Regrsión lineal multiple

IVAN ALDAIR CONDE SALINAS

```
rm(list = ls(all.names = TRUE))
gc()
```

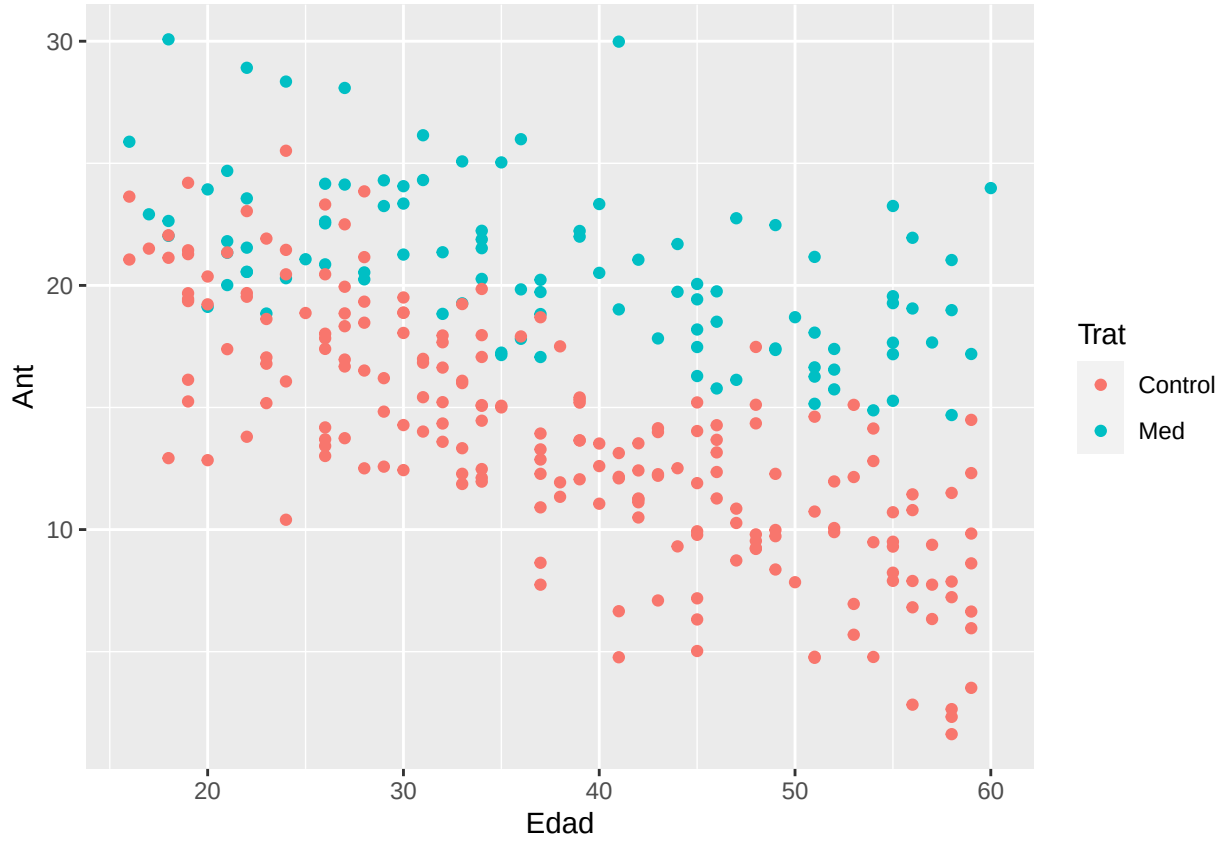
```
##          used (Mb) gc trigger (Mb) max used (Mb)
## Ncells 420285 22.5     878958   47   643711 34.4
## Vcells 767528  5.9    8388608   64  1712058 13.1
```

```
setwd('C:/Users/aldai/OneDrive/Documentos/R')
datos <- read.csv("Ex5.csv")
str(datos)
```

```
## 'data.frame':    300 obs. of  3 variables:
##  $ Ant : num  24.3 21.2 22.2 15.3 17.7 ...
##  $ Trat: chr  "Med" "Med" "Med" "Med" ...
##  $ Edad: int  29 51 34 55 57 18 39 55 40 36 ...
```

```
datos$Trat=factor(datos$Trat)

library(ggplot2)
ggplot(datos, aes(Edad, Ant, color = Trat)) +
  geom_point()
```



I) ANÁLISIS DESCRIPTIVO

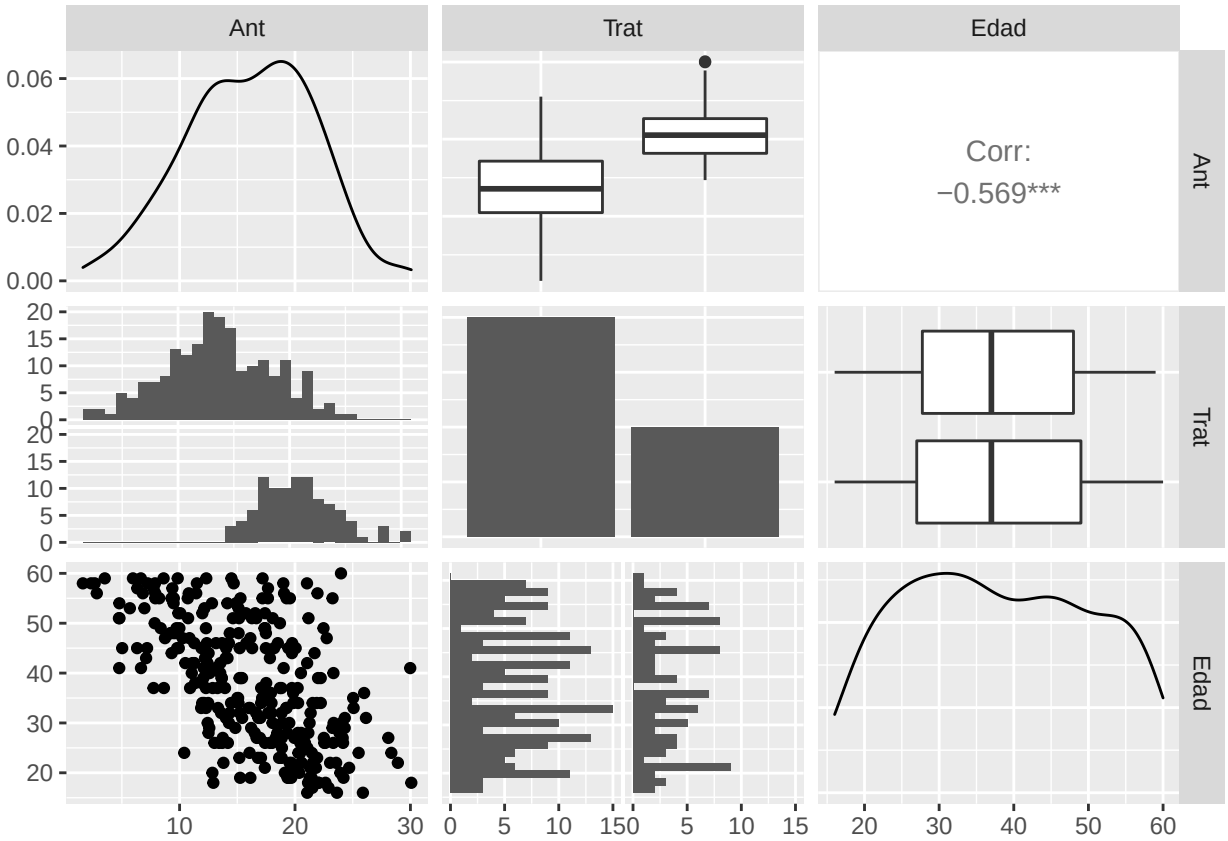
```
library(GGally)
```

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

```
ggpairs(datos)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Los datos nos proporcionan una variable categórica (Trat) con dos niveles ("Control","Med") y unas variables continuas (Edad). La correlación de la variable de respuesta con la continua es considerablemente alta, lo cual podemos notar en el gráfico de dispersión.

La distribución de la variable respuesta respecto a las categóricas, se puede observar que el comportamiento es bastante similar en ambas categorías. Las medianas son distintas y se encuentran alrededor de los 17 y el rango también es diferente

II) ajuste de modelo

```
fit <- lm(Ant ~ Edad * Trat, data = datos)
summary(fit)
```

```
##
## Call:
## lm(formula = Ant ~ Edad * Trat, data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.9856 -1.9133 -0.0627  1.8837  9.7037
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  25.43075    0.68736  36.998  < 2e-16 ***
## Edad         -0.30914    0.01722 -17.956  < 2e-16 ***
```

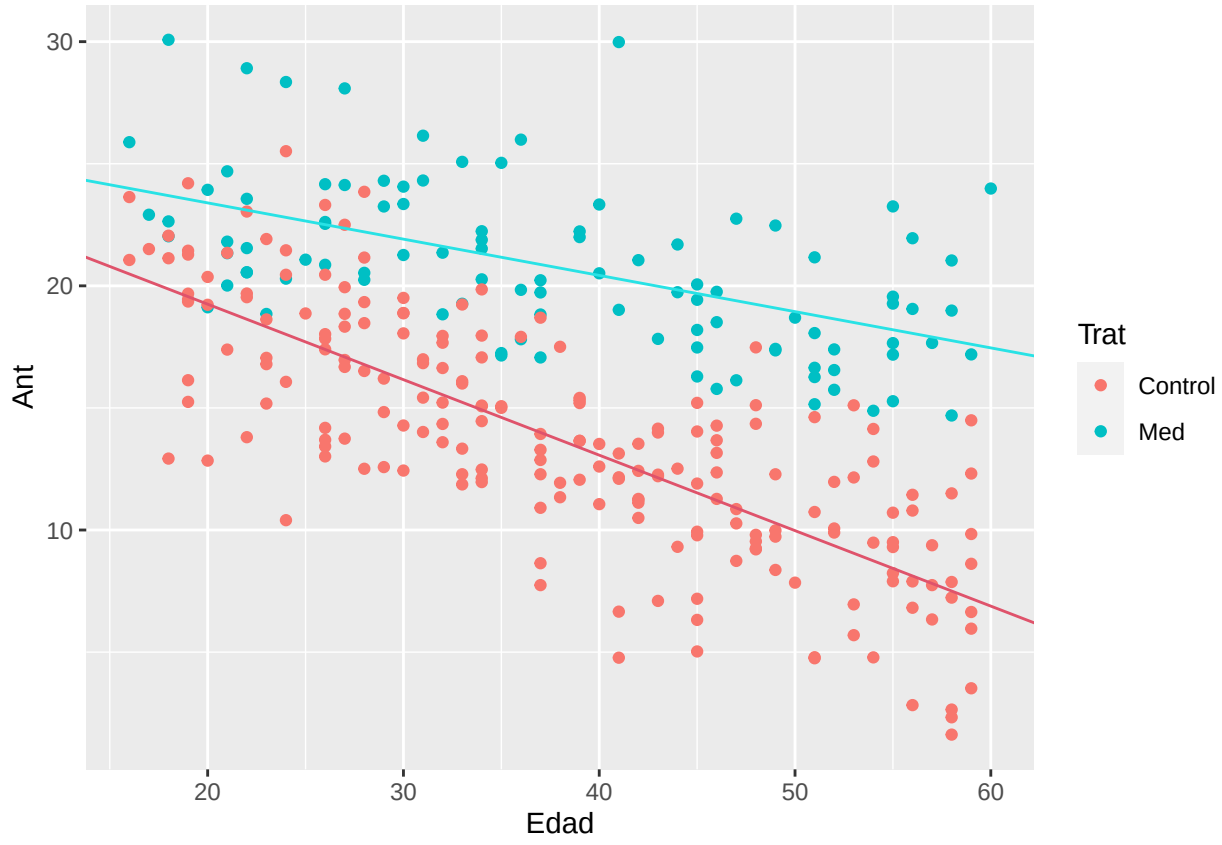
```
## TratMed       0.93511    1.17635   0.795    0.427
## Edad:TratMed  0.16069    0.02947   5.452 1.05e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.978 on 296 degrees of freedom
## Multiple R-squared:  0.7129, Adjusted R-squared:   0.71
## F-statistic:   245 on 3 and 296 DF,  p-value: < 2.2e-16
```

Se rechaza H_0 en la prueba asociada a la tabla ANOVA, por lo tanto, el modelo tiene sentido.

El modelo con interacción es

$$Ant = \beta_0 + \beta_1 Edad + \beta_3 Trat + \beta_4 Edad * Trat$$

```
ggplot(datos, aes(Edad, Ant, color = Trat)) +
  geom_point() +
  geom_abline(intercept = (fit$coefficients[1] + fit$coefficients[3]) ,
              slope = fit$coefficients[2] + fit$coefficients[4],color=5) +
  geom_abline(intercept = (fit$coefficients[1] ) ,
              slope = fit$coefficients[2],color=2)
```



III) promedio de anticuerpos Recordemos que el modelo es

$$Ant = \beta_0 + \beta_1 Edad + \beta_3 Trat + \beta_4 Edad * Trat$$

Para el grupo de control $Trat = 0$, entonces $E(Ant; Trat = control, Edad) = \beta_0 + \beta_1 Edad$

Para el de tratamiento medico $Trat = 1$, entonces

$$E(Ant; Tat = Med, Edad) = \beta_0 + \beta_1 Edad + \beta_2 + \beta_3 Edad = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) Edad$$

IV) Edad afecta igual a la generación de anticuerpos en el grupo control que en el grupo que recibe el medicamento

El cambio en el valor promedio de los anticuerpos que se obtiene al aumentar en una unidad la variable Edad en el tratamiento médico es

$$E(Ant; Trat = Med, Edad{+}1) - E(Ant; Trat = Med, Edad) = (\beta_0{+}\beta_2){+}(\beta_1{+}\beta_4)(Edad{+}1) - ((\beta_0{+}\beta_2){+}(\beta_1{+}\beta_4)(Edad)) = \beta$$

es decir, $\beta_1 + \beta_3$ es el promedio del cambio en la generación de anticuerpos al aumantar la edad que tiene el paciente en el grupo al que se le aplicó tratamiento médico.

Para el tratamiento de control

$$E(Ant; Trat = Control, Edad{+}1) - E(Ant; Trat = Control, Edad) = \beta_0{+}\beta_1(Edad{+}1) - (\beta_0{+}\beta_1(Edad)) = \beta_1$$

Entonces, β_1 es el promedio del cambio en la generación de anticuerpos al cambiar la edad en un año al paciente en un grupo de control.

Para saber si la edad afecta de la misma forma la generación de anticuerpos en el grupo control que en el grupo que recibe el medicamento, la hipótesis nula propuesta es $H_o: \ \beta_1 = \beta_1 + \beta_3$ que equivale a

$$H_0 : \beta_3 = 0 \quad \text{vs} \quad \beta_3 \neq 0$$

$$Ant = \beta_0 + \beta_1 Edad + \beta_2 Trat + \beta_3 Edad * Trat$$

```
library(multcomp)
```

```
## Loading required package: mvtnorm
```

```
## Loading required package: survival
```

```
## Loading required package: TH.data
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'TH.data'
```

```
## The following object is masked from 'package:MASS':
##
##     geyser
```

```r
K=matrix(c(0,0,0,1), ncol=4, nrow=1, byrow=TRUE)
m=c(0)
summary(glht(fit, linfct=K, rhs=m), test=Ftest())
```

```
##
##   General Linear Hypotheses
##
## Linear Hypotheses:
##        Estimate
## 1 == 0   0.1607
##
## Global Test:
##       F DF1 DF2    Pr(>F)
## 1 29.73   1 296 1.049e-07
```

Se rechaza H_0, encontramos evidencia estadística que indica que el coeficiente asociado a la interacción entre las variables edad y tratamiento (β_3) es significativo. Es decir, encontramos evidencia que nos dice que la edad afecta de distinta forma la generación de anticuerpos en el grupo control que en el grupo que recibe el medicamento.

v) Ajuste del modelo incluyendo la interpretación de cada uno de los coeficientes.

El modelo ajustado es

$$Ant = 25.43075 - 0.3091356Edad + 0.9351099Trat + 0.1606855Edad * Trat$$

el coeficiente $\beta_0 = 25.43075$ nos dice la cantidad de anticuerpos que tiene una persona de cero años dado que esta en el grupo de control. Como ya se había mencionado anteriormente, el coeficiente $\beta_1 = -0.3091356$ es el promedio del cambio en la generación de anticuerpos al aumentarle la edad en un año al paciente en un grupo de control.

El coeficiente $\beta_2 = 0.9351099$ Al ver el summary vemos la posibilidad de que el coeficiente β_2=tratMed sea 0 porque tiene una significancia baja.

```r
library(multcomp)
K=matrix(c(0,0,1,0), ncol=4, nrow=1, byrow=TRUE)
m=c(0)
summary(glht(fit, linfct=K, rhs=m), test=Ftest())
```

```
##
##   General Linear Hypotheses
##
## Linear Hypotheses:
##        Estimate
## 1 == 0   0.9351
##
## Global Test:
##        F DF1 DF2 Pr(>F)
## 1 0.6319   1 296 0.4273
```

Rechaza la prueba F, por lo tanto, hemos encontrado evidencia que nos permite tomar el coeficiente $\beta_2 = 0$, por lo tanto, no tiene sentido interpretarlo.

El coeficiente β_3 se puede interpretar como cuánto cambia el promedio de anticuerpos en un año al cambiar de un grupo a otro y como vimos en el inciso anterior es distinto de cero.

```
fitred <- lm(Ant ~ Edad  + I(Edad*(Trat=="Med")), data = datos)
summary(fitred)
```

```
##
## Call:
## lm(formula = Ant ~ Edad + I(Edad * (Trat == "Med")), data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.9928 -1.9251 -0.1039  1.8968  9.7185
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                25.750017   0.557464   46.19   <2e-16 ***
## Edad                       -0.316748   0.014299  -22.15   <2e-16 ***
## I(Edad * (Trat == "Med"))   0.182958   0.009131   20.04   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.976 on 297 degrees of freedom
## Multiple R-squared:  0.7123, Adjusted R-squared:  0.7104
## F-statistic: 367.7 on 2 and 297 DF,  p-value: < 2.2e-16
```

El modelo reducido con $\beta_2* = \beta_3$ es

$$Ant = \beta_0 + \beta_1 Edad + \beta_2 * Trat * Med$$

Todos los coeficinetes nos aportan al modelo

vi) Argumente en contra o a favor de la afirmación:

```
edad <- seq(from = 16, to = 60, by = .5)
length(edad)
```

```
## [1] 89
```

Para una banda para la recta del grupo de control $E(Y; Trat = control, Edad) = b0 + b1Edad$

```
Kc <- cbind(1, edad, 0)
```

Para una banda para la recta del tratamiento médico $E(Y; Tat = med, Edad) = b0 + b1Edad + b2Edad = (b0 + b2) + (b1 + b4)Edad$

```
Km <- cbind(1, edad, edad)
```

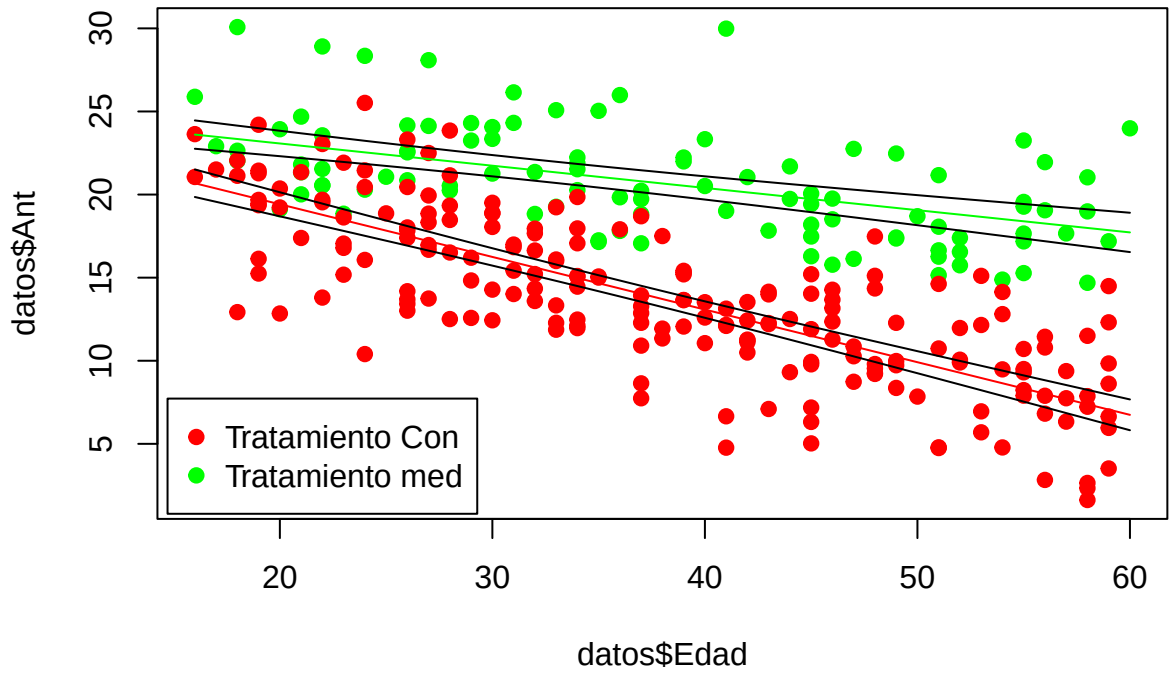
```
K=rbind(Kc, Km)

fitE <- glht(fitred, linfct = K)
fitci <- confint(fitE, level = 0.90)
```

visualización del modelo

```
plot(datos$Edad, datos$Ant, pch=19, col = c("red", "green")[datos$Trat])
legend("bottomleft", c("Tratamiento Con", "Tratamiento med"),
       col = c("red", "green"), pch = 19, inset = 0.01)
lines(edad, coef(fitE)[1:89], col="red")
lines(edad, fitci$confint[1:89,"upr"])
lines(edad, fitci$confint[1:89,"lwr"])

lines(edad, coef(fitE)[90:178], col="green")
lines(edad, fitci$confint[90:178,"upr"])
lines(edad, fitci$confint[90:178,"lwr"])
```



Con ayuda de la prueba de hipótesis simultanea, con un nivel del 90%, podemos afirmar que el medicamento aumenta el número de anticuerpos, pues entre las edades de 16 y 60 años la banda de confianza del tratamiento se encuentra por encima de la banda de confianza del grupo de control.