UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Capítulo 20

Regresión Lineal Múltiple

Jordi Cortés, José Antonio González y Erik Cobo
Pilar Muñoz, Ángel Ruiz y Nerea Bielsa

Marzo 2015

MEDICINA CLINICA
equator network
TRIALS
ELSEVIER DOYMA
Fundació Patronat Científic
Col·legi de Metges Illes Balears
CLÍNIC BARCELONA Hospital Universitari
ASOCIACION ESPAÑOLA DE FISIOTERAPEUTAS
Departament d'Estadística i Investigació Operativa
UNIVERSITAT POLITÈCNICA DE CATALUNYA

Regresión Lineal Múltiple

Regresión Lineal Múltiple 2

Presentación 3

1 Modelo general 4

2 Modelo con un factor categórico 5

2.1 Más de 2 categorías * 9

2.2 Codificaciones alternativas de las variables cualitativas * 10

3 Modelo con 2 factores 12

3.1 Modelo con interacción entre 2 factores 18

4 Modelo aditivo para una numérica y una categórica. 23

4.1 Modelo con interacción entre numérica y categórica* 25

5 Utilidad en investigación de salud 28

5.1 La predicción puede ser simultánea o futura. 29

5.2 Coeficientes ajustados: Especulando sobre cómo cambiar el futuro 30

5.3 Confusión entre pronóstico y etiología 30

Soluciones a los ejercicios 32

Presentación

Este capítulo modela una variable respuesta Y con cierta combinación de variables predictoras que podrán ser X (intervenciones) o Z (condiciones). El principal objetivo de este capítulo es interpretar sus coeficientes según estas predictoras sean numéricas o categóricas. Estudia, además, el caso en el que el coeficiente de un predictor es el mismo para cada valor de las restantes predictoras (modelo aditivo) y el caso en el que cambia (modelo con interacción o multiplicativo).

Por ejemplo, la Figura 1 contiene el ajuste de un modelo de regresión múltiple en un ensayo clínico de artritis. La variable respuesta Y es el ángulo de flexión del codo al final del seguimiento y las 2 variables predictoras son el tratamiento y el ángulo de flexión previo. También aparece un término de interacción entre estas dos variables que responde a la pregunta de si el efecto del tratamiento es común ("independiente") para cualquier flexión previa.

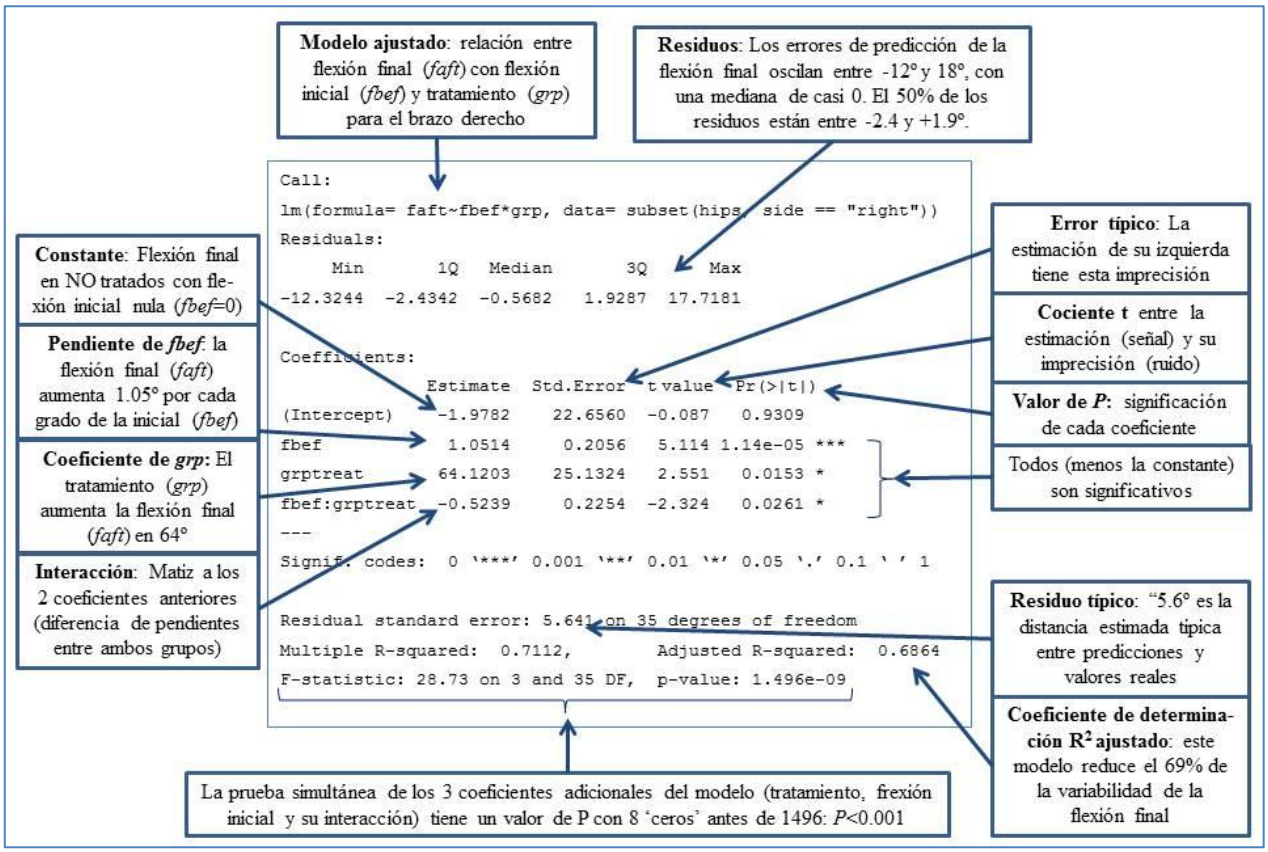


Figura 1: Interpretación de la regresión del nivel de flexión final (*faft*) según su nivel inicial (*fbef*), el tratamiento (*grp*) y su interacción en el codo derecho (*side=="right"*).

Este tema aborda la interpretación y significación estadística de los resultados. El siguiente estudia las premisas y cautelas necesarias para interpretarlos.

Contribuciones: basado en apuntes previos elaborados por PM, JAG, JC y EC; AR, JC y EC lo actualizaron; JAG lo revisó y NB lo editó.

1 Modelo general

En Regresión Lineal Simple (RLS) el modelo contiene una variable predictora:

$$Y = \beta_0 + \beta_1 Z + \varepsilon$$

La Regresión Lineal Múltiple (RLM) suma las contribuciones lineales de **k** predictoras:

$$Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \cdots + \beta_k Z_k + \varepsilon$$

Ejercicio 1.1

¿Cuántas 'betas' β tiene el modelo lineal con k variables?

Igual que en RLS, su traslado a los valores muestrales origina 'n' ecuaciones para los 'n' casos, representados por el subíndice 'i':

$$y_i = b_0 + b_1 z_{1i} + b_2 z_{2i} + \cdots + b_k z_{ki} + e_i$$

Ejercicio 1.2

Diga en palabras qué significa cada símbolo: Y, β, β_0, β_1, Z_1, β_k, Z_k, ε, y_i, b_0, b_1, z_{1i}, b_k, z_{ki}, e_i.

Además, podrán aparecer términos de interacción entre dos predictoras: Por ejemplo, ¿el efecto del tratamiento X en la respuesta Y depende del género Z? O también, ¿la relación de la presión arterial inicial Z con la final Y es la misma para todos los estratos de glicemia inicial?

A continuación, veremos la RLM con 2 predictoras según sean numéricas o factores dicotómicos.

NOTA: Es habitual llamar a la variable respuesta Y 'dependiente', precisamente porque es la que 'depende' de las variables Z, y a éstas últimas 'independientes', porque en la ecuación no dependen de terceras variables y sus contribuciones pueden sumarse, son "aditivas". Aquí evitamos esta ambigüedad usando el término de variables predictoras. Cuando se trate de intervenciones discutiremos si esta predicción de Z en Y puede elevarse a 'efecto' de X en Y.

Ejercicio 1.3

Busque en ambos documentos de STROBE (la declaración y el largo explicativo) cuántas veces aparece el término "independent variable". Repita en el doc E&E de TRIPOD. ¿Y cuántas aparece 'predictor' en la 1ª página de este último?

Recuerde

Use variable 'predictora' en lugar de 'independiente'.

Progresivamente iremos introduciendo variables en este modelo.

Historieta: En este capítulo, las predictoras serán "independientes" entre ellas. En el siguiente, lo generalizamos a variables independientes que no son independientes entre sí.

2 Modelo con un factor categórico

En el caso de una categoría con 2 valores, el modelo es:

$$Y = \beta_0 + \beta_1 Z_1 + \varepsilon$$

La variable Z_1 representa un factor con dos categorías. Para convertirla en "numérica", usamos un indicador (*dummy*) que valdrá 0 para la categoría de referencia y 1 para la otra.

Ejemplo 2.1: Sea Y la altura en centímetros de adultos sanos de Barcelona. Sea Z la variable género: mujer (0) y hombre (1). Se propone el siguiente modelo:

$$y_i = 165 + 10 \cdot z_i + e_i$$

La altura para una mujer vendrá dada por la substitución z_i por el valor 0:

$$y_i = 165 + 10 \cdot 0 + e_i = 165 + e_i \rightarrow Modelo\ para\ las\ mujeres$$

$$y_\iota = 165 \rightarrow Valor\ predicho\ para\ las\ mujeres$$

Para los hombres, la altura esperada es 175 cm.

$$y_i = 165 + 10 \cdot 1 + e_i = 175 + e_i \rightarrow Modelo\ para\ los\ hombres$$

$$y_\iota = 175 \rightarrow Valor\ predicho\ para\ los\ hombres$$

Ejemplo 2.2: El conjunto de datos *hips* del paquete *faraway* de R sobre Espondilitis Anquilosante contiene los grados de flexión de los codos antes y después de cierto tratamiento en 39 pacientes. La Figura 2.1 muestra la recta que pasa por las medias de la flexión final derecha en ambos grupos de tratamiento.

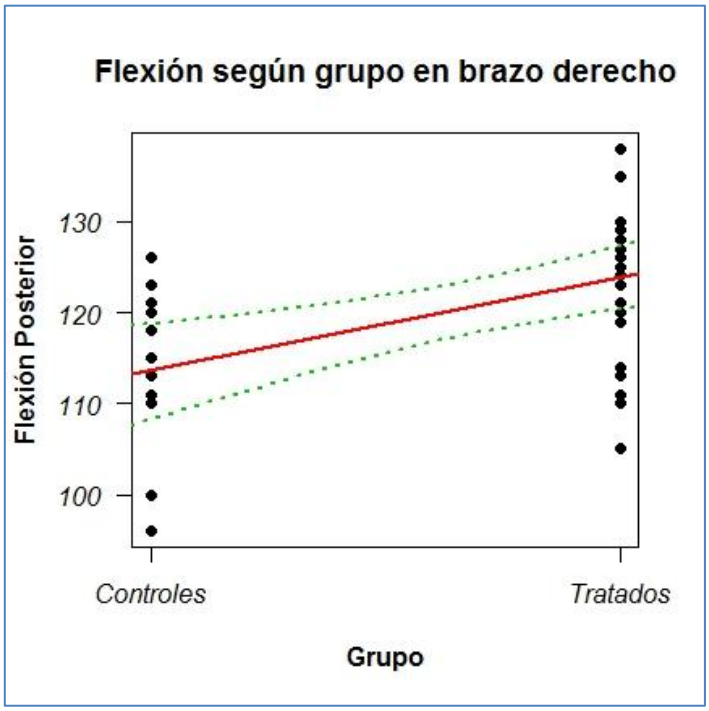


Figura 2.1. Flexión final en brazo derecho según tratamiento. La linea (roja) es la recta de regresión estimada con sus bandas de confianza (verde).

Ejercicio 2.1

En la regresión del grado de flexión final según el grupo de tratamiento, diga qué representan los 2 coeficientes del modelo.

```
>install.packages("faraway")
>library("faraway")
>hips
Call:
lm(formula = faft ~ grp, data = hips)
> lm(faft~grp,data=hips)
Coefficients:
(Intercept)     grptreat
     113.75        10.23
```

Recuerde

Si usa el indicador 0, 1, la constante es la media en el grupo de referencia (codificado 0) y la pendiente, la diferencia entre ambas medias.

Nota: Más adelante se comentan las implicaciones de codificaciones alternativas.

Ejemplo 2.1 (cont): Vimos que una mujer tendrá una altura esperada de 165 cm. Sin embargo, no todas las mujeres miden 165 cm. De aquí, la presencia del término *particular* ε: la señora Abigail Abad (primer caso: i=1) mide 167 cm. Por tanto, e_1 vale

$$y_1 = 167;\ y_1 = 165;\ \rightarrow\ e_1 = y_1 - y_1 = 167 - 165 = 2$$

Su idiosincrasia vale 2 cm, que generará un residuo o error en la predicción de 2 cm

Para los hombres, la altura esperada es 175 cm. Sin embargo, el señor Abraham Abadesa (caso i=2) mide 174 cm, por lo que su particularidad ε vale -1:

$$y_2 = 174;\ y_2 = 175;\ \rightarrow\ e_2 = y_2 - y_2 = 174 - 175 = -1$$

Nota: Poner a mujeres en 0 es arbitrario. Tan sólo es más cómodo hablar en positivo: "los hombres miden 10 cm más".

Recuerde

Elija la categoría de referencia ("0") para facilitar la interpretación.

Como en RLS, la instrucción *lm* estima los coeficientes del modelo RLM.

Ejemplo R

```
> # Datos (4 hombres y 4 mujeres)
> y <- c(165,171,164,149,169,179,175,184)
> z <- factor(c('M','M','M','M','H','H','H','H'),levels=c('M','H'))

> # Descriptiva (tapply realiza el 'summary' de y estratificado por z)
> tapply(y,z,summary)
$M
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  149.0   160.2   164.5   162.2   166.5   171.0
$H
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  169.0   173.5   177.0   176.8   180.2   184.0

> # Modelo estimado
> lm(y~z)
Call:
lm(formula = y ~ z)
Coefficients:
(Intercept)           zH
      162.2         14.5
```

Ejercicio 2.2

Diga, para este ejemplo, qué significan 162.2 y 14.5.

El modelo lineal descansa en la premisa de linealidad: diferentes incrementos de una unidad en la variable predictora Z_k van seguidos del mismo incremento β_k en la variable respuesta Y. En el caso de una dicotomía, sólo hay un solo incremento entre las dos categorías y, por tanto, la premisa de linealidad no es necesaria.

Nota: Por 2 puntos siempre pasa una recta. No tiene mérito alinear 2 medias en una recta, lo que sí lo tendría sería alinear 3 o más.

Una dicotomía, como la representada por una variable *dummy*, puede interpretarse en cualquier escala de medida, incluso numérica.

Ejemplo 2.3 (cont. Ejemplo 3.2): Podemos mirar al indicador (*dummy*) de género como "número de cromosomas Y": '0' para las mujeres y "1" para los hombres. La diferencia de medias observada, 14.5 cms, puede interpretarse como el incremento de altura asociado a un incremento de 1 unidad en el número de cromosomas Y.

Recuerde

Un indicador (*dummy*) representa en números a una dicotomía.

Ejercicio 2.3

Las siguientes instrucciones cargan los datos *faraway* de Espondilitis Anquilosante y guardan en *mod.lm1* la RL de la flexión final (*faft*) en función del tratamiento (*grp*) en el brazo derecho.

```
> install.packages("faraway")
> library(faraway)
> mod.lm1 <- lm(faft~grp,data=subset(hips,side=='right'))
```

Ejecute estas instrucicones y obtenga el *summary* de mod.lm1, luego: a) Interprete a nivel predictivo los 2 coeficientes del modelo; b) discuta su significación estadística; c) interprete la capacidad de anticipación del modelo; d) obtenga el IC95% del coeficiente de GRP; y e) asumiendo que se trata de un ECA bien diseñado y ejecutado, sin riesgos de sesgos, ¿qué coeficiente podría interpretar causalmente? Interprételo.

Ejercicio 2.4

Diga cuáles son ciertas y, en caso contrario, exprese correctamente diciendo las razones.

a) El mejor nombre para las variables a la derecha del "=" es *independientes*.

b) Strobe sugiere llamar intervenciones y confusoras a las variables a la derecha del símbolo "=".

c) Como en el modelo RLS, las mayúsculas representan a los valores; y las minúsculas, las variables.

d) Como en el modelo RLS, las letras griegas representan a los valores estimados en las muestras; y las latinas, los parámetros desconocidos de la población.

e) RLM usa 2 subíndices: k para los casos; y i para las variables.

f) Como en RLS, el término 'e' representa la particularidad de cada caso, es decir, aquello que puede ser modelado de forma común con los demás casos. Representa aquello predecible por el modelo estudiado y suele llamarse residuo o incluso error.

g) En una RL con un indicador (*dummy*) de valores 0, 1; la constante proporciona la diferencia de medias ; y la pendiente a la media del grupo "1".

h) Como en RLS, en una *dummy* de valores 0, 1; el coeficiente que estudia la relación entre la predictora Z y la respuesta Y es la constante (*intercept*).

i) Si hago el promedio de los valores 0, 1 de una *dummy*, obtengo la suma de '1' dividida por el total de casos, es decir, el promedio de "unos", que no es nada más que la proporción de casos que tienen el valor 1.

2.1 Más de 2 categorías *

La variable categórica puede tener más de 2 modalidades. En este caso, se tendrá un parámetro adicional por cada modalidad extra. El motivo, es que se crean tantas variables "dummies" con 2 categorías (con valor 0 si no se pertenece a determinada modalidad y 1 si se pertenece) como modalidades tenga la variable original menos 1 (la de referencia).

Ejemplo 2.4: En una variable que representa la edad categorizada con las modalidades *joven* (referencia), *adulto y anciano*, al hacer la regresión se transformará automáticamente en dos variables *dummies* representando las categorías *adulto* y *anciano*.

Variable original	Dummy Adulto	Dummy Anciano
joven	0	0
adulto	1	0
anciano	0	1

Por tanto, en estos casos, para una variable con k categorías, el modelo ajustará (k-1) pendientes que se interpretarán cada una de ellos como el "efecto" en la respuesta de esa modalidad concreta respecto a la de referencia.

Ejemplo R : flexión final según 3 niveles de flexión inicial

```
> # Cargamos la libreria y los datos
> library(faraway)
> data(hips)

> # Nueva variable con 3 categorias → (87,112];(112,122];(122,139]
> hips$fbef.cat <- cut(hips$fbef,br=c(87,112,122,139))
> lm(faft~fbef.cat,subset(hips,side=="right"))
[…]
Coefficients:
      (Intercept)  fbef.cat(112,122]  fbef.cat(122,139]
           107.80              14.14              22.65
```

La flexión posterior (*faft*) en el brazo derecho es 14.14 grados superior en aquellos con una flexión inicial (*fbef*) moderada (entre 113 y 122) y 22.65 grados superior en aquellos con flexión inicial elevada (entre 123 y 139) respecto a aquellos con una flexión inicial pobre (entre 88 y 112). El segundo incremento es un 60% superior al primero.

Ejercicio 2.5

Repita el ejemplo anterior con los datos del brazo derecho, pero para el nivel de rotación (*raft* y *rbef*). Recodifique la variable de rotación inicial *rbef* en 4 categorías con puntos de corte en los percentiles 25%, 50% y 75%. [Pista: use la instrucción quantile]. Interprete. Discuta la proporcionalidad de los incrementos.

2.2 Codificaciones alternativas de las variables cualitativas *

Como se ha explicado, R usa por defecto la codificación de 0's y 1's. No obstante, ésta no es la única codificación posible. Otra alternativa habitual es usar -1's y +1's. Usar una u otra codificación cambia la interpretación de los coeficientes.

R permite cambiar la codificación que se emplea en los factores.

Ejemplo R

```
> # Cargamos la libreria y los datos
> library(faraway)
> data(hips)

> # Codifición con 0's y 1's ('contr.treatment')
> options(contrasts=c('contr.treatment','contr.poly'))
> lm(raft~grp,subset(hips, side == "right"))
[…]
Coefficients:
(Intercept)     grptreat
      29.50         3.13

> # Codifición con -1's y 1's ('contr.sum')
> options(contrasts=c('contr.sum','contr.poly'))
> lm(raft~grp,subset(hips, side == "right"))
[…]
Coefficients:
(Intercept)         grp1
     31.065       -1.565

> # Volvemos a la codificación inicial
> options(contrasts=c('contr.treatment','contr.poly'))
```

Note que con la codificación -1 y 1, el coeficiente del tratamiento (*grp*) es la mitad del efecto del tratamiento.

Recuerde

Usar una codificación alternativa para las variables categóricas cambia el valor de los coeficientes de la recta.

Nota: En este caso, no cambiaría su significación estadística ni su R^2, pero esta propiedad no aplica a los modelos multivariantes.

Utilice la codificación por defecto e interprete los coeficientes de la manera explicada. Si tiene dudas utilice la instrucción *predict* de R para interpretar.

3 Modelo con 2 factores

Con 2 factores, el modelo es:

$$Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \varepsilon$$

Ejercicio 3.1

Convierta las 2 variables *grp* y *side* en una única Z de 4 valores mediante:

```
> z <- with(hips,factor(paste(grp,side,sep=""),
                        labels=c("Ref_I","Ref_D","T_I","Trt_D")))
```

Haga `plot(faft~z,data=hips)` e interprete.

Ejemplo 3.1 (cont. Ejemplo 2.2): La Figura 3.1 contiene las dos rectas de regresión que se ajustarían para cada brazo del paciente. Si asumimos que la calidad del estudio permite interpretar causalmente los resultados se puede ver que el efecto del tratamiento es el mismo en ambos brazos. En ambos brazos el efecto aproximado de cambiar de C a T es de 10 grados (en el brazo izquierdo de 112 a 122 y en el brazo derecho de 110 a 120).

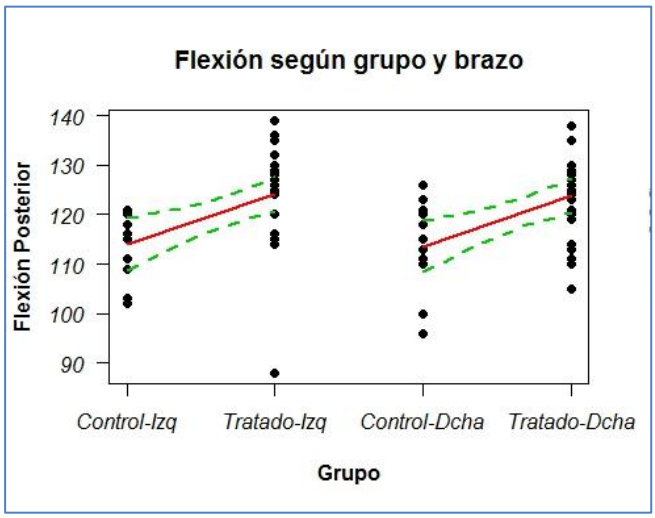


Figura 3.1. Rectas de regresión para ambos brazos.

Al añadir variables aumentan las formas de preguntar a los datos. Cada una tendrá sus matices lógicos. Estimar un efecto propio de cada centro tiene una utilidad limitada. Y se basa en menos casos. Pero una pregunta más ambiciosa sería estimar un efecto único común para los 2 centros — siempre que tenga sentido creer que el efecto del tratamiento es el mismo en ambos.

Ejemplo 3.2: En un EC sobre el efecto de un consejo dietético-higiénico profundo (T) frente al convencional (C), se han obtenido mediciones en 2 centros de atención primaria (A y B). La respuesta es PAD tras 8 semanas:

Ejemplo R

```
#Lectura de datos con R que estan accesibles via web:
>w<-'http://www-eio.upc.edu/teaching/best/datos-ejemplos/PAD.txt'
> datos<-read.table(url(w),header=TRUE)

#Descriptiva por centro y grupo:
> with(datos,by(PAD1,list(Tratamiento,Centro),summary))
> boxplot(PAD1~Tratamiento+Centro,datos)

#Ajuste del modelo G:
> modG <- lm(PAD1 ~ Tratamiento + Centro,data=datos)
> summary(modG)
Call:
lm(formula = PAD1 ~ Tratamiento + Centro, data = datos)
Residuals:
    Min      1Q  Median      3Q     Max
-16.600  -4.525  -0.100   5.300  12.200
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    84.500      1.936  43.652  < 2e-16 ***
TratamientoT  -15.800      2.235  -7.069  2.3e-08 ***
CentroB         6.100      2.235   2.729  0.00966 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 7.068 on 37 degrees of freedom
Multiple R-squared: 0.6081,     Adjusted R-squared:  0.5869
F-statistic: 28.71 on 2 and 37 DF,  p-value: 2.976e-08
```

A partir de la significación de la variable *TratamientoT* se puede afirmar que el tratamiento T tiene efecto en ambos centros. La variable *CentroB* también es significativa, esto quiere decir que existen diferencias entre el centro A (incluído en el *Intercept*) y el Centro B.

Ejercicio 3.2

(1) Recupere y compare los valores del residuo típico S y del coeficiente de determinación R^2 en los 3 modelos G, A, y B anteriores; ¿coinciden sus cambios?

(2) Discuta si estas 3 estimaciones de los residuos típicos significan lo mismo;

(3) ¿Bajo qué condiciones estos residuos representarían las 'particularidades' de los casos?

Para saber si el Tratamiento T tiene efecto en los centros por separado hay que crear un primer modelo en el que sólo se seleccionen los datos del Centro A y un segundo modelo en el que sólo se seleccionen los datos del Centro B

```
#Ajuste del modelo en el centro A; mediante  la función subset
seleccionamos los datos del centro A
> modA <- lm(PAD1 ~ Tratamiento ,data=subset(datos,Centro=='A'))
> summary(modA)
Call:
lm(formula = PAD1 ~ Tratamiento, data = subset(datos, Centro ==
    "A"))
Residuals:
   Min     1Q Median     3Q    Max
-12.00  -3.45  -0.10   5.00  10.80
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    84.200      1.828  46.057  < 2e-16 ***
TratamientoT  -15.200      2.585  -5.879 1.44e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 5.781 on 18 degrees of freedom
Multiple R-squared:  0.6576,    Adjusted R-squared:  0.6385
F-statistic: 34.56 on 1 and 18 DF,  p-value: 1.445e-05

#Ajuste del modelo en el centro B; mediante  la función subset
seleccionamos los datos del centro B
> modB <- lm(PAD1 ~ Tratamiento ,data=subset(datos,Centro=='B'))
> summary(modB)
Call:
lm(formula = PAD1 ~ Tratamiento, data = subset(datos, Centro ==
    "B"))
Residuals:
   Min     1Q Median     3Q    Max
-16.90  -5.05  -0.20   6.35  12.50
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    90.900      2.628  34.586  < 2e-16 ***
TratamientoT  -16.400      3.717  -4.412 0.000336 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 8.311 on 18 degrees of freedom
Multiple R-squared:  0.5196,    Adjusted R-squared:  0.4929
F-statistic: 19.47 on 1 and 18 DF,  p-value: 0.0003362
```

De nuevo, para saber si el tratamiento T tiene efecto en el Centro A hay que mirar la significación de las variables. La variable *TratamientoT* es significativa, por lo tanto el Tratamiento T tiene efecto en el centro A; lo mismo sucede con la significación de la variable *TratamientoT* en los datos del Centro B.

Ejercicio 3.3

(1) Recupere y compare los valores del residuo típico S y del coeficiente de determinación R^2 en los 3 modelos G, A, y B anteriores; ¿coinciden sus cambios?

(2) Discuta si estas 3 estimaciones de los residuos típicos significan lo mismo;

(3) ¿Bajo qué condiciones estos residuos representarían las 'particularidades' de los casos?

Recuerde

Estimar un único coeficiente ("efecto", si procede) requiere asumir que su valor es común en todos los casos.

Ejercicio 3.4

Con los datos del Ejemplo 3.2 obtenga el $IC_{95\%}$ del efecto, sin considerar el centro,e interprete.

Pero el centro podría ser también una fuente de variabilidad: podría ser que los aparatos fueran de distinta marca o no estuvieran igual calibrados; o los pacientes podrían tener distintas condiciones, quizás de edad o de hábitos saludables. Por si fuera así, podríamos querer descontar del residuo típico todo aquello que pudiera ser explicado por el centro. En ese caso, pondremos ambas variables en el mismo modelo.

Ejercicio 3.5

De nuevo, con los datos del Ejemplo 3.2 obtenga el $IC_{95\%}$ del efecto ajustando por centro e interprete.

Ejemplo 3.3 (cont. Ejemplo 3.2): Los residuos típicos proporcionados por R en los modelos que incluyen: (A) sólo el tratamiento; y (B) centro y tratamiento; son:

(A)

```
lm(formula = PAD ~ Tratamiento)
Residual standard error: 7.645 on 38 degrees of freedom
```

(B)

```
lm(formula = PAD ~ Tratamiento + Centro)
Residual standard error: 7.068 on 37 degrees of freedom
```

Aunque las 2 estimaciones son parecidas, el modelo que incluye ambas variables proporciona un valor menor para las 'particularidades': la oscilación alrededor del valor predicho para ese centro y ese tratamiento es de 7.068. Tiene 37 grados de libertad ya que se dispone de la información de 40 casos y se han estimado 3 parámetros.

Ejercicio 3.6

Con los datos del Ejemplo 3.2: En un EC sobre el efecto de un consejo dietético-higiénico profundo (T) frente al convencional (C), se han obtenido mediciones en 2 centros de atención primaria (A y B).compare los coeficientes del Centro y del Tratamiento obtenidos en los modelos que incluyen sólo una predictora y el modelo que incluye ambas.

Recuerde

Si en el modelo múltiple se añade una variable completamente independiente de la previa, la estimación puntual no cambia.

Nota técnica: La estimación del residuo típico es exactamente:

$$S = \sqrt{\frac{\sum e_i^2}{n-p-1}} \rightarrow \begin{matrix} e_i: residuo\ i-ésimo \\ n: número\ de\ observaciones \\ p: número\ de\ variables\ predictoras \end{matrix}$$

Al añadir un coeficiente en el modelo, el denominador desciende en 1 punto. El grado de descenso del numerador dependerá de la mejora predictiva. En el peor caso quedaría igual, por lo que al disminuir el denominador en un punto el residuo típico podría aumentar.

Recuerde

Cuando baja el residuo típico disminuye la oscilación de la estimación.

Ejercicio 3.7

Con los datos del Ejemplo 3.2: En un EC sobre el efecto de un consejo dietético-higiénico profundo (T) frente al convencional (C), se han obtenido mediciones en 2 centros de atención primaria (A y B).compare los errores típicos de los coeficientes de las variables centro y tratamiento. ¿Qué sucede? ¿Por qué cree que sucede? [Nota: el error típico de la pendiente coincide, en este caso, con el de la diferencia de 2 medias (comprobar y poner fórmula), usando como S, el valor del residuo típico.]

Ejercicio 3.8

Con los datos del Ejemplo 3.2: En un EC sobre el efecto de un consejo dietético-higiénico profundo (T) frente al convencional (C), se han obtenido mediciones en 2 centros de atención primaria (A y B).calcule a mano los valores predichos para cada tratamiento y centro de acuerdo con el modelo que incluye ambas variables. Compruebe que las siguientes instrucciones de R le proporcionan el mismo resultado:

```
> #Cree las 4 combinaciones para hacer la predicción.
> data.predict <- data.frame(Tratamiento=c("C","C","T","T"),
                             Centro=c("A","B","A","B"))
> predict(mod,data.predict)
```

Ejemplo 3.4: Al comparar los valores de los coeficientes de determinación de los modelos del Ejercicio 3.5 observe como cuando la variabilidad residual disminuye aumenta la variabilidad explicada (R-squared)

	Multiple R-squared	**Adjusted R-squared**	**Residual standard error**
mod (PAD ~ Tratamiento + Centro)	0.6081	0.5869	7.068
mod1 (PAD ~ Tratamiento)	0.5292	0.5168	7.645
mod2 (PAD ~ Centro)	0.07888	0.05464	10.69

En el modelo estudiado, los coeficientes correspondientes a las variables estudiadas se suman: "*son aditivos*". Ello es así porque en este modelo se ha considerado que el efecto de la intervención es el mismo en ambos centros. El próximo apartado relaja esta premisa.

Ejercicio 3.9

Diga cuáles son ciertas y, en caso contrario, exprese correctamente diciendo las razones.

a) Cabe esperar que al incluir más variables en el modelo, mejore la predicción y baje la estimación del residuo típico.

b) Cuando dos variables predictoras introducidas en el modelo son independientes entre sí, la estimación puntual del coeficiente de una es la misma cuando está la otra que cuando no está.

c) Cuando baja el residuo típico sube la precisión de las estimaciones de los parámetros.

d) Cuanto más baja el residuo típico más sube R^2.

3.1 Modelo con interacción entre 2 factores

Ejemplo 3.5: En un EC sobre el efecto de un fármaco (T) frente a un grupo Control (C) en la PAD a las 8 semanas, se han obtenido datos de 2 centros de atención primaria (A y B). La Figura 3.2 muestra los resultados hipotéticos de 4 posibles estudios con sus respectivas pendientes del efecto del tratamiento (T frente a C) en todas las situaciones.

Las líneas rojas enlazan las medianas de los box-plot: en las situaciones a y b, el descenso de la PAD es el mismo en ambos centros.

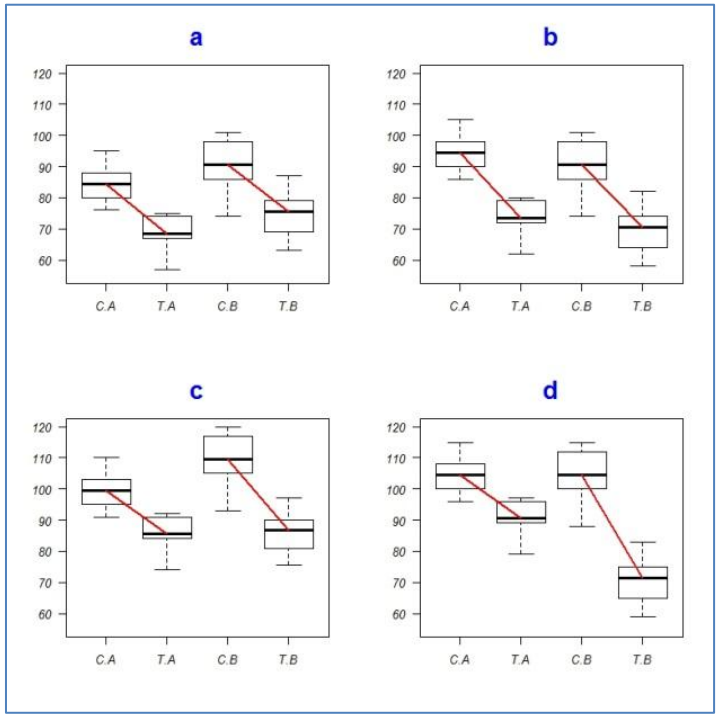


Figura 3.2

En la b es algo mayor en el segundo centro. Pero en el caso c y d la diferencia entre ambos efectos es mucho mayor. Este es un ejemplo de interacción, ya que el efecto del tratamiento es diferente según el centro: es decir en los casos c y d, hay interacción entre las variables tratamiento y centro.

La instrucción de R, *interaction.plot* proporciona un análisis gráfico directo de la interacción. La Figura 3.3 los muestra para los casos b y d anterior.

Ejemplo R

```
> # Lectura de datos
> w <- 'http://www-eio.upc.edu/teaching/best/datos-ejemplos/PAD.txt'
> datos <- read.table(url(w),header=TRUE)

> # Interaction.plot con las variables en este orden: X, Z, Y
> par(mfrow=c(1,2))
> with(datos,interaction.plot(Tratamiento,Centro,PAD2,
                              main="Datos de PAD2",ylim=c(70,110)))
> with(datos,interaction.plot(Tratamiento,Centro,PAD4,
                              main="Datos de PAD4",ylim=c(70,110)))
```

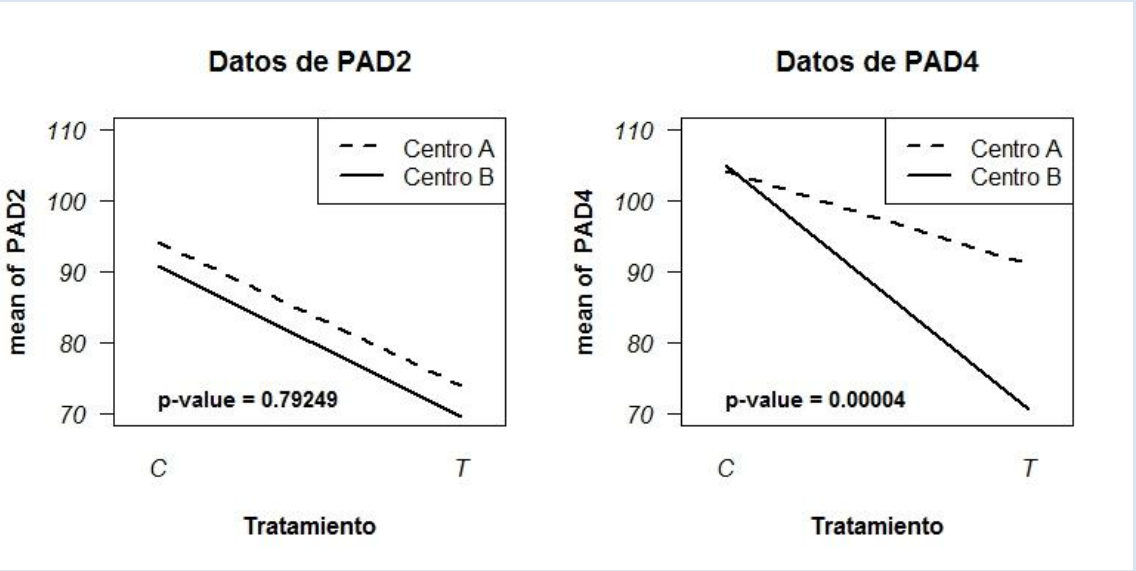


Figura 3.3

Ejercicio 3.10

Interprete los gráficos de interacción anteriores.

En términos poblacionales, las medias de los 4 grupos del Ejemplo 3.2: En un EC sobre el efecto de un consejo dietético-higiénico profundo (T) frente al convencional (C), se han obtenido mediciones en 2 centros de atención primaria (A y B).serán:

	T	C	
Centro A	$\mu_{T,A}$	$\mu_{C,A}$	μ_A
Centro B	$\mu_{T,B}$	$\mu_{C,B}$	μ_B
	μ_T	μ_C	

Así, las siguientes diferencias se corresponden con

$\mu_{T,A} - \mu_{C,A}$ → "efecto" de T respecto a C en el centro A.

$\mu_{T,B} - \mu_{C,B}$ → "efecto" de T respecto a C en el centro B.

$\mu_T - \mu_C$ → "efecto" de T respecto a C, en general, sin condicionar.

Las herramientas ($IC_{95\%,}$ valor de P) de la inferencia estadística permiten el salto a la población. Para ello ajustaremos el mismo modelo anterior añadiendo un término que indique la interacción.

Ejercicio 3.11

Lea con R los datos del último caso anterior (d) que estan accesibles via web, haga la descriptiva de los 2 grupos de tratamiento; de los 2 centros; de los 4 grupos; rellene la tabla inferior con todas las medias; estime puntualmente el efecto global y en cada centro con R; y explique a qué diferencia de medias corresponde en la tabla. Finalmente, discuta si tiene sentido estimar el efecto global.

```
> w='http://www-eio.upc.edu/teaching/best/datos-ejemplos/PAD.txt'
> datos<-read.table(url(w),header=TRUE)
> with(datos,mean(PAD4))
> with(datos,by(PAD4,list(Tratamiento),summary))
> with(datos,by(PAD4,list(Centro),summary))
> with(datos,by(PAD4,list(Tratamiento,Centro),summary))
> boxplot(PAD4~Tratamiento+Centro,datos)
```

	C	T	Todos
A			
B			
Todos			

El ejercicio anterior muestra que, en presencia de interacción, mirar un efecto global —que promedia efectos que son diferentes entre sí— tiene una utilidad muy limitada.

Recuerde

En presencia de interacción, un efecto global tiene poco sentido.

Ejercicio 3.12

Obtenga el modelo con interacción para el caso (d), compruebe si tiene el mismo valor anterior e intente deducir qué estima cada coeficiente en este modelo; es decir, a qué diferencia entre medias corresponde.

```
> mod.interaccion <- lm(PAD4 ~ Tratamiento * Centro,data=datos)
> summary(mod.interaccion)
```

La interacción puede ser definida mediante la diferencia entre los efectos de la intervención en ambos centros. A partir de:

$\mu_{T,A} - \mu_{C,A}$ → "efecto" de T respecto a C en el centro A.

$\mu_{T,B} - \mu_{C,B}$ → "efecto" de T respecto a C en el centro B.

$(\mu_{T,A} - \mu_{C,A}) - (\mu_{T,B} - \mu_{C,B})$ → "interacción del centro en el efecto" de T respecto a C

Recuerde

En caso de interacción, reporte los efectos en cada grupo por separado.

Ejercicio 3.13

Compare los errores típicos de estimación de los efectos de la intervención de los modelos del ejercicio 3.12. Observe también el error típico del término de interacción.

La estimación de un efecto global junta los casos de los grupos y tiene un menor error típico de la estimación. Así, si puede asumir que el efecto es el mismo, la estimación es más precisa. Además, disponer de una única medida del efecto, sin necesidad de matizar su valor según los grupos en comparación, hace la vida más fácil.

Recuerde

Un efecto homogéneo en los grupos es más preciso y fácil de comunicar.

La estimación de la interacción tiene un mayor error típico: dispone de más información para estimar los efectos de la intervención que para comprobar si son estables a lo largo de los grupos.

Nota: Ello es debido a que el efecto es la diferencia de 2 medias, pero la interacción le da una vuelta más: es la diferencia de los efectos en los subgrupos; es decir, la diferencia de 2 diferencias. Al comparar 4 estimaciones, cada una con su error muestral tiene mayor imprecisión. Además, cada estimación se basa en un subgrupo de menor tamaño, con mayor oscilación muestral.

Recuerde

Hay menos precisión para estudiar la interacción que los efectos.

Este mayor error de estimación conduce a mayores $IC_{95\%}$. Pero, quizás más relevante, dificulta encontrar diferencias significativas: desciende la potencia del contraste que pone a prueba la interacción.

Recuerde

La prueba de la interacción tiene menos potencia.

Ejercicio 3.14

Compruebe que se pueden reproducir las medias de cada grupo combinanado los coeficientes del modelo con interracción. Utilice las siguiente comanda para obtener la tabla de medias:

```
> install.packages("MASS")
> library(MASS)
> data(birthwt)
> birthwt
> summary(birthwt)
> tapply(birthwt$bwt, list(birthwt$smoke, birthwt$low), mean)
```

La interacción puede ser una hipótesis que se desea estudiar y poner a prueba.

Ejemplo 3.6: la combinación de 2 antibióticos tiene un efecto sinérgico.

O también, casi al revés, la homogeneidad del efecto puede ser una premisa muy conveniente para estudiar el efecto de una intervención en una población más amplia.

Ejemplo 3.7: Los ensayos clínicos hacen el cálculo del tamaño para estimar un único efecto. Implícitamente están asumiendo que, dentro de las condiciones del estudio, los criterios de elegibilidad definen una población en la que el efecto de la intervención es constante.

Ejercicio 3.15

Diga cuáles son ciertas y, en caso contrario, exprese correctamente diciendo las razones.

a) La interacción es la diferencia del efecto de la intervención en los grupos estudiados.

b) En presencia de interacción, los coeficientes del modelo incluidos en la misma son más fáciles de interpretar.

c) La prueba de hipótesis de la interacción tiene más potencia estadística que la del efecto de la intervención.

d) La interacción puede ser vista como una hipótesis a ser contrastada, pero también como una premisa necesaria para la hipótesis principal.

4 Modelo aditivo para una numérica y una categórica.

Ejemplo 4.1: Recupere el ejemplo del peso y la altura de los adultos varones sanos (Ejemplo 2.1), pero tenga en cuenta también, ahora, a las mujeres. Una simple ecuación podría ser:

$Peso_i = -100Kg + 1Kg/cm \cdot altura_i - 5Kg$ si mujer $+ \varepsilon_i$

Así, la predicción para una mujer de 170cms es 65Kg. La representación gráfica es (Figura 4.1):

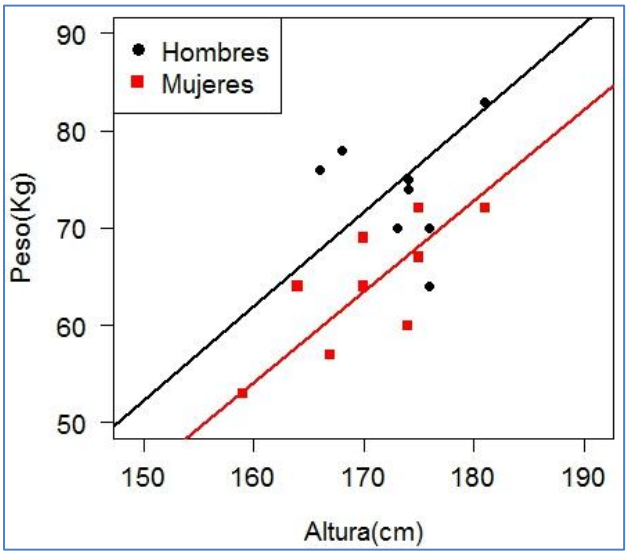


Figura 4.1

Si además cuantifica en $\sigma\varepsilon = 5Kg$ a la dispersión de la particularidad εi, el modelo estará completamente especificado.

Este modelo "aditivo" que simplemente suma los coeficientes tiene una interesante implicación: la suma de un coeficiente es la misma para cualquier valor de la otra variable.

Recuerde

Al reportar el modelo debe informar sobre la dispersión de ε_i.

Ejemplo 4.1 (cont): sea cual sea el valor de la altura, siempre resta 5Kg a todas las mujeres.

Así, aplicar el coeficiente de una variable arroja siempre el mismo resultado, "independientemente" del valor de la otra variable.

Ejercicio 4.1

Continuando con los datos *hips* del paquete *faraway*, obtenga el summary del siguiente modelo que compara el efecto del tratamiento (grp) en la respuesta (faft) teniendo en cuenta el nivel inicial (fbef) y conteste, teniendo en cuenta los $IC_{95\%}$: (1) Por cada grado inicial de movilidad, ¿cuánto mayor es la movilidad final? (2) ¿Qué hubiera significado que el coeficiente para fbef hubiera sido 1? (3) ¿cuántos grados de movilidad aumenta la intervención? (4) Este efecto del tratamiento, ¿varía según la movilidad inicial? (5) ¿Cuál es la capacidad predictiva de este modelo?

```
> mod.lm<-lm(faft~grp+fbef,data=subset(hips,side=='right'))
```

Recuerde

En el modelo aditivo, un coeficiente es "*independiente*" de las otras variables.

La instrucción *lm(...)* proporciona el ajuste del modelo (al igual que con la regresión lineal simple).

Ejercicio 4.2

A partir de las instrucciones siguientes, obtenga e interprete el modelo para predecir el peso en función de altura y género.

```
> # Lectura de datos
> w<-'http://www-eio.upc.edu/teaching/best/datos-ejemplos/peso.txt'
> datos<-read.table(url(w),header=TRUE)

> # Ajuste del modelo
> mod.lm1 <- lm(peso~altura+genero,datos)
> summary(mod.lm1)
> # IC para los coeficientes
> confint(mod.lm1)
```

Recuerde

Interprete los coeficientes con sus $IC_{95\%}$.

Ejercicio 4.3

El recuadro muestra la salida de R para la regresión de la Presión Arterial Sistólica (PAS) en función de la edad y desglosada por género (datos inventados). Interprete los coeficientes y obtenga su intervalo de confianza del 95%.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 103.0133     5.7636  17.873  < 2e-16 ***
edad          0.4194     0.1094   3.835 0.000264 ***
generoMujer  -8.9092     2.6729  -3.333 0.001351 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.65 on 73 degrees of freedom
Multiple R-squared: 0.2576,     Adjusted R-squared: 0.2373
F-statistic: 12.67 on 2 and 73 DF,  p-value: 1.897e-05
```

4.1 Modelo con interacción entre numérica y categórica*

Hasta ahora la premisa ha sido que la dicotomía no altera la relación entre las otras dos variables.

Ejemplo 4.2: La pendiente (el incremento de peso por cada cm de altura) era el mismo en hombres y mujeres. Suponga ahora que no es así y que Vd. conoce el auténtico modelo, que es:

Hombres $\rightarrow Peso_i = -130 + 1.2 \cdot altura_i + e_i$

Mujeres $\rightarrow Peso_i = -100 + 1.0 \cdot altura_i + e_i$

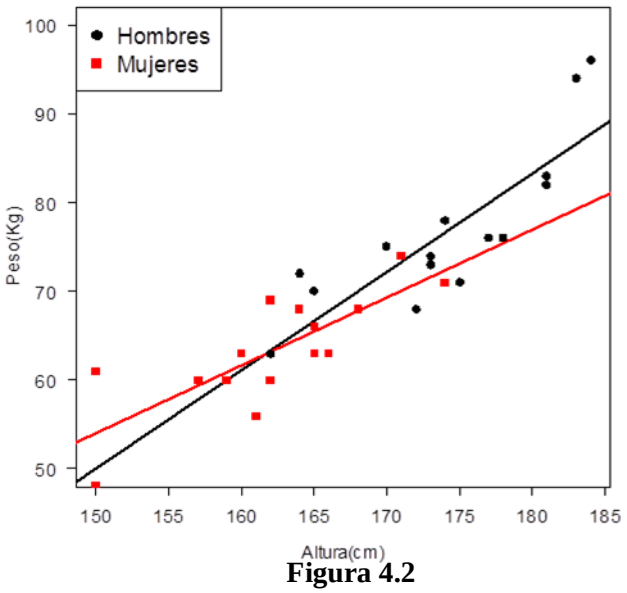


Figura 4.2

La obtención de este modelo con R puede llevarse a cabo de 2 maneras. La primera trabaja con todas las variables indicando con un "*" que el término de interacción estará incluido:

Ejemplo R

Los siguientes datos son un ejemplo imaginario de peso y altura.

```
# datos
> w<-'http://www-eio.upc.edu/teaching/best/datos-ejemplos/peso2.txt'
> datos<-read.table(url(w),header=TRUE)
> mod.lm2 <- lm(peso~altura*genero,datos)
> summary(mod.lm2)

[…]
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        -93.3929    45.9291  -2.033   0.0589 .
altura               0.9705     0.2597   3.737   0.0018 **
generoMujer         -2.1492    59.4328  -0.036   0.9716
altura:generoMujer  -0.0356     0.3427  -0.104   0.9186
[…]
```

Nota: Fíjese en la equivalencia de este modelo con los dos posteriores:

$Peso_i = \beta_0 + \beta_1 \cdot altura + \beta_2 \cdot género + \beta_{12} \cdot altura{:}genero + \varepsilon_i$

Si género es igual a 0 (Hombre), entonces:

$Peso_i = \beta_0 + \beta_1 \cdot altura \quad \rightarrow \quad \beta_0 = \beta_{0H} = -116.0 \;\; ; \;\; \beta_1 = \beta_{1H} = 1.1$

Si género es igual a 1 (Mujer), entonces:

$Peso_i = \beta_0 + \beta_1 \cdot altura + \beta_2 + \beta_{12} \cdot altura = (\beta_0 + \beta_2) + (\beta_1 + \beta_{12}) \cdot altura \rightarrow \beta_0 + \beta_2 = \beta_{0M} = -60.7$

$\beta_1 + \beta_{12} = \beta_{1M} = 0.76$

Una posibilidad es ajustar un modelo para cada género:

Hombres: $Peso_i = \beta_{0H} + \beta_{1H} \cdot altura + \varepsilon_i$

Mujeres: $Peso_i = \beta_{0M} + \beta_{1M} \cdot altura + \varepsilon_i$

El modelo para los hombres se podría obtener ajustando lm sólo para los hombres

Ejemplo R

```
> mod.lmH <- lm(pesoH~alturaH)
> summary(mod.lmH)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -116.0495    34.3974  -3.374  0.00499 **
alturaH        1.1071     0.1974   5.609  8.5e-05 ***
> confint(mod.lmH)
                  2.5 %     97.5 %
```

```
(Intercept) -190.3604799 -41.738538
alturaH        0.6806569   1.533541
```

El IC95% para β_{0H} es [-190.4 a -41.7] e incluye el verdadero valor (-130) del modelo.

El IC95% para β_{1H} es [0.68 a 1.53] e incluye la pendiente real (1.2) del modelo.

El modelo para las mujeres sería:

Ejemplo R

```
> mod.lmM <- lm(pesoM~alturaM)
> summary(mod.lmM)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -60.6949    26.6159  -2.280 0.040089 *
alturaM       0.7643     0.1639   4.664 0.000443 ***

> confint(mod.lmM)
                  2.5 %    97.5 %
(Intercept) -118.1950383 -3.194763
alturaM        0.4102724  1.118424
```

El IC95% para β_{0M} es [-118.2 a -3.2] e incluye el verdadero valor (-100) del modelo.

El IC95% para β_{1M} es [0.41 a 1.11] e incluye la pendiente real (1) del modelo.

Ejercicio 4.4

En el siguiente gráfico (**Figura 4.3**) se vuelve a representar la PAS en función de la edad y estratificada por género (datos inventados). En el fichero *PAS.txt* encontrará los datos que han generado este gráfico (en el conjunto de datos los hombres se codifican con un 1 y las mujeres con un 2). Analice con R y responda:

1) ¿Cuál es el $IC_{95\%}$ para la pendiente (incremento en la PAS por año de edad) en los hombres?

2) ¿Cuál es el $IC_{95\%}$ para la pendiente en las mujeres?

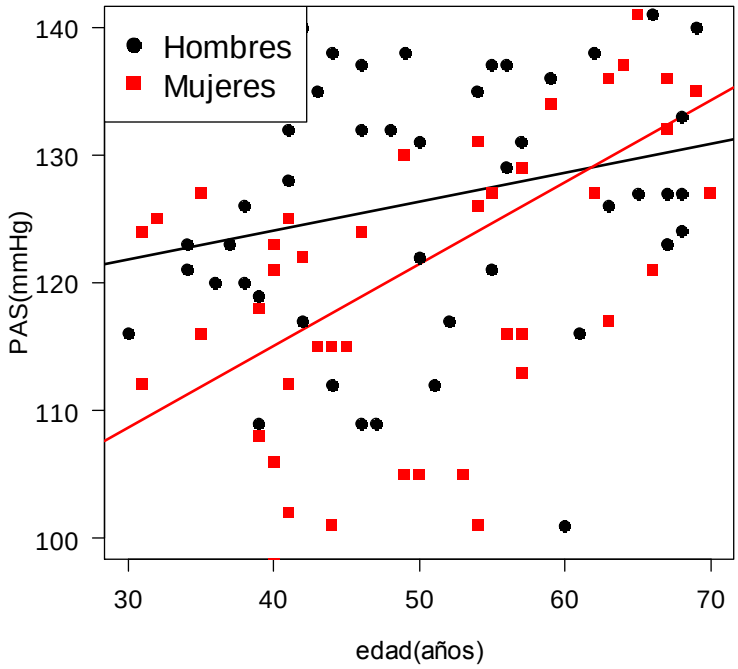


Figura 4.3

5 Utilidad en investigación de salud

Repitamos otra vez que, en los estudios observacionales, los modelos con múltiples predictoras tienen 2 grandes usos: (1) cuantificar la capacidad predictiva de una escala pronóstica y (2) explorar el papel etiológico de las variables predictivas.

Para el primer objetivo, el indicador más importante es el coeficiente de determinación R^2, que cuantifica, precisamente, la reducción en la incertidumbre de la respuesta Y a partir de las variables predictoras Z.

En cambio, para el segundo objetivo, el indicador más importante es la pendiente β, que permite especular sobre el futuro efecto en la respuesta Y que se obtendrá cuando se consiga cambiar la variable predictora X_i en una unidad mientras se dejan fijas todas las restantes.

Recuerde

La regresión múltiple permite cuantificar: (1) o bien la capacidad para anticipar la respuesta mediante el R^2; (2) o bien el hipotético efecto β causal en Y de cambiar X en una unidad.

Las guías TRIPOD y STROBE abordan el modelado aplicado al primer y al segundo objetivo, respectivamente.

5.1 La predicción puede ser simultánea o futura.

En un estudio diagnóstico puede desearse conocer hasta qué punto puede sustituirse a una variable actual o simultánea obtenida en un estudio transversal. En cambio, el pronóstico requiere un lapso de tiempo recogido en un diseño longitudinal.

Ejercicio 5.1

Busque en la red Equator la guía TRIPOD sobre modelos múltiples y diga si aplica a diagnóstico, a pronóstico, a ambos o a ninguno.

Ejemplo 5.1: La Tabla 5.1 muestra el modelo de regresión lineal múltiple para predecir la masa del ventrículo izquierdo[1].

TABLA 5. **Análisis de regresión múltiple entre el índice de masa del ventrículo izquierdo y los valores de interleucina-6**

Variable	Coeficiente	IC del 95%	p
Intercepto	94,51	76,2-112,78	< 0,001
IL-6 (log)	42,41	15,47-69,35	0,0031
Varón	36,41	14,75-58,07	0,0017
Fase III de la enfermedad de Chagas	92,62	66,40-118,84	< 0,001

Se presentan las variables remanentes después de la eliminación hacia atrás de las no significativas.
r^2 ajustada = 0,66; n = 64.

Tabla 5.1

Este modelo explica el 66% de la respuesta con las 3 variables predictoras.

Para ver qué implica R^2 en una situación concreta hay que recordar, 1º, que es una medida sobre errores cuadrados; y, 2º, que los intervalos tienen mayor amplitud para valores de las predictoras Z más alejados de sus medias.

Recuerde

Sólo si el modelo es independiente a la muestra, el estudio permite contrastar dicha hipótesis, confirmando o rechazando su capacidad predictiva.

Ejercicio 5.2

¿Por qué es necesario el lapso de tiempo para un estudio pronóstico pero no para uno diagnóstico.

Ejercicio 5.3

Según la guía TRIPOD, el modelo múltiple obtenido en un estudio ¿es una propuesta por confirmar o un resultado ya validado?

5.2 Coeficientes ajustados: Especulando sobre cómo cambiar el futuro

Otro atractivo de este modelo es que representa la relación entre la variación en una unidad para una variable predictora (mientras se dejan igual las restantes) con el incremento en la respuesta Y.

Ejemplo 5.2: ¿Cuál sería el cambio en la respuesta Y cuando la variable X_1 aumenta 1 unidad y todas las demás variables permanecen constantes? Nótese que el cambio asociado a la respuesta Y será, precisamente de β_1 unidades. En general, Y aumenta β_1 veces el incremento experimentado por la variable X_1.

Recuerde

Los coeficientes β_k representan el cambio esperado en la respuesta Y por un aumento unitario en X_k, mientras permanecen fijas todas las demás variables.

Estos modelos permiten fijar matemáticamente a las terceras variables Z, para así estimar la relación de la variable X habiendo suprimido las posibles interferencias de las terceras variables Z. Nótese que este "fijar" es conceptual: intenta emular qué se hubiera observado en un estudio que, por criterios de selección controlara a esas variables.

Ahora bien, si X está relacionada con Y a nivel fijo de Z, ¿podemos interpretar etiológicamente el coeficiente β_X? ¿Podemos decir que X es una causa de Y? ¿Podemos decir que β_X es el 'efecto' en Y cuando cambiamos X en una unidad y dejamos fijas el resto de variables? El próximo capítulo discute las premisas que hacen razonable esta "emulación".

5.3 Confusión entre pronóstico y etiología

En ocasiones, podría suceder que, inconscientemente, cambiemos entre los objetivos de pronóstico y de etiología. Por ejemplo, ¿cuál es el objetivo de establecer la relación entre la edad de los padres y el riesgo de síndrome de Down: intervenir o predecir? Si lo que se pretende es cambiar la edad de la madre o del padre para disminuir el riesgo, se trata de una intervención, pero si se desea

seleccionar aquellos embarazos con elevada probabilidad de síndrome de Down, entonces el objetivo es predecir. Para este objetivo, es más conveniente usar la edad de la madre, aunque sólo sea por discreción.

NOTA: Si el objetivo fuera intervenir, para poder establecer la necesaria relación causal, la mejor respuesta vendría de un diseño experimental en el que se estudiara la edad de cada uno de los padres dejando equilibrada la del otro. Por ejemplo, se aparearían tanto las madres de 20 como las de 40 años, por igual, con padres de 20 y de 40 años. Este diseño no sería ético, por supuesto, pero es que además, en nuestro entorno sociocultural, este estudio no tendría sentido práctico, ya que una madre o un padre no buscan pareja «independientemente» de su propia edad.

Ejercicio 5.4

Diga si es correcto o corrija en caso contrario.

a) Si definimos la solidez de un estudio como la menor necesidad de premisas adicionales, podemos decir que un ensayo clínico aleatorizado es más sólido que un estudio observacional.

b) Un estudio transversal permite estimar modelos múltiples pronósticos.

c) Un estudio longitudinal pretende estimar modelos múltiples diagnósticos.

d) Pronóstico, etiología e intervención hablan de variables en diferentes momentos del tiempo.

e) Hablar de pronóstico, etiología o intervención en un estudio transversal tiene pinta de ser un marrón horrible.

Soluciones a los ejercicios

1.1. Tiene k+1: el término independiente β_0 y k coeficientes para las k pendientes de las k predictoras: $\beta_1, \beta_2, \ldots, \beta_k$.

1.2. Y: Variable respuesta. β: Coeficientes de las variables predictoras. β_0: Intercept, constante o término independiente. β_1: Pendiente que indica el aumento de Y que acompaña a un aumento de Z_1 en una unidad. Z_1: Primera variable predictora. β_k: Coeficiente de la variable predictora k. Z_k: Variable predictora k. ε: término aleatorio. y_i: Valor de la variable respuesta para el caso iésimo. b_0: Estimación concreta de β_0. b_1: Estimación de β_1. z_i: Valor de Z del caso iésimo. b_k: Coeficiente estimado de la variable k. z_{ki}: Valor de la variable k en el caso iésimo. e_i: Término aleatorio (diferencia entre valor real y predicho) del caso iésimo.

1.3. En la declaración STROBE no aparece. Sí aparece en el ítem 7 de E&E, pero no aconseja utilizarlo, así como tampoco aconseja utilizar el término "variables explicativas". En su lugar sugiere exposiciones y confusoras. En TRIPOD E&E 'independent variable' aparece sólo 1 vez, por 5 'predictor, sólo en la 1ª página.

2.1. El coeficiente 113.75 representa la media estimada para el grupo de control y el coeficiente 10.23 representa la diferencia estimada de medias.

2.2. La media estimada de la altura en las mujeres es 162.2cms y los hombres son 14.5 cm más altos, en media.

2.3. `>summary(mod.lm1)`

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  113.583      2.591  43.840  < 2e-16 ***
grptreat      10.269      3.114   3.298  0.00216 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 8.975 on 37 degrees of freedom
Multiple R-squared:  0.2272,      Adjusted R-squared:  0.2063
```

a) La flexión final esperada en los controles es de 114º y se espera un incremento de 10º en los tratados; b) Ambos coeficientes son significativamente diferentes de 0; c) La capacidad predictiva del modelo es del 20% (conocer el grupo reduce un 20% la incertidumbre sobre la respuesta) d) use `confint(mod.lm1)` o, simplemente sume y reste 1.96 veces el error estándar (3.11) para obtener que la pendiente poblacional está entre 3.96 y 16.58 con una confianza del 95%; e) Ser tratado aumenta el ángulo de flexión posterior en el brazo derecho unos 10º, con una incertidumbre de este valor que oscila entre 4 y 17º)

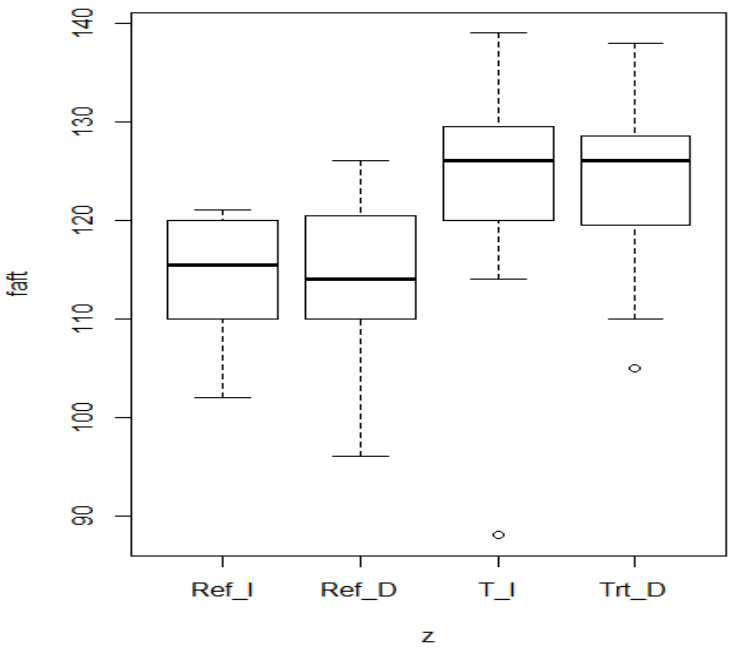
2.4. a) Falso, TRIPOD usa predictoras y STROBE exposiciones o confusoras. b) Falso, sí confusoras, pero no intervenciones. c) Falso, al revés, mayúsculas indica a toda la variable y minúsculas a su valor en un caso concreto. d) Falso, al revés. e) Falso, al revés, *k* es para las variables e *i* para los casos (número de individuo). f) Falso, representa aquello no modelado y no predecible por el modelo. g) Falso, la constante representa la media estimada del grupo 0 y la pendiente representa la diferencia de medias. h) Falsa, es la pendiente (que representa la diferencia de medias) la que resume la relación entre ambas. i) Cierta

```
> data(hips)
> # Cálculo de cuantiles
> quantile(hips$rbef,c(0,0.25,0.5,0.75,1))
   0%   25%   50%   75%  100%
 2.00 20.25 25.00 31.50 48.00
> # Categorización
> hips$rbef.cat <- cut(hips$rbef,br=c(2,20.25,25,31.5,48),
                       include.lowest = TRUE)
> # Modelo
> lm(raft~rbef.cat,subset(hips, side == "right"))
[…]
Coefficients:
```

```
(Intercept)  rbef.cat(20.2,25]  rbef.cat(25,31.5]  rbef.cat(31.5,48]
     24.250             5.167             9.194            14.450
```

La rotación posterior (raft) en el brazo derecho es 5.2 grados superior en aquellos con una rotación inicial (rbef) comprendida entre 20 y 25 grados; 9.2 grados superior en aquellos con una rotación basal entre 25 y 31.5; y 14.4 grados mayor para rotaciones previas superiores a 31.5 respecto a los pacientes con una rotación inicial inferior o igual a 25 grados. Los incrementos para cuartiles sucesivos son 5.2, 4.0 (9.2-5.2) y 5.2 (14.4-9.2), bastante similares, y es coherente que el incremento entre las clases centrales (4.0) sea menor ya que estas se encuentran más concentradas.

3.1. Mediante el gráfico se puede ver que los individuos que han tenido el tratamiento (*grp*) de referencia *control* tienen una flexión final inferior a los que han seguido el tratamiento *treat*. Además puede ver que tanto el tratamiento de referencia como el tratamiento *treat* parece que funcionan igual en ambos brazos ya que no hay diferencias significativas entre el brazo izquierdo y el brazo derecho en ninguno de los dos grupos.



3.2. (1) En G, S=7.068 , R^2=0.6081; en A, S=5.781 , R^2=0.658; y en B, S=8.311, R^2=0.520. Cuanto mayor es S, menor es R^2. El modelo que deja menos por predecir (S) y predice más (R^2) es B.

(2) Son modelos distintos que estiman S diferentes: significarían lo mismo si asumiéramos que en ambos centros los pacientes tienen la misma variabilidad; y en este caso, la mejor estimación sería la de G, que dispone de más casos (información) y sería más estable.

(3) Sólo asumiendo que hemos 'descontado' todo lo explicable, lo que quede sería propio del caso y no compartido con otros. Ello requiere que el modelo incluya (y haya descontado) todo lo que sea común; es decir, como veremos en el próximo capítulo, que el modelo sea completo en el sentido de incluir todas las variables que explican la variabilidad [Nótese lo exigente de esta premisa.].

3.3.

```
#Ajustamos el modelo solo con la variable Tratamiento, para que R entienda que no se quiere
distinguir por centro.
>mod1<-lm(PAD1 ~  Tratamiento,data=datos)
#Mediante la función confint se obtiene el IC95% para la variable Tratamiento
> confint(mod1)
                  2.5 %    97.5 %
(Intercept)   84.08954  91.01046
TratamientoT -20.69382 -10.90618
```

El $IC_{95\%}$ para el efecto del tratamiento es [-20.7 -10.9], esto quiere decir que el tratamiento T provoca una disminución de la PAD media de entre 20.7 y 10.9 mmHg., con una confianza del 95%, si no distinguimos por centro.

3.4. En este caso, se realiza el ajuste del modelo con las dos variables (Tratamiento y Centro) y se obtiene el $IC_{95\%}$ del mismo modo que en el ejercicio anterior:

```
> mod <- lm(PAD1 ~ Tratamiento + Centro,data=datos)
> confint(mod)
                  2.5 %    97.5 %
(Intercept)   80.577780  88.42222
TratamientoT -20.328989 -11.27101
CentroB        1.571011  10.62899
```

El IC95% para el efecto del tratamiento es [-20.33 -11.27], esto quiere decir que en el 95% de los casos el tratamiento T provoca una disminución de la PAD media de entre 20.33 y 11.27 mmHg, si distinguimos por centro.

Tenga en cuenta que al incluir una variable que reduce el residuo, baja la oscilación de la estimación; y ello disminuye el numerador del error típico de la estimación de la pendiente, por eso en este segundo ejercicio, si se compara con el anterior el IC es más estrecho.

3.5.
```
> mod <- lm(PAD ~ Tratamiento + Centro,data=datos)
> summary(mod)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    84.500      1.936  43.652  < 2e-16 ***
TratamientoT  -15.800      2.235  -7.069  2.3e-08 ***
CentroB         6.100      2.235   2.729  0.00966 **
--
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 7.068 on 37 degrees of freedom
Multiple R-squared:  0.6081,    Adjusted R-squared:  0.5869
---------------------------------------------------------------

> mod1 <- lm(PAD ~ Tratamiento,data=datos)
> summary(mod1)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    87.550      1.709  51.217  < 2e-16 ***
TratamientoT  -15.800      2.417  -6.536 1.05e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 7.645 on 38 degrees of freedom
Multiple R-squared:  0.5292,    Adjusted R-squared:  0.5168
---------------------------------------------------------------

> mod2 <- lm(PAD ~ Centro,data=datos)
> summary(mod2)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   76.600      2.391  32.036   <2e-16 ***
CentroB        6.100      3.381   1.804   0.0792 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 10.69 on 38 degrees of freedom
Multiple R-squared:  0.07888,    Adjusted R-squared:  0.05464
```

Son idénticos. Esto sucede en este caso porque la RLM garantiza que la estimación puntual de un coeficiente no cambia al añadir una predictora que sea independiente de la variable en estudio. En este caso, centro y tratamiento son independientes entre sí, ya que cada combinación tiene 10 casos.

3.6. Son iguales porque el residuo típico es el mismo para ambos coeficientes, así como las 'n' de las muestras.

Nota: el error típico de la pendiente coincide, en este caso, con el de la diferencia de 2 medias usando como S, el valor del residuo típico. $\frac{\sigma^2}{x_i - x^{-2}} = \frac{\sigma^2}{(1-0.5)^2 \cdot n} = \frac{\sigma^2}{(0.5)^2 \cdot n} = \frac{\sigma^2}{0.25 \cdot n} = \frac{\sigma^2}{\frac{n}{4}} = \frac{\sigma^2}{\frac{n}{2}} + \frac{\sigma^2}{\frac{n}{2}}$

teniendo en cuenta que será una suma de 0's y 1's, con tantos 0's como 1's

3.7. `> predict(mod,data.predict)`

```
    1    2    3    4
84.5 90.6 68.7 74.8
```

A mano, se pueden hallar, substituyendo las variables por 0's o 1's según corresponda:

$PAS_{CA} = 84.5 - 15.1 \cdot 0 + 6.1 \cdot 0 = 84.5$

$PAS_{CB} = 84.5 - 15.1 \cdot 0 + 6.1 \cdot 1 = 90.6$

$PAS_{TA} = 84.5 - 15.1 \cdot 1 + 6.1 \cdot 0 = 68.7$

$PAS_{TB} = 84.5 - 15.1 \cdot 1 + 6.1 \cdot 1 = 74.8$

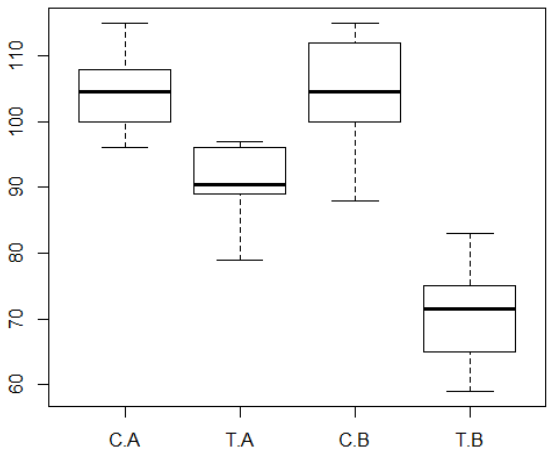
Observe que ambos coeficientes son aditivos.

3.8. a) Cierta. b) Cierta. c) Cierta. d) Cierta Son ciertas a), b) y d)] Escrib bien c.

3.9. En la izquierda puede verse que el efecto (aproximado) de cambiar de C a T es 20 mmHg en ambos centros: en A baja de 95 a 75 y en B de 90 a 70. En cambio, en la figura de la derecha hay interacción, ya que en el centro A baja 15 mmHg (105 a 90) y en el centro B 35 mmHg (105 a 70).

3.10.

```
#Descriptiva por grupo de tratamiento
> with(datos,mean(PAD4))
[1] 92.65
> with(datos,by(PAD4,list(Tratamiento),summary))
: C
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   88.0   100.0   104.5   104.6   110.2   115.0
------------------------------------------------
: T
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  59.00   72.75   81.50   80.75   90.25   97.00
> with(datos,by(PAD4,list(Centro),summary))
: A
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  79.00   90.75   96.50   97.60  104.20  115.00
------------------------------------------------------
----------
: B
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  59.00   72.75   85.50   87.70  103.80  115.00
#Descriptiva por grupo de tratamiento y centro
> with(datos,by(PAD4,list(Tratamiento,Centro),summary))
: C
: A
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   96.0   100.2   104.5   104.2   107.5   115.0
------------------------------------------------
: T
: A
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  79.00   89.25   90.50   91.00   96.00   97.00
------------------------------------------------
: C
: B
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   88.0   100.5   104.5   104.9   111.8   115.0
------------------------------------------------
: T
: B
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  59.00   65.50   71.50   70.50   74.75   83.00
```



	C	T	Todos

A	104.2	91.0	97.6
B	104.9	70.50	87.7
Todos	104.6	80.75	92.65

Para estimar el efecto en global y en cada centro seguimos los mismos pasos que en el Ejemplo 3.2: En un EC sobre el efecto de un consejo dietético-higiénico profundo (T) frente al convencional (C), se han obtenido mediciones en 2 centros de atención primaria (A y B).:

```
#Efecto global:
> mod6<-lm(PAD4~Tratamiento+Centro,datos)
> summary(mod6)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   109.500      2.453  44.638  < 2e-16 ***
TratamientoT  -23.800      2.833  -8.402 4.21e-10 ***
CentroB        -9.900      2.833  -3.495  0.00125 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 8.957 on 37 degrees of freedom
Multiple R-squared:  0.6912,      Adjusted R-squared:  0.6745
--------------------------------------------------

#Efecto Centro A:
> mod4 <- lm(PAD4 ~ Tratamiento ,data=subset(datos,Centro=='A'))
> summary(mod4)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   104.200      1.828  56.997  < 2e-16 ***
TratamientoT  -13.200      2.585  -5.106 7.39e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 5.781 on 18 degrees of freedom
Multiple R-squared:  0.5915,      Adjusted R-squared:  0.5688
--------------------------------------------------

#Efecto Centro B:
> mod5 <- lm(PAD4 ~ Tratamiento ,data=subset(datos,Centro=='B'))
> summary(mod5)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   104.900      2.628  39.912  < 2e-16 ***
TratamientoT  -34.400      3.717  -9.255  2.9e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 8.311 on 18 degrees of freedom
Multiple R-squared:  0.8263,      Adjusted R-squared:  0.8167
```

Observe en este caso que el efecto global es -23.8=80.8-104.6; el del centro A, -13.2=91.0-104.2; y el del centro B, -34.4=70.5-104.9.

En este caso, como ya ha podido ver gráficamente, existe interacción. Si existe interacción hablar de efecto global indicará un efecto intermedio de los dos centros (en este caso), pero no da información útil de ninguno de ellos en particular.

3.11. `> mod.interaccion <- lm(PAD4 ~ Tratamiento * Centro,data=datos)`

```
> summary(mod.interaccion)
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          104.200      2.264  46.028  < 2e-16 ***
TratamientoT         -13.200      3.202  -4.123  0.00021 ***
CentroB                0.700      3.202   0.219  0.82816
TratamientoT:CentroB  -21.200      4.528  -4.682 3.95e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 7.159 on 36 degrees of freedom
```

```
Multiple R-squared:  0.8081,      Adjusted R-squared:  0.7921
```

En este caso, el Intercept es la PAD esperada para una persona que esté en el centro A y se le aplique el tratamiento C (categorías '0' de cada *dummy*); el coeficiente de TratamientoT es el cambio de tratamiento C al T en el centro A (*dummy* de centro =0),; el coeficiente de CentroB es el cambio de A a B en los controles (*dummy* de trat=0); por último, el efecto de TratamientoT:CentroB es el efecto adicional del trat T enaquellas personas que están en el centro B: es la diferencia de los efectos, es decir (TB - TA) - (CB - CA)

3.12. `#Modelo aditivo`

```
> summary(mod.ad)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   109.500      2.453  44.638  < 2e-16 ***
TratamientoT  -23.800      2.833  -8.402 4.21e-10 ***
CentroB        -9.900      2.833  -3.495  0.00125 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 8.957 on 37 degrees of freedom
Multiple R-squared:  0.6912,      Adjusted R-squared:  0.6745

#Modelo con interacción
> summary(mod.in)

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         104.200      2.264  46.028  < 2e-16 ***
TratamientoT        -13.200      3.202  -4.123  0.00021 ***
CentroB               0.700      3.202   0.219  0.82816
TratamientoT:CentroB  -21.200      4.528  -4.682 3.95e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 7.159 on 36 degrees of freedom
Multiple R-squared:  0.8081,      Adjusted R-squared:  0.7921
```

Los errores típicos del modelo aditivo son más bajos que los del modelo con interacción porque cuando se estima un efecto global los errores típicos de las estimaciones son menores, ya que se asume que el efecto es el mismo para los diferentes grupos y la estimación es más precisa. Si observa el error típico de la variable TratamientoT:CentroB (término de interacción) verá que es mayor que el resto de errores típicos. Esto se debe a que la varianza de esta diferencia es mayor que la varianza de cada término.

3.13. `> tapply(birthwt$bwt, list(birthwt$smoke, birthwt$low), mean)`

```
         0        1
0 3394.802 2050.069
1 3200.705 2143.033
#Cálculo del modelo
> modelo<-lm(bwt~low*smoke, data=birthwt)
> summary(modelo)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3394.80      48.34  70.233   <2e-16 ***
low         -1344.73      96.26 -13.970   <2e-16 ***
smoke        -194.10      83.08  -2.336   0.0206 *
low:smoke     287.06     143.28   2.003   0.0466 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 448.3 on 185 degrees of freedom
Multiple R-squared:  0.6282,      Adjusted R-squared:  0.6221

#Intercept = 3394.80   ([0,0] en la tabla)
#Efecyo low ([0,1] en la tabla)
> 3394.80-1344.73
[1] 2050.07
```

```
#Efecto smoke
> 3394.80 -  194.10 ([1,0] en la tabla)
[1] 3200.7
#Efecto interacción low smoke ([1,1] en la tabla)
> 3394.80 -1344.73  - 194.10 +   287.06
[1] 2143.03
```

3.14. a) Cierto. b) Falso, cuando hay interacción los coeficientes del modelo incluidos en la misma son más complicados de interpretar y se recomienda no realizar una interpretación del efecto global, si no realizar la interpretación del efecto por grupos. c) Falso, la potencia de la prueba de hipótesis de la interacción tiene menor potencia estadística que la del efecto de la intervención. d) Cierta.

4.1.
```
> summary(mod.lm)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 45.91158    9.97406   4.603 5.02e-05 ***
grptreat     5.90756    2.16784   2.725  0.00986 **
fbef         0.61566    0.08937   6.889 4.60e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.976 on 36 degrees of freedom
Multiple R-squared:  0.6666,      Adjusted R-squared:  0.6481

> confint(mod.lm)
                 2.5 %     97.5 %
(Intercept) 25.6832527 66.1399007
grptreat     1.5109824 10.3041452
fbef         0.4344046  0.7969238
```

(1) Para saber cuánto mayor es la movilidad final según el grado inicial de movilidad hay que mirar el coeficiente de *fbef*=0.61566; esto quiere decir que por cada grado inicial de movilidad se espera un aumento de la movilidad final de 0.61566 grados. El $IC_{95\%}$ para esta variable va de [0.43 0.80], por lo tanto en el 95% de los casos se espera que por cada grado inicial de movilidad que tenga el individuo su movilidad final aumente entre 0.43 y 0.80 grados. (2) Querría decir que por cada grado de movilidad inicial que el individuo tenga aumenta en un grado la movilidad final. (3) El hecho de aplicar la intervención (*treat*) implica un aumento esperado en la movilidad final de 5.9 grados; el $IC_{95\%}$ va de 1.511 a 10.304, esto implica que en el 95% de los casos aplicar el tratamiento provocará un aumento en los grados de movilidad final de entre 1.511 y 10.604 grados. (4) En caso de modelo aditivo no, ya que el valor de los coeficientes es independiente del resto de variables. (5) La capacidad predictiva del modelo es de un 66.66%.

4.2.
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -89.7815    29.1252  -3.083  0.00675 **
altura        0.9501     0.1645   5.776 2.23e-05 ***
generoMujer  -8.3142     3.0175  -2.755  0.01352 *
---
> # IC para los coeficientes
> confint(mod.lm1)
                   2.5 %     97.5 %
(Intercept) -151.2303425 -28.332661
altura         0.6030784   1.297108
generoMujer  -14.6806971  -1.947753
```

La estimación de la constante es -89.8, con un IC95% = [-151.2 a -28.3], muy amplio. La estimación del coeficiente de la altura es 0.95, IC95% = [0.6 a 1.3]: el peso aumenta entre 0.6 y 1.3 Kg por cada cm adicional de la altura. Al no incluir el valor 0 de independencia, peso y altura están relacionadas: la altura contribuye a disminuir la incertidumbre sobre el peso. Para el género la estimación es -8.3, IC95% = [-14.7 a -1.9], también significativa al no incluir el 0. Estos intervalos de confianza son amplios por las pocas observaciones de las que se dispone. El

residuo típico o variabilidad no explicada por el modelo (σε) es 6.02. Representa el 20% de la variabilidad total, ya que la explicada es el 80% (R^2 = 0.80).]

4.3. Por cada año más que el individuo tenga (indiferente si es hombre o mujer) se espera un aumento de la PAS de 0.4194 unidades; si además, es individuo es de sexo femenino se espera que la PAS sea 8.91 unidades más baja que en los individuos de sexo masculino.

Los IC serían: coeficiente *Intercept*: $IC_{95\%}$: 103.01 ± 5.76 = [97.25 , 108.78]; coeficiente *edad*: $IC_{95\%}$: 0.4194 ± 0.1094 = [0.31 , 0.53]; coeficiente *generoMujer:* $IC_{95\%}$: -8.9092 ± 2.67 = [-11.58 , -6.24].

4.4.
```
> mod<-lm(PAS~edad*genero,datos)
> summary(mod)
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      100.71935    8.21482  12.261  < 2e-16 ***
edad               0.46547    0.16051   2.900  0.00495 **
generoMujer       -4.56463   11.34610  -0.402  0.68865
edad:generoMujer  -0.08687    0.22041  -0.394  0.69464
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.72 on 72 degrees of freedom
Multiple R-squared:  0.2592,      Adjusted R-squared:  0.2283

> confint(mod)
                      2.5 %      97.5 %
(Intercept)       84.3434084 117.0953004
edad               0.1454887   0.7854463
generoMujer      -27.1826610  18.0534073
edad:generoMujer  -0.5262433   0.3525024
```

1) En este caso el IC que interesa es el del Intercept, por lo tanto el $IC_{95\%}$ para el incremento en la PAS por año de edad en los hombres va de 84.34 a 117.10. 2) El $IC_{95\%}$ para las mujeres va de -27.18 a 18.05.

5.1. Aplica a ambos.

5.2. Simplemente porque el pronóstico hace referencia al futuro, no al presente.

5.3. Una vez se tiene el modelo debe realizarse una validación interna (los datos utilizados son los que provienen de la muestra del estudio; los métodos más comunes son *bootstrapping* o validación cruzada) y una validación externa (datos diferentes a los utilizados para la estimación del modelo).

5.4. a) Cierta. b) Falsa, un estudio transversal permite estimar modelos múltiples diagnósticos. c) Falsa, Un estudio longitudinal pretende estimar modelos múltiples pronósticos. d) Cierta. e) Cierta, al querer aplicar a diferentes momentos del tiempo resultados obtenidos en un solo tendremos que hacer un montón de premisas adicionales. Por ejemplo, es bien conocido que para hablar de causalidad se requiere discutir primero qué variable tira de qué variable.