Seminario de Estadística

Tema 1: Modelos Lineales Generalizados

Repaso de Regresión Lineal Múltiple - Problemas tipo ANCOVA

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/Sweave/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)
```

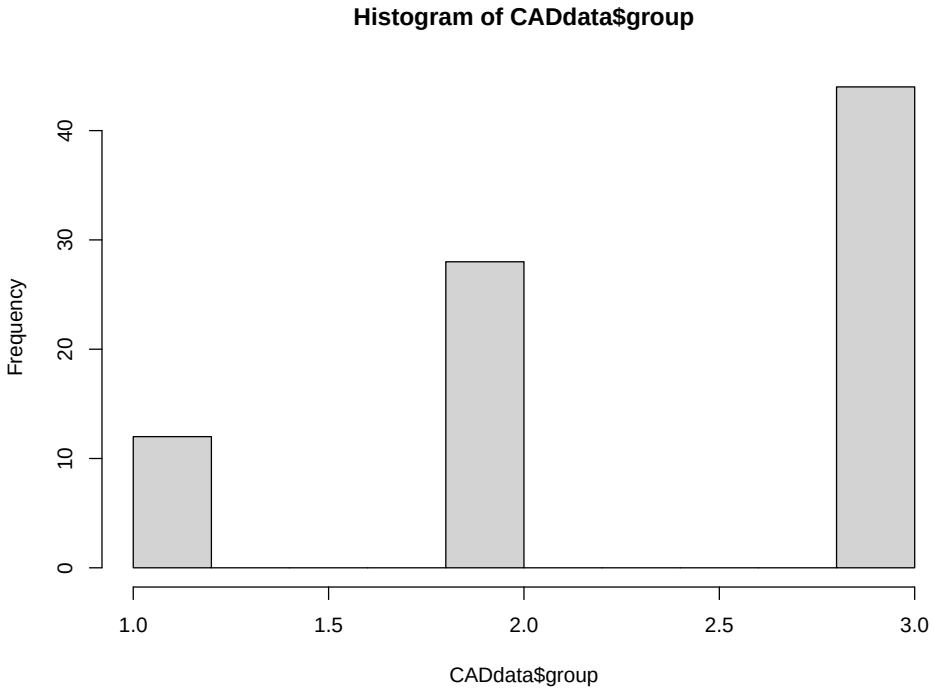


Figure 1: Some caption.

```
> 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))
```

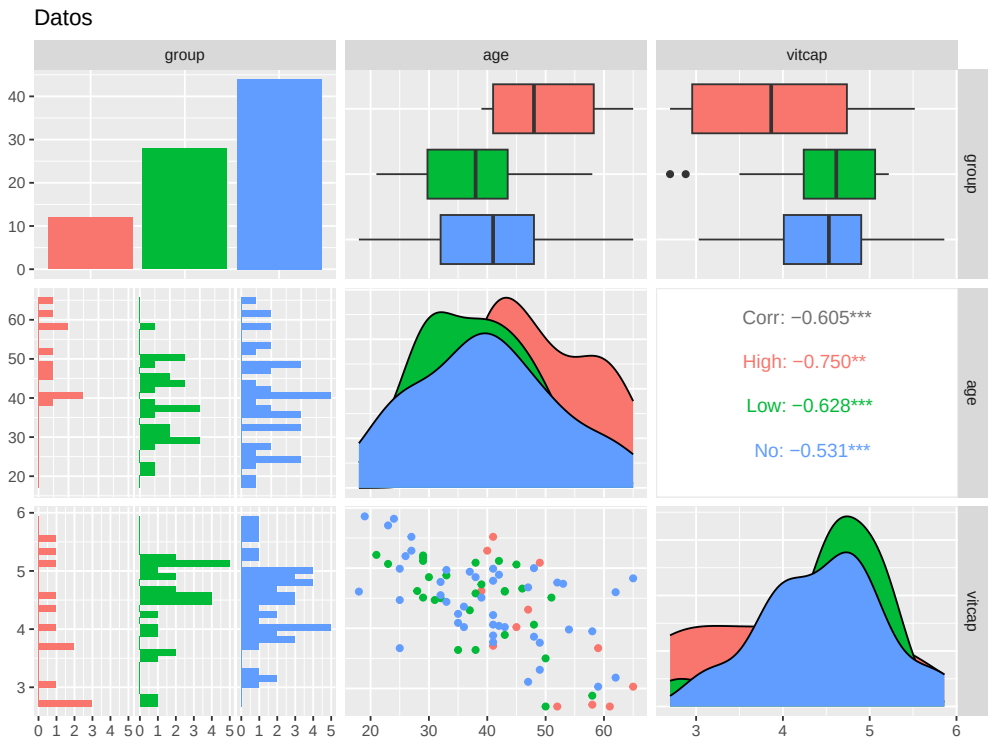


Figure 2: Some caption.

```
> #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)
```

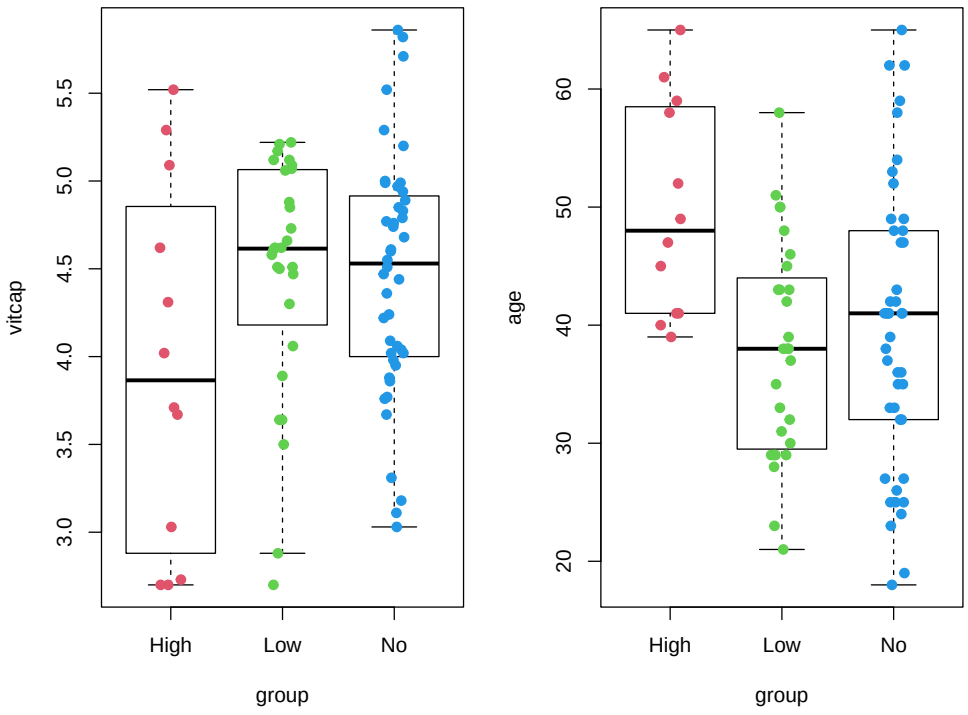


Figure 3: Some caption.

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)
```

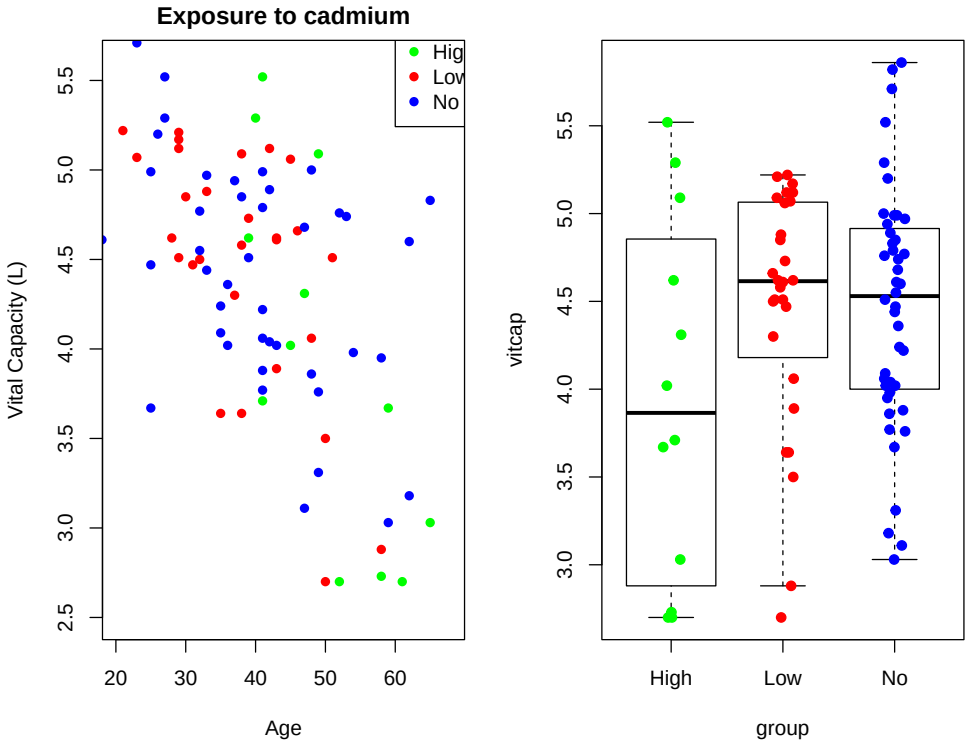


Figure 4: Some caption.

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.

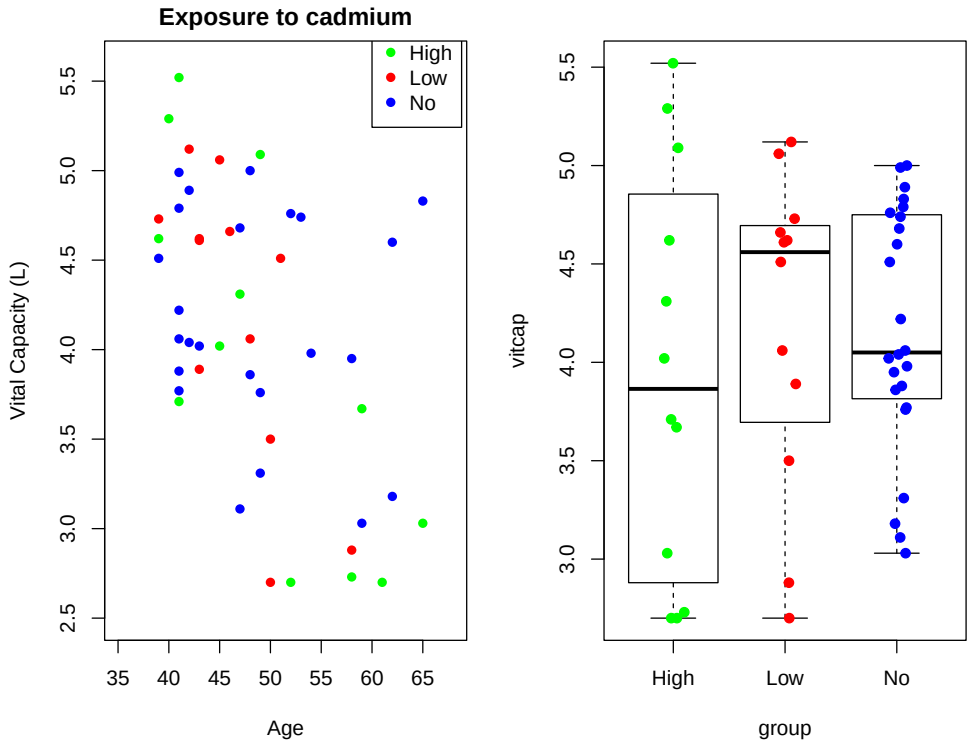
```
> summary(CADdata[CADdata$group=="High",])
```

```
 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)
```



Veamos 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".

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

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

En un problema tipo ANCOVA se inicia con el modelo con interacciones y despues se trata de hacer hipotesis para simplificarlo, este modelo se conoce como modelo "saturado", pues cuenta con todas las variables y todas las posible interacciones.
Ajuste del modelo con interacciones:
E(y;x)= b0 + b1 age + b2 Low + b3 No + b4(age*Low) + b5(age*No)
Para cada valor de las variables dummy se tiene lo siguiente:
E(Y;group=High, age)= b0 + b1 age
E(Y;group= Low, age)= b0 + b1 age +b2 + b4 age = (b0 + b2) + (b1 + b4) age
E(Y;group= No, age)= b0 + b1 age +b3 + b5 age = (b0 + b3) + (b1 + b5) age
Este modelo considera tres diferentes rectas, una para cada grupo

```
> fit <- lm(vitcap ~ age * group, data = CADdata)
> summary(fit)
```

```
Call:
lm(formula = vitcap ~ age * group, data = CADdata)

Residuals:
    Min      1Q  Median      3Q     Max
-1.1017 -0.3456 -0.0324  0.5152  1.0770

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    8.1834     1.0641    7.69  1.5e-09 ***
age           -0.0851     0.0211   -4.04  0.00022 ***
groupLow       1.2369     2.0332    0.61  0.54622
groupNo       -3.3940     1.3446   -2.52  0.01547 *
age:groupLow  -0.0273     0.0426   -0.64  0.52586
age:groupNo    0.0722     0.0269    2.69  0.01031 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.636 on 42 degrees of freedom
Multiple R-squared:  0.394,        Adjusted R-squared:  0.321
F-statistic: 5.45 on 5 and 42 DF,  p-value: 0.000592
```

Despues de este paso lo inmediato es la validación de supuestos, aunque en este ejemplo lo dejaremos para más adelante.

El paso siguiente es hacer la prueba asociada a la tabla ANOVA. De la salida de R vemos que se rechaza H0 en la prueba asociada a la tabla ANOVA con un p-value de 0.000592, entonces el modelo tiene sentido, es decir, es decir, la edad o el nivel de exposición ayudan a modelar E(Y)

Buscando simplificar el modelo (que las rectas tengan la misma pendiente). Realizaremos la prueba de igualdad de pendientes i.e. coeficientes asociados a las interacciones.
Si no se rechaza H0 se puede optar por un modelo con igualdad de pendientes o rectas paralelas (facilita la interpretación).
H0:b4=0 y b5=0 vs Ha: b4!=0 o b5!=0

```
> library(multcomp)
> K=matrix(c(0,0,0,0,1,0,
+            0,0,0,0,0,1), ncol=6, nrow=2,byrow=TRUE)
> m=c(0,0)
> summary(glht(fit, linfct=K, rhs=m),test=Ftest())
```

```
         General Linear Hypotheses

Linear Hypotheses:
       Estimate
1 == 0  -0.0273
2 == 0   0.0722

Global Test:
     F DF1 DF2 Pr(>F)
1 5.27   2  42 0.0091
```

De la salida de R se rechaza H0 (p-value:0.0091) no se puede considerar el modelo con rectas paralelas. Al menos una de las rectas tiene distinta pendiente. Ahora usaremos pruebas de hipotesis simultáneas buscando reducir el modelo.
Procedemos a ver si las tres rectas tienen pendiente diferente, con la prueba simultánea que sigue a la prueba lineal general.
Notar que aquí se incluyen las 3 hipótesis individuales asociadas a:
H0_1: pendientes de High y Low iguales (b4=0)
H0_2: pendientes de High y No iguales (b5=0)
H0_3: pendientes de Low y No iguales (b4-b5=0)
Esta prueba sólo detecta las diferencias más evidentes, la lectura se debe hacer con cuidado considerando que esta prueba tiene más chance de cometer el error tipo II.
¿Todas las pendientes difieren?
Al omitir el argumento test=Ftest() en la función summary() nos arrojara las pruebas simultaneas.

```
> K=matrix(c(0,0,0,0,1, 0,
+            0,0,0,0,0, 1,
+            0,0,0,0,1,-1), ncol=6, nrow=3, byrow=TRUE)
> m=c(0,0,0)
> summary(glht(fit, linfct=K, rhs=m))
```

```
         Simultaneous Tests for General Linear Hypotheses

Fit: lm(formula = vitcap ~ age * group, data = CADdata)

Linear Hypotheses:
       Estimate Std. Error t value Pr(>|t|)
1 == 0  -0.0273     0.0426   -0.64    0.795
2 == 0   0.0722     0.0269    2.69    0.026 *
3 == 0  -0.0995     0.0406   -2.45    0.046 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

La interpretación de la salida de R es que se rechaza la prueba de hipótesis simultáneas si al menos uno de los p-values reportados son <0.05
De forma simultánea podemos identificar que se rechaza H0 por dos diferencias:
1. Se rechaza b5=0, pues p-value < 0.026 Hay evidencia sobre pendientes de High y No diferentes
2. Se rechaza b4-b5=0, pues p-value < 0.046 Hay evidencia sobre pendientes de Low y No diferentes
Con esta prueba no se rechaza b4=0, por lo que podríamos optar por considerar un modelo con b4=0 , es decir sin la interacción (age*Low)
La utilidad de esta prueba es que nos dice donde esta la mayor evidencia que hizo que se rechazara la prueba.

Ajustemos el modelo reducido que no considera la intereacción (age*No)
Este nuevo modelo corresponde a
E(y;x)= b0 + b1 age + b2 Low + b3 No + b4(age*No)
La esperanza dado cada uno de los valores de la variables group:
E(Y;group=High, age)= b0 + b1 age

E(Y;group= Low, age)= b0 + b1 age +b2 = (b0 + b2) + b1 age
E(Y;group= No, age) = b0 + b1 age +b3 + b4 age = (b0 + b3) + (b1 + b4) age

```
> fitred <- lm(vitcap ~ age + group + I(age*(group=="No")), data = CADdata)
> summary(fitred)
```

```
Call:
lm(formula = vitcap ~ age + group + I(age * (group == "No")),
    data = CADdata)

Residuals:
    Min      1Q  Median      3Q     Max
-1.1738 -0.3758 -0.0677  0.5384  1.0720

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)                8.5149     0.9230    9.22  9.4e-12 ***
age                       -0.0918     0.0182   -5.05  8.7e-06 ***
groupLow                  -0.0524     0.2647   -0.20   0.8439
groupNo                   -3.7255     1.2323   -3.02   0.0042 **
I(age * (group == "No"))   0.0789     0.0246    3.20   0.0026 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.632 on 43 degrees of freedom
Multiple R-squared:  0.388,        Adjusted R-squared:  0.331
F-statistic: 6.81 on 4 and 43 DF,  p-value: 0.000245
```

Se rechaza H0 en la prueba asociada a la tabla ANOVA (p-value: 0.000245) por lo cual al menos una de las betas es distinta de 0

Ahora podríamos proceder a reducir un poco más el modelo para tratar de facilitar la interpretación o bien a contestar las preguntas sobre los investigadores.
De la salida de fitred, se puede observar que una opción es considerar b2=0 pues la prueba t asociada no rechaza la hipotesis de b2=0
El modelo reducido quedaría como:
E(y;x)= b0 + b1 age + b2 No + b3(age*No)
Condicionando al valor de la variable group quedaría como:
E(Y;group=High, age)= b0 + b1 age
E(Y;group= Low, age)= b0 + b1 age
E(Y;group= No, age)= b0 + b1 age +b2 + b3 age = (b0 + b2) + (b1 + b3) age
Las rectas asociadas a la exposicion "High" y "Low" son la misma, por otra parte, la exposición de tipo "No" tiene su propia recta.

```
>  fitred2 <- lm(vitcap ~ age + I(group=="No") + I(age*(group=="No")), data =CADdata)
>  summary(fitred2)
```

```
Call:
lm(formula = vitcap ~ age + I(group == "No") + I(age * (group ==
    "No")), data = CADdata)

Residuals:
    Min      1Q  Median      3Q     Max
-1.2015 -0.3863 -0.0677  0.5428  1.0975

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)                8.4499     0.8534    9.90  9.1e-13 ***
age                       -0.0910     0.0175   -5.19  5.2e-06 ***
```

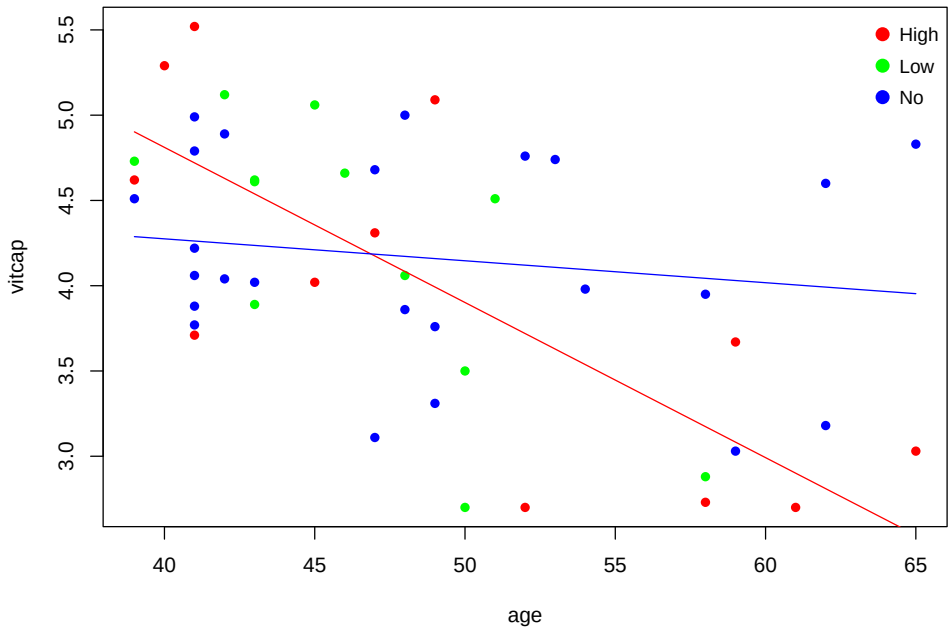
```
I(group == "No")TRUE      -3.6605     1.1748   -3.12   0.0032 **
I(age * (group == "No"))   0.0781     0.0240    3.25   0.0022 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.625 on 44 degrees of freedom
Multiple R-squared:  0.387,        Adjusted R-squared:  0.345
F-statistic: 9.27 on 3 and 44 DF,  p-value: 7.26e-05
```

Se rechaza H0 en la prueba asociada a la tabla ANOVA (p-value: 7.26e-05). También se puede observar que no se podría reducir más el modelo.
Con este modelo reducido ya se puede trabajar.

Las rectas ajustadas son:
$\hat{E}$(Y;group=High, age)= 8.4499 - 0.0910 age
$\hat{E}$(Y;group=Low, age) = 8.4499 - 0.0910 age
$\hat{E}$(Y;group=No, age)= (8.4499-3.6605)+(-0.0910+0.0781) age = 4.789 - 0.0129 age
Se puede observar que la pendiente de los expuestos a cadmio es mayor en valor absoluto y negativa con lo que podemos interpretar que la capacidad vital decrece más rápido en ese grupo comparado con los no expuesto.
Esto se puede observar más fácilmente en una gráfica

```
> fitredHyL <- function(X2) {fitred2$coefficients[1]+ fitred2$coefficients[2]*X2}
> fitredN <- function(X2) {fitred2$coefficients[1]+fitred2$coefficients[3]+
+     (fitred2$coefficients[2]+fitred2$coefficients[4])*X2}
```

```
> with(CADdata, plot(age, vitcap, col=c("red", "green", "blue")[CADdata[,1]] ,
+                    pch = c(16,16,16)))
> legend("topright",levels(CADdata[,1]), col=c("red", "green", "blue"),
+        pch =c(16,16,16), pt.cex=1.5,cex = .9, y.intersp = 1.4 , bty="n" )
> curve(fitredHyL, from = min(CADdata$age), to = max(CADdata$age),
+ col = "red", add = T)
> curve(fitredN, from = min(CADdata$age), to = max(CADdata$age),
+ col = "blue", add = T)
```



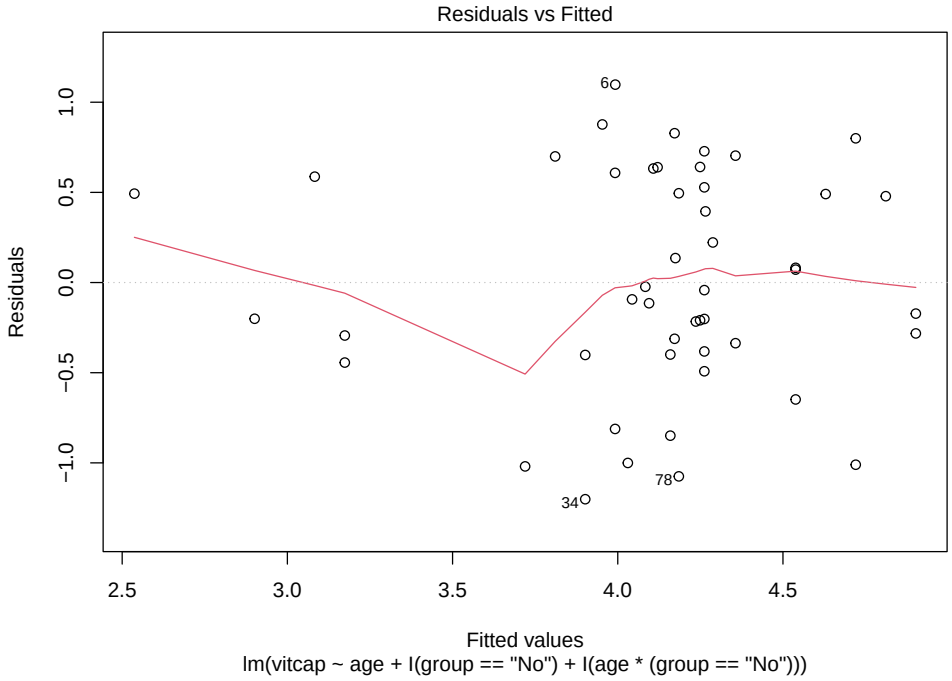
Esta grafica faltaria complementarla no es la definitiva

Toda esta interpretación se basa en que se cumplan los supuestos, veamos una revisión rápida de los mismos.

SUPUESTO DE LINEALIDAD

Este supuesto permite validar si hay algún problema con el modelado de la esperanza de y. En el eje x está la estimación de Y, en el eje y están los errores observados (residuales o diferencias entre lo real y lo estimado). Lo ideal seria que los datos estén sobre la linea punteada (y=0), que no haya curvas que suban o bajen alrededor del 0. En este caso no hay tales curvas, lo unico es que que una recta baja pero se debe a que se tienen pocos datos.

```
>  plot(fitred2, 1)
```



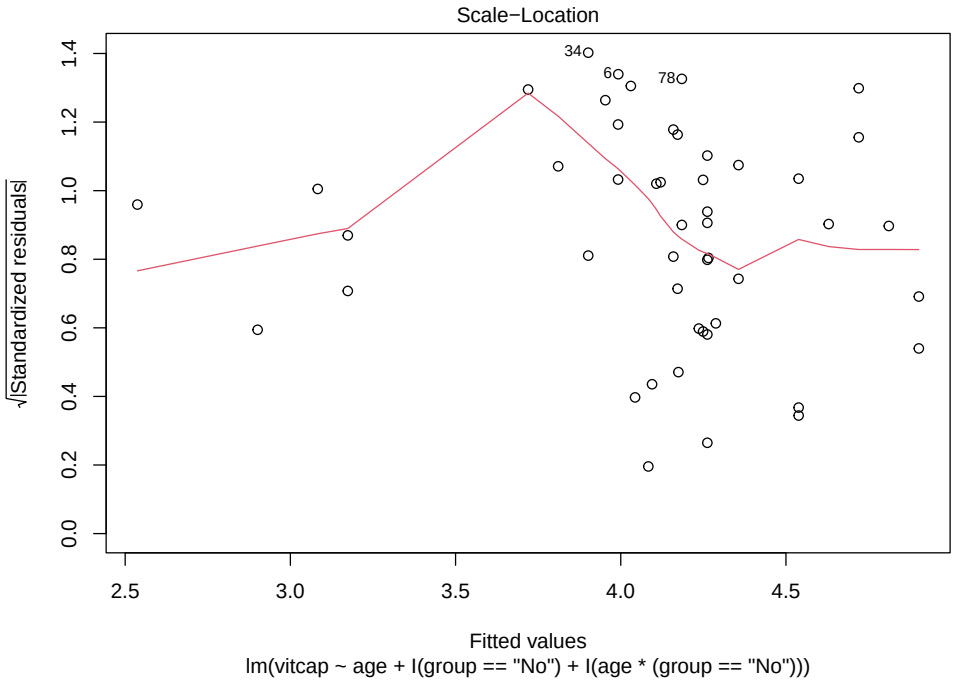
SUPUESTO DE HOMOCEDASTICIDAD

Nota: Eje Y representa los residuales estandarizados/estudentizados y se les aplica raiz del valor absoluto. El eje X representa la Y estimada.

Si se cumple la homocedasticidad las nubes de puntos deben estar encerradas entre bandas horizontales, y para cada valor del eje x, los puntos estarían cayendo entre esas bandas.

En este caso vemos que en el eje X valores menores a 3.5 se tienen puntos no tan dispersos verticalmente, para valores de x mayores a 3.5 los puntos tienen mayor variabilidad, esto nos deja en duda, mas adelante veremos algunas pruebas de hipótesis para validar este supuesto.

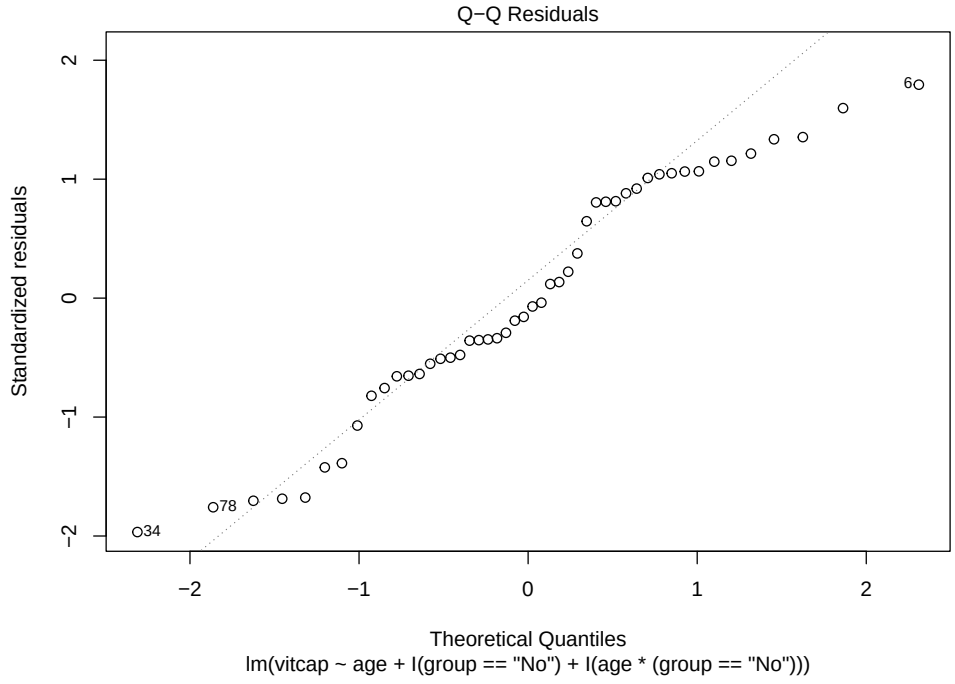
```
> plot(fitred2, 3)
```



SUPUESTO DE NORMALIDAD

Hay algunos puntos que se salen pero son pocos, los demás si se ajustan bien. La evidencia en contra no es tan fuerte.

```
> plot(fitred2, 2)
```



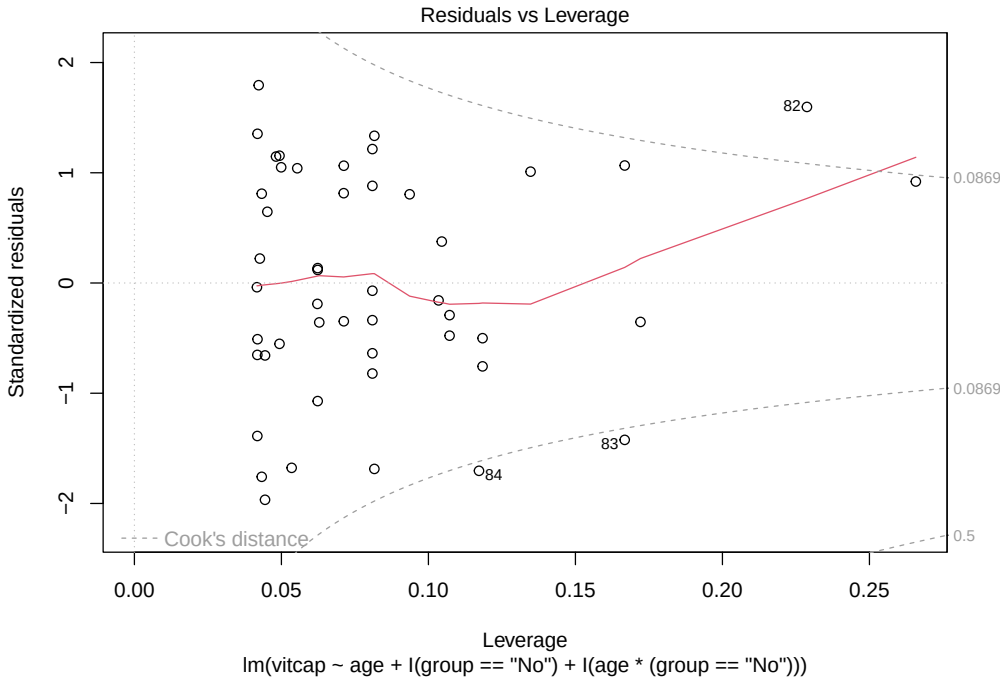
Además de estos supuestos es importante revisar los posibles outliers

En el eje X se grafica la medida "leverage" que indica qué tan alejados están los puntos de la media de la variable, entre más alejados pueden tener mas impacto. En el eje Y se grafican los errores estandarizados. Lo deseable seria que en las esquinas superior e inferior derecha no haya muchos puntos que se salgan.

Las lineas punteadas en forma de embudo representan la distancia de Cook, estas son las de libro y son mas conservadoras, las lineas punteadas mas anchas representan la distancia de Cook del software R las cuales son mas laxas (En la grafica no solo se alcanza a apreciar la del la esquina inferior derecha).

En base a esto deberiamos revisar la observacion 82

```
> plot(fitred2, 5, cook.levels = c(4/(dim(CADdata)[1]-2), 0.5, 1.0))
```



Con lo que hemos visto hasta ahora no encontramos evidencia muy fuerte en contra del cumplimiento de los supuestos. Recordemos que en este tipo de problemas nos interesa que no haya evidencia fuerte para poder seguir trabajando.