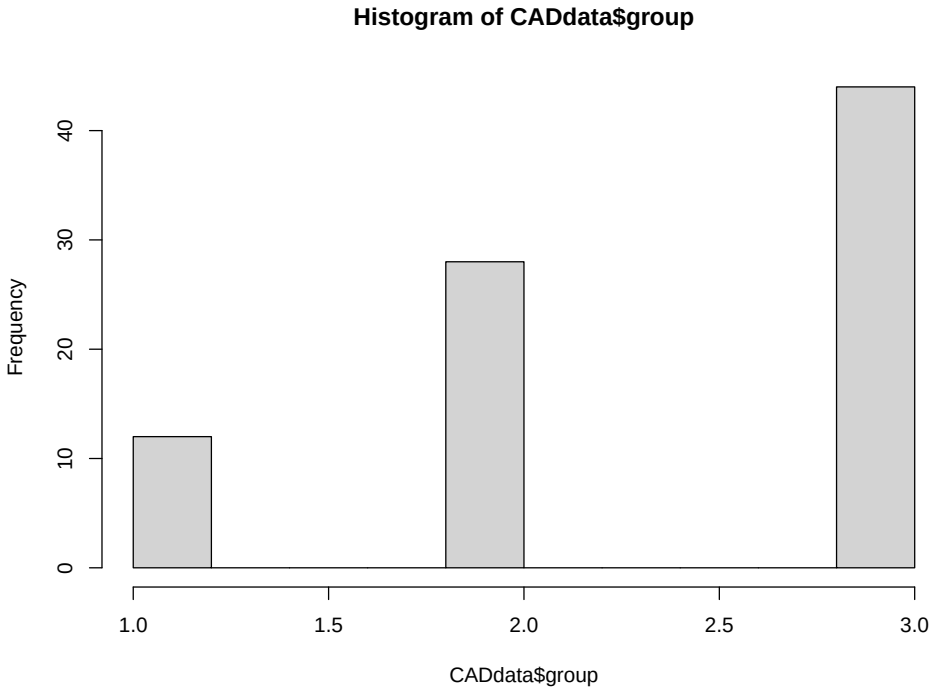


Figure 1: Some caption.

```
> CADdata$group=factor(CADdata$group, levels=c(1,2,3), labels=c("High","Low","No") )
> str(CADdata)
```

```
'data.frame':        84 obs. of  3 variables:
 $ group : Factor w/ 3 levels "High","Low","No": 1 1 1 1 1 1 1 1 1 1 ...
 $ age   : int  39 40 41 41 45 49 52 47 61 65 ...
 $ vitcap: num  4.62 5.29 5.52 3.71 4.02 5.09 2.7 4.31 2.7 3.03 ...
```

Una gráfica global que representan pares de variables

```
> #X11()
> library(GGally)
> ggpairs(data=CADdata, title="Datos", aes(colour = group))
```

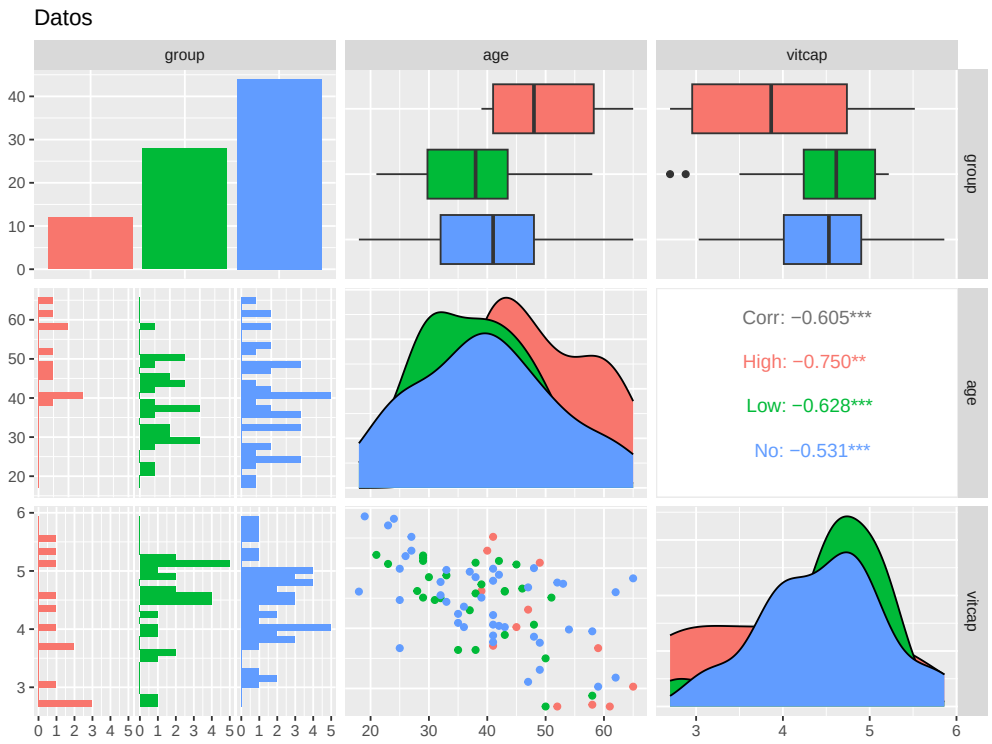


Figure 2: Some caption.

```
> #Otras gráficas
> #X11()
> par(mfrow=c(1,2)); par(mar=c(4,4,2,1.5))
> boxplot(vitcap ~ group, data = CADdata, col = "white", outline=FALSE)
> stripchart(vitcap ~ group, data = CADdata,
+            method = "jitter",
+            pch = 19,
+            col = 2:4,
+            vertical = TRUE,
+            add = TRUE)
> boxplot(age ~ group, data = CADdata, col = "white", outline=FALSE)
> stripchart(age ~ group, data = CADdata,
+            method = "jitter",
+            pch = 19,
+            col = 2:4,
+            vertical = TRUE,
+            add = TRUE)
```

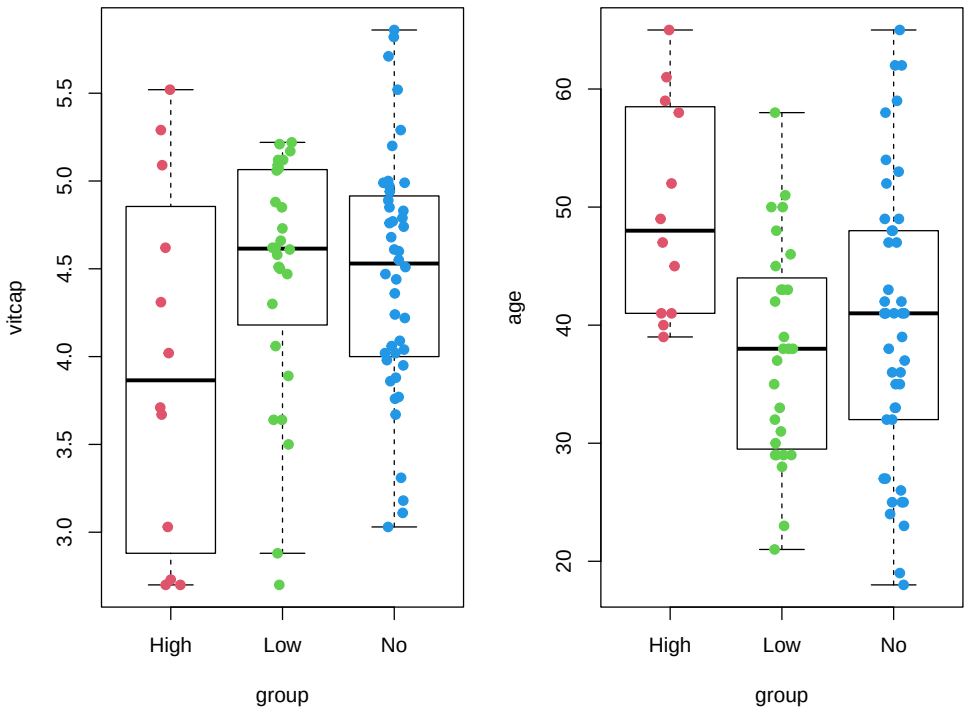


Figure 3: Some caption.

La gráfica de interés en todos los datos

```
> #X11()
> par(mfrow=c(1,2)); par(mar=c(4,4,2,1.5))
> plot(CADdata$age, CADdata$vitcap, col=c("green","red","blue")[CADdata$group]
+      ,xlab = "Age",
+      ylab = "Vital Capacity (L)",xlim=c(20,68), ylim=c(2.5,5.6),
+      pch=c(16,16,16), main = "Exposure to cadmium")
> legend(60,5.8, levels(CADdata$group),
+        col=c("green","red","blue"), pch = c(16,16,16) )
> boxplot(vitcap ~ group, data = CADdata, col = "white", outline=FALSE)
> stripchart(vitcap ~ group, data = CADdata,
+            method = "jitter",
+            pch = 19,
+            col = c("green","red","blue"),
+            vertical = TRUE,
+            add = TRUE)
```

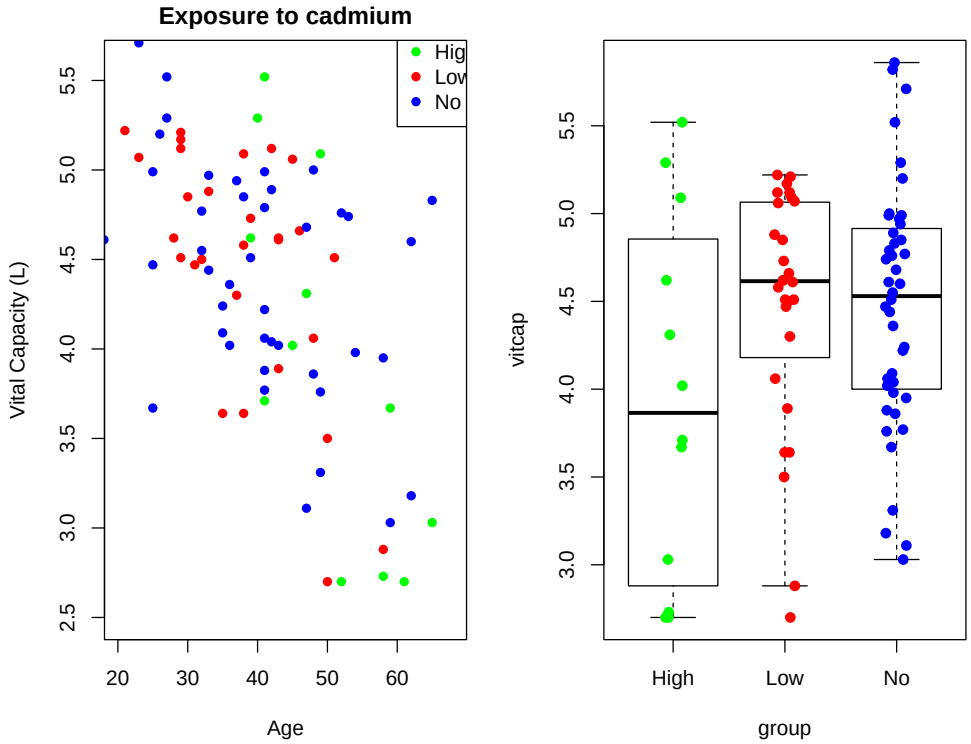


Figure 4: Some caption.

Filtraremos los datos por edad
Aplicamos este filtro ya que al ser un problema tipo ANCOVA se quiere que la muestra sea homogénea en cada uno de los grupos a estudiar, en este caso vemos que del grupo de personas con alta exposición al cadmio no se tienen personas menores a 30 años, entonces filtramos la edad para que todos los grupos sean homogeneos en edad.
En los problema tipo ANCOVA se requiere que las poblaciones sean lo más homogeneas posible y que las unicas diferencias entre cada grupo se deban a cierta exposición a alguno tratamiento.

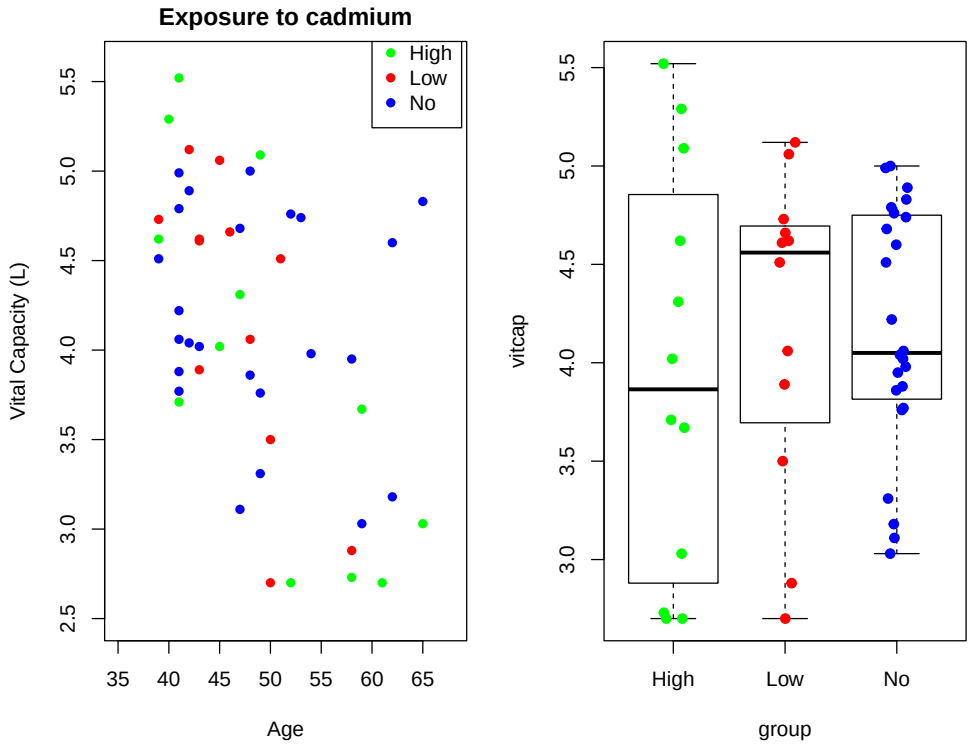
```
> summary(CADdata[CADdata$group=="High",])
```

```
  group         age            vitcap
 High:12   Min.   :39.0   Min.   :2.70
 Low : 0   1st Qu.:41.0   1st Qu.:2.96
 No  : 0   Median :48.0   Median :3.87
           Mean   :49.8   Mean   :3.95
           3rd Qu.:58.2   3rd Qu.:4.74
           Max.   :65.0   Max.   :5.52
```

```
> CADdata=CADdata[CADdata$age>=39,]
```

Nuevamente visualizamos la gráfica

```
> #X11()
> par(mfrow=c(1,2)); par(mar=c(4,4,2,1.5))
> plot(CADdata$age, CADdata$vitcap, col=c("green","red","blue")[CADdata$group],
+      xlab = "Age", ylab = "Vital Capacity (L)",xlim=c(35,68), ylim=c(2.5,5.6),
+      pch=c(16,16,16), main = "Exposure to cadmium")
> legend(60,5.8, levels(CADdata$group),
+        col=c("green","red","blue"), pch = c(16,16,16) )
> boxplot(vitcap ~ group, data = CADdata, col = "white", outline=FALSE)
> stripchart(vitcap ~ group, data = CADdata, method = "jitter",
+            pch = 19, col = c("green","red","blue"), vertical = TRUE,
+            add = TRUE)
```



Importante analizar el metadato de niveles, pues el primero será el nivel de referencia en los modelos
Aquí el nivel referencia es High, es decir, R va a crear variables dummy para las categorias "Low" y "No".
Se define la categoria de referencia como "High"

```
> levels(CADdata$group)
```

```
[1] "High" "Low"  "No"
```