ANCOVA: Ejemplo **colesterol**

Supongamos que estamos interesados en medir el efecto de un nuevo medicamento para bajar el nivel de colesterol en sangre. Nos interesa saber si este medicamento produce mejores resultados que el medicamento estándar. Podríamos tener dos grupos de pacientes:

- uno recibe la medicación nueva (grupo tratamiento)
- el otro recibe el tratamiento estándar (grupo control)

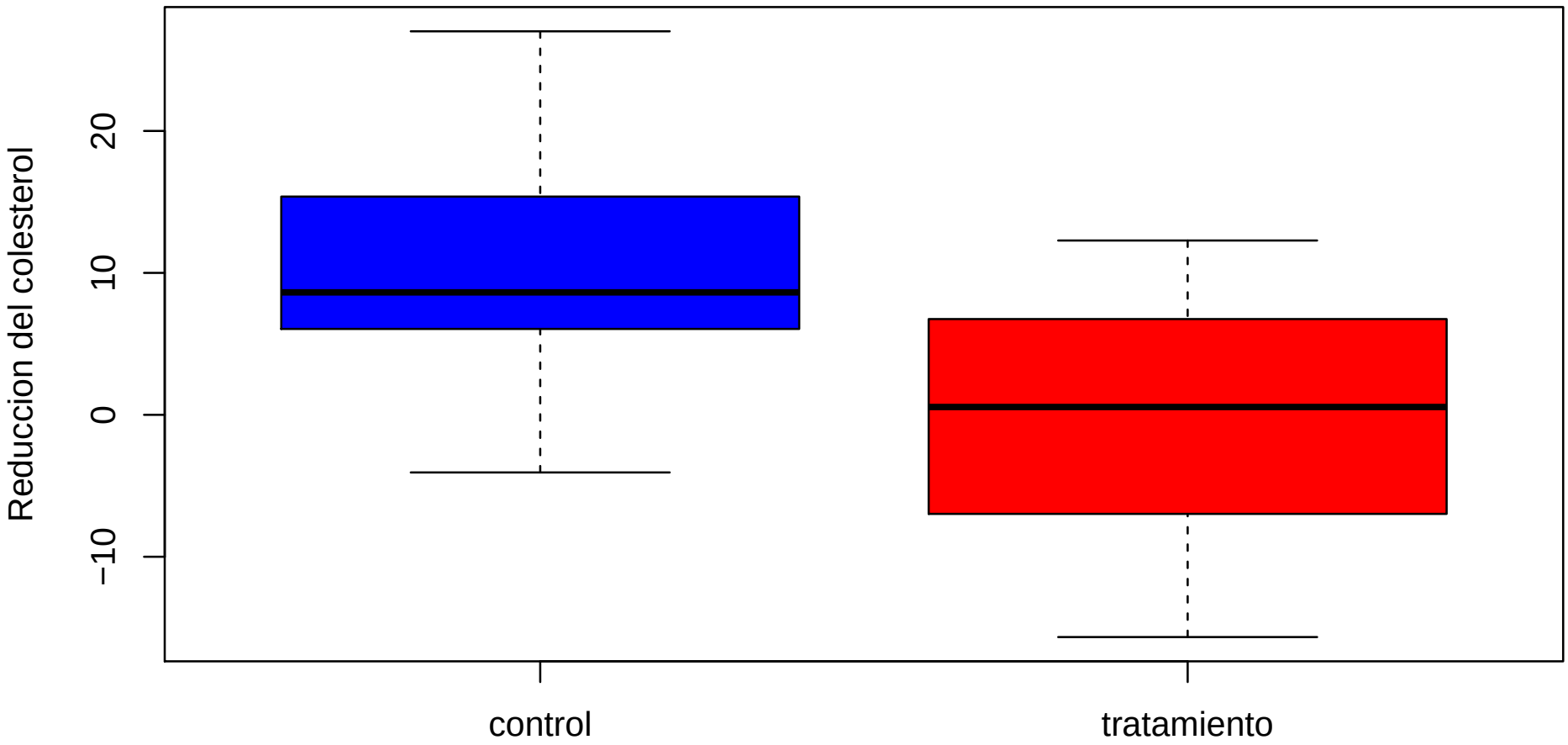
Para cada paciente se registra la edad y la variable respuesta: reducción porcentual en el nivel de colesterol luego del tratamiento, `reducccol`

`reducccol` = reducción porcentual de colesterol

Si supiéramos que los dos grupos difieren con respecto a la edad, no podríamos tratar esto como un simple problema de comparación de dos muestras ya que la edad afecta el nivel de colesterol. Veamos un ejemplo simulado en lo que sigue. Datos `colesterol.txt`. Script `ancovasimu2.R`

(Ejemplo simulado siguiendo a Faraway, J. *Linear Models with R*, Chapman and Hall CRC, 2009)

Ejemplo **colesterol**: Boxplot de reducción del colesterol por grupos



ANCOVA: Ejemplo **colesterol**

Promedio (sd) de cada variable en la muestra de 20 pacientes por grupo

	control	tratamiento
`reducccol`	10.58 (7.75)	0.01 (8.65)
`edad`	39.50	59.80

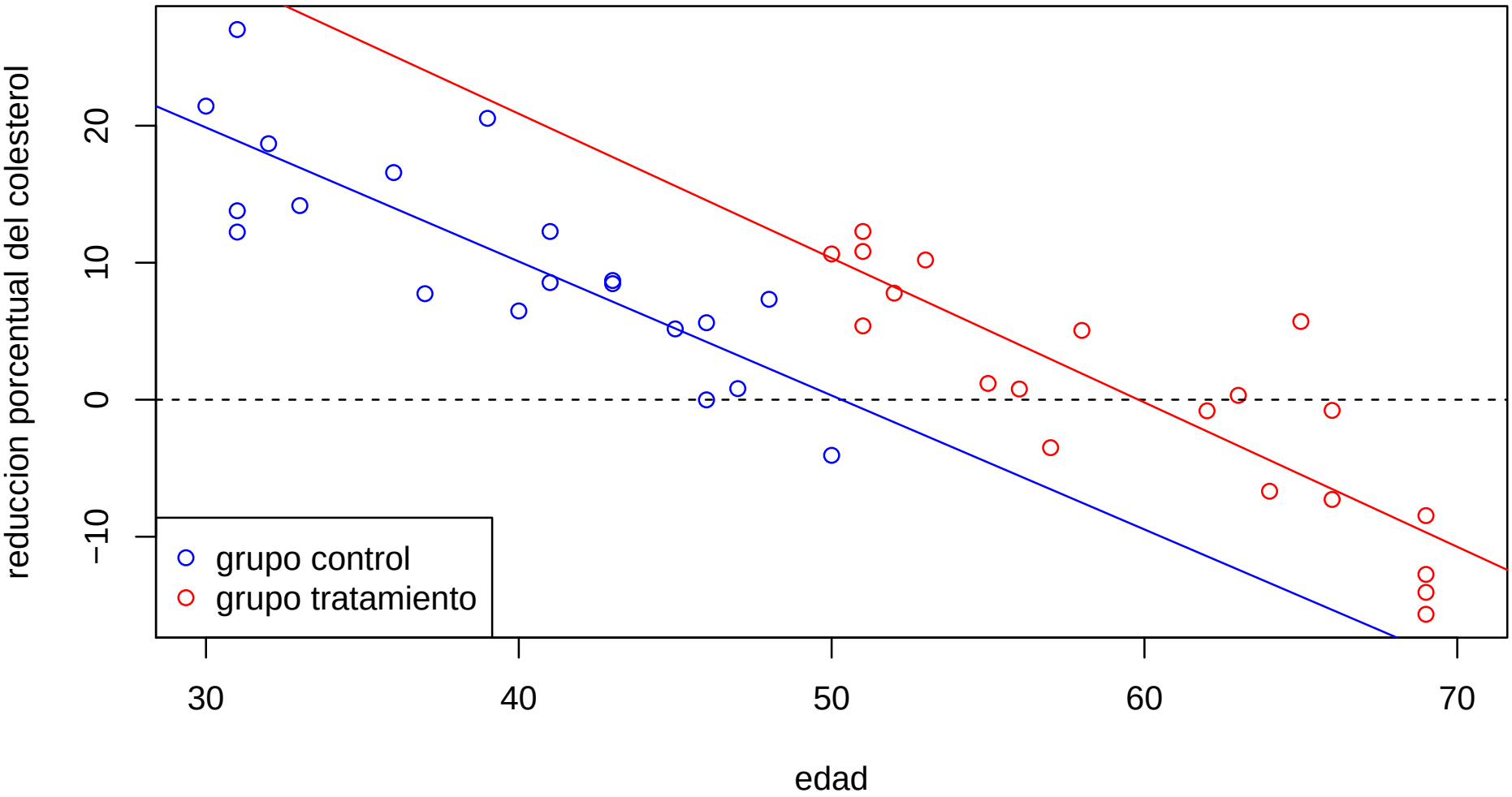
Para los pacientes que recibieron la medicación, la reducción media en el nivel de colesterol fue del 0%, mientras que para los que no lo hicieron, la reducción media fue del 10%. Luego, concluiríamos que es mejor no ser tratado. Sin embargo, la edad promedio de los pacientes del grupo tratamiento fue de 60 años, mientras que la edad promedio de los pacientes del grupo control fue de 40 años. Queremos hacer una comparación entre tratamientos *que tenga en cuenta la edad*. O, que *controle por edad*.

ANCOVA: Ejemplo **colesterol**

```
> t.test(reducccol[trat==1],reducccol[trat==0], var.equal = T)
Two Sample t-test
data:  reducccol[trat == 1] and reducccol[trat == 0]
t = -4.0679, df = 38, p-value = 0.0002307
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
-15.82639  -5.30861
sample estimates:
mean of x mean of y
0.0095   10.5770
```

Grafiquemos la respuesta por edad, en colores distintos segun el grupo al que pertenece el paciente. Superponemos la recta de ajuste por mínimos cuadrados que se obtiene con los 20 pacientes de cada grupo.

ANCOVA: Ejemplo **colesterol**



ANCOVA: Ejemplo **colesterol**

Definimos una variable indicadora del tratamiento ($\texttt{trat} = 1$ si el paciente pertenece al grupo tratamiento, 0 sino). El modelo más general que proponemos es (con interacción)

$$Y_i = \beta_0 + \beta_1 \texttt{edad}_i + \beta_2 \texttt{trat}_i + \beta_{1:2}(\texttt{trat} \cdot \texttt{edad})_i + \varepsilon_i, \qquad (1)$$

Es decir, que para los pacientes del grupo control (i.e. $\texttt{trat}_i = 0$) tenemos

$$Y_i = \beta_0 + \beta_1 \texttt{edad}_i + \varepsilon_i,$$

y que para los pacientes tratados ($\texttt{trat}_i = 1$) tenemos

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 \texttt{edad}_i + \beta_2 + \beta_{1:2}\texttt{edad}_i + \varepsilon_i \\ &= \beta_0 + \beta_2 + (\beta_1 + \beta_{1:2})\texttt{edad}_i + \varepsilon_i \end{aligned}$$

¡rectas no necesariamente paralelas! ¿Cómo se relacionan las ordenadas al origen entre sí? O sea, 2 rectas libres (sin ninguna atadura entre ambos modelos).

Colesterol: Ajuste del modelo completo

```
> trat.f<-factor(tratamiento)
> modelo1<-lm(reducccol~ edad*trat.f)
> summary(modelo1)

             Estimate Std. Error t value Pr(>|t|)
(Intercept)   49.2062     6.3317   7.771 3.28e-09 ***
edad          -0.9780     0.1583  -6.179 4.01e-07 ***
trat.f1       13.8116    10.7917   1.280    0.209
edad:trat.f1  -0.0757     0.2148  -0.352    0.727
---
Residual standard error: 4.496 on 36 degrees of freedom
Multiple R-squared:  0.8023,Adjusted R-squared:  0.7858
F-statistic:  48.7 on 3 and 36 DF,  p-value: 9.36e-13
```

El coeficiente de la interacción no es significativo. ¿Qué quiere decir?

Colesterol: Ajuste del modelo completo

```
> interac<-(edad*tratamiento)
> cor(cbind(edad,tratamiento,interac))
                edad tratamiento   interac
edad       1.0000000   0.8366529 0.8908923
tratamiento 0.8366529   1.0000000 0.9868524
interac     0.8908923   0.9868524 1.0000000
```

La interacción y el tratamiento están muuuy correlacionadas (repiten información). Probemos ajustar un modelo sin interacción

Colesterol: modelo aditivo

El modelo que proponemos es

$$Y_i = \beta_0 + \beta_1 \texttt{edad}_i + \beta_2 \texttt{trat}_i + \varepsilon_i, \tag{2}$$

Es decir, que para los pacientes del grupo control (i.e. $\texttt{trat}_i = 0$) tenemos

$$Y_i = \beta_0 + \beta_1 \texttt{edad}_i + \varepsilon_i,$$

y que para los pacientes tratados ($\texttt{trat}_i = 1$) tenemos

$$Y_i = \beta_0 + \beta_2 + \beta_1 \texttt{edad}_i + \varepsilon_i$$

¡rectas paralelas! ¿Cómo son sus ordenadas al origen?

Colesterol: ajuste del modelo aditivo

```
> modelo2<-lm(reducccol~ edad+trat.f)
> summary(modelo2)
Estimate Std. Error t value Pr(>|t|)
(Intercept)  50.8301     4.2919  11.843 3.76e-14 ***
edad         -1.0191     0.1057  -9.641 1.23e-11 ***
trat.f1      10.1195     2.5648   3.946 0.000342 ***
---
Residual standard error: 4.442 on 37 degrees of freedom
Multiple R-squared:  0.8016,Adjusted R-squared:  0.7909
F-statistic: 74.76 on 2 and 37 DF,  p-value: 1.007e-13
```

Todos los coeficientes son significativos. Vemos que una vez que se tiene en cuenta la edad, la diferencia entre tratamiento y control nuevamente es del 10%, pero esta vez a favor del tratamiento.

Colesterol: modelo aditivo

El modelo ajustado es

$$\widehat{Y}_i = 50.83 - 1.02 \cdot \texttt{edad}_i + 10.12 \cdot \texttt{trat}_i,$$

Es decir, que para los pacientes del grupo control (i.e. $\texttt{trat}_i = 0$) tenemos

$$\widehat{Y}_i = 50.83 - 1.02 \cdot \texttt{edad}_i$$

y que para los pacientes tratados ($\texttt{trat}_i = 1$) tenemos

$$\widehat{Y}_i = 60.95 - 1.02 \cdot \texttt{edad}_i$$

¿Colesterol: por qué usar un modelo lineal

El análisis de la covarianza, utiliza la relación entre la variable de respuesta (reducción en el nivel de colesterol, en nuestro ejemplo) y una o más variables cuantitativas para las cuales hay información disponible (edad, en nuestro ejemplo) para reducir la variabilidad del término del error y permitir que la comparación entre los grupos sea más poderosa. Se lo suele denominar *controlar por la variable edad*.

El interés está puesto en la comparación de la respuesta en ambos grupos, pero los grupos difieren en características que pueden ser tenidas en cuenta en el modelo.

Esto sucedió en el ejemplo del colesterol, en el cual si nos quedamos con el análisis sin covariables (comparación de medias con el `t.test` y boxplot) concluimos equivocadamente que conviene quedarse con el tratamiento estándar, sin embargo, cuando incluimos a la edad como covariable ("controlamos por la edad") podemos llegar a la conclusión correcta: el nuevo tratamiento reduce más el colesterol que el estándar.

Comparación con dos modelos lineales por separado

¿Qué diferencia hay con usar dos modelos lineales, uno para cada grupo de pacientes?

1. El modelo aditivo (2) asume pendientes iguales y la misma varianza del error de para cada observación. En consecuencia, la pendiente común β_1 se puede estimar mejor usando la información en la muestra conjunta. Ojo, este modelo no debería usarse si no se cree que este supuesto sea correcto para los datos a analizar.
2. Se puede testear si el modelo de rectas paralelas es correcto
3. Usando el modelo aditivo otras inferencias, como por ejemplo las realizadas sobre β_0 yβ_2 resultarán más precisas pues se dispone de más observaciones para estimarlas y para estimar a σ^2 la varianza del error (lo que se traduce en más grados de libertad para estimarlos).

Dos ajustes simples

Grupo control

```
> summary(lm(reducccol[tratamiento==0]~ edad[tratamiento==0]))
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             49.2062     6.3794   7.713 4.11e-07 ***
edad[tratamiento == 0]  -0.9780     0.1595  -6.133 8.59e-06 ***
---
Residual standard error: 4.53 on 18 degrees of freedom
Multiple R-squared:  0.6763,Adjusted R-squared:  0.6584
F-statistic: 37.61 on 1 and 18 DF,  p-value: 8.591e-06
```

Grupo tratamiento

```
> summary(lm(reducccol[tratamiento==1]~ edad[tratamiento==1]))
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             63.0178     8.6726   7.266 9.38e-07 ***
edad[tratamiento == 1]  -1.0537     0.1441  -7.314 8.58e-07 ***
---
Residual standard error: 4.462 on 18 degrees of freedom
Multiple R-squared:  0.7482,Adjusted R-squared:  0.7342
F-statistic: 53.49 on 1 and 18 DF,  p-value: 8.578e-07
```

Otro ejemplo, pág. 926 Neter

`datos.txt`, analizados en el script `ancovasimu2.R`

Una empresa estudió los efectos de tres tipos diferentes de promociones en las ventas de sus galletitas de agua.

- Tratamiento 1: Entrega de muestras gratis del producto a los clientes en la tienda, y espacio habitual/regular en los estantes del comercio
- Tratamiento 2: Espacio adicional en los estantes del comercio, en la ubicación regular
- Tratamiento 3: Estantes de exhibición especiales en los extremos del pasillo, además del espacio en los estantes regulares

Se seleccionaron quince comercios para el estudio. A cada tienda se le asignó al azar uno de los tipos de promoción, con cinco tiendas asignadas a cada tipo de promoción. Otras condiciones relevantes bajo el control de la empresa, como el precio y la publicidad, se mantuvieron iguales para todas las tiendas del estudio.

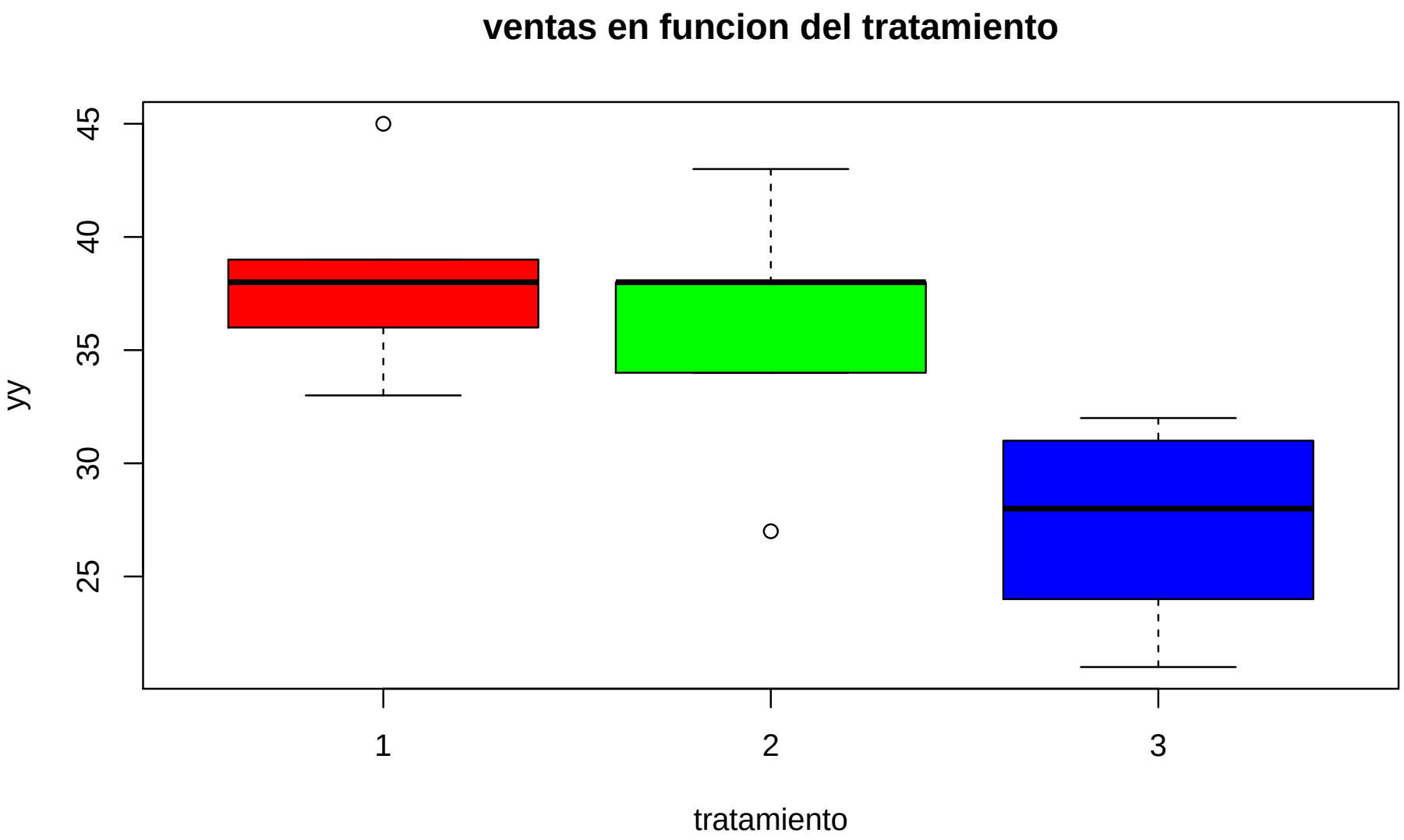
Otro ejemplo, pág. 926 Neter

La variable de interés es el número de unidades del producto vendidas durante el período promocional, denotado por Y. Se cuenta con los datos sobre las ventas del producto en el período anterior al promocional, denotado por X. Utilizaremos las ventas en el período anterior se como variable explicativa. ¿Por qué es razonable esto?

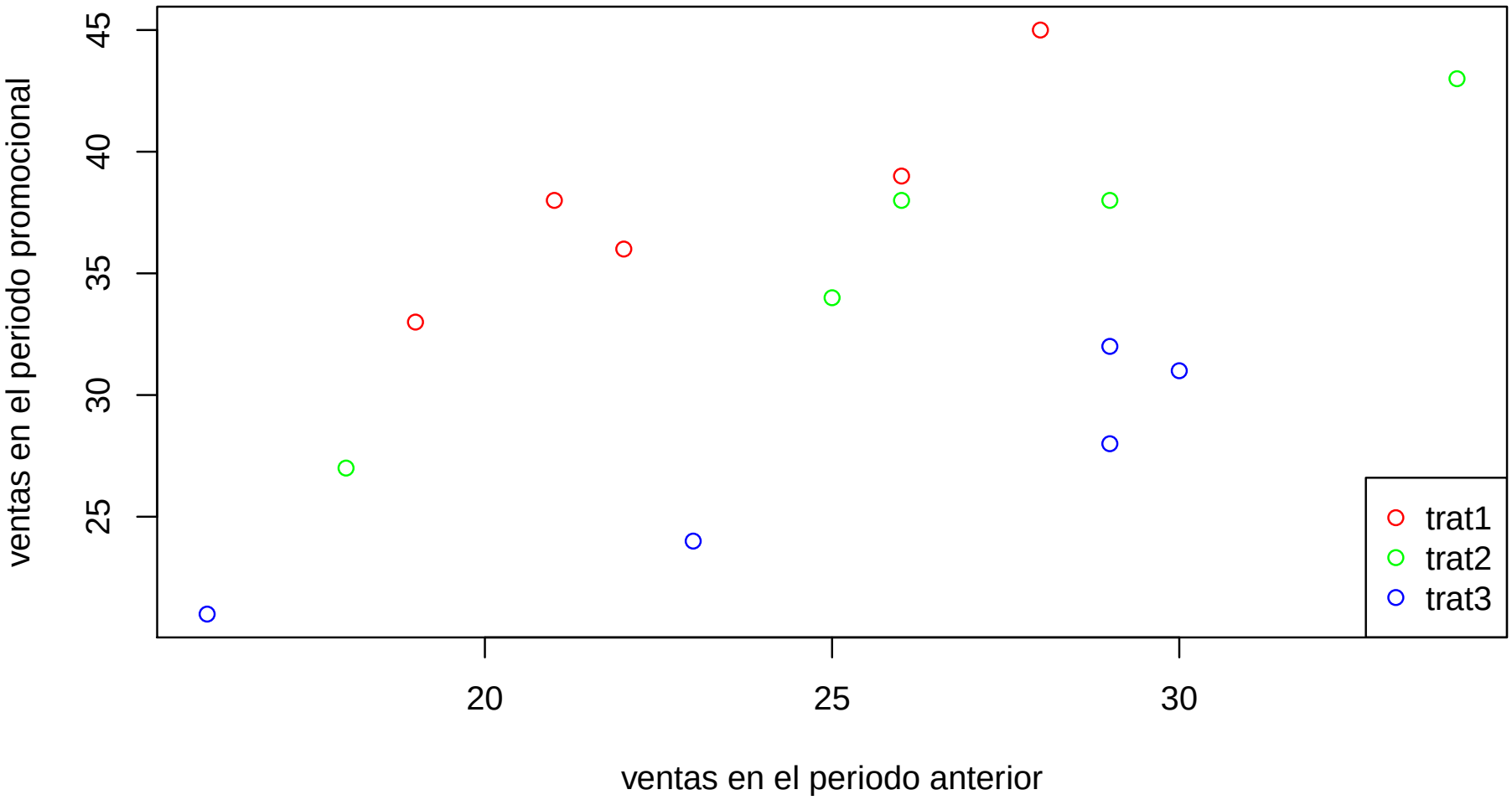
Los datos son

com	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
trat	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
xx	21	34	23	26	26	29	22	29	30	28	18	16	19	25	29
yy	38	43	24	39	38	32	36	38	31	45	27	21	33	34	28

Boxplot por tratamiento



Y versus X, separados por tratamiento



Ajuste del modelo aditivo

```
> ajuste2<-lm(yy ~ xx+trat.f)
> summary(ajuste2)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.3534     2.5230   6.878 2.66e-05 ***
xx            0.8986     0.1026   8.759 2.73e-06 ***
trat.f2      -5.0754     1.2290  -4.130  0.00167 **
trat.f3     -12.9768     1.2056 -10.764 3.53e-07 ***
---
Residual standard error: 1.873 on 11 degrees of freedom
Multiple R-squared:  0.9403,Adjusted R-squared:  0.9241
F-statistic: 57.78 on 3 and 11 DF,  p-value: 5.082e-07
```

¿Qué hipótesis nos interesa testear? ¿Qué hipótesis se testean en esta salida?

Evaluamos el efecto tratamiento en el modelo aditivo

Tarea: hacerlo usando los tests vistos antes, multiplicando por matrices adecuadas, y comparar lo obtenido con estas dos salidas automáticas

```
> anova(ajuste2)
Analysis of Variance Table
Response: yy
Df Sum Sq Mean Sq F value    Pr(>F)
xx          1 190.68 190.678  54.379 1.405e-05 ***
trat.f      2 417.15 208.575  59.483 1.264e-06 ***
Residuals 11  38.57   3.506
---
> ajuste3<-lm(yy ~ xx)
> anova(ajuste3,ajuste2)
Analysis of Variance Table
Model 1: yy ~ xx
Model 2: yy ~ xx + trat.f
Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     13 455.72
2     11  38.57  2    417.15 59.483 1.264e-06 ***
```

Comparaciones de a pares

Intervalos Scheffé para la comparacion de los tres grupos

```
AA<-matrix(rbind(c(0,0,-1,0),c(0,0,0,-1),c(0,0,1,-1)),ncol=4)
> AA
[,1] [,2] [,3] [,4]
[1,]    0    0   -1    0
[2,]    0    0    0   -1
[3,]    0    0    1   -1

library(Matrix)
rr<-rankMatrix(AA)[1]   # rr = 2, rango de AA
k<-ajuste2$rank    # = 4, cant de parametros estimados
n<-length(xx)   # = 15, cantidad de datos
alfa<-0.05
sqrt(rr*qf(1-alfa,rr,n-k))
delt<-sqrt(rr*qf(1-alfa,rr,n-k))*sqrt(diag(AA%*%
vcov(ajuste2)%*%t(AA)))
```

Comparaciones de a pares

Intervalos Scheffé para la comparacion de los tres grupos

```
ICScheff<-data.frame(cbind(AA%*%coef(ajuste2)-delt,AA%*%
coef(ajuste2),AA%*%coef(ajuste2)+delt)
,row.names = c("1-2","1-3","2-3"))
names(ICScheff)<-c("ext.inf","dife","ext.sup")
ICScheff
> ICScheff
     ext.inf      dife   ext.sup
1-2 1.607052  5.075390  8.543728
1-3 9.574367 12.976831 16.379294
2-3 4.546608  7.901441 11.256273
```

Las tres medias difieren

Anova un factor

```
> ajuste1<-lm(yy~ trat.f)
> summary(ajuste1)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    38.200      2.264  16.871 1.01e-09 ***
trat.f2        -2.200      3.202  -0.687  0.50511
trat.f3       -11.000      3.202  -3.435  0.00494 **
---
Residual standard error: 5.063 on 12 degrees of freedom
Multiple R-squared:  0.5241,Adjusted R-squared:  0.4448
F-statistic: 6.609 on 2 and 12 DF,  p-value: 0.01161
```

Anova un factor

```
> anova(ajuste1)
Analysis of Variance Table
           Df Sum Sq Mean Sq F value  Pr(>F)
trat.f      2  338.8 169.400  6.6086 0.01161 *
Residuals 12  307.6  25.633
```

Intervalos Scheffe para la comparacion de los tres grupos en el modelo anova un factor sin covariable.

	ext.inf	dife	ext.sup
1-2	-6.73	2.20	11.13
1-3	2.07	11.00	19.93
2-3	-0.13	8.80	17.73

¡No detecta diferencias entre los tratamientos 1 y 2. Tampoco entre los 2 y 3!

Ajuste del modelo completo (descartemos la interacción)

```
> ajuste3<-lm(yy ~ xx*trat.f)
> summary(ajuste3)
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.88321    5.92451   2.175  0.05768 .
xx           1.09124    0.25281   4.317  0.00194 **
trat.f2     -3.05231    7.32063  -0.417  0.68649
trat.f3     -4.31947    7.19742  -0.600  0.56322
xx:trat.f2  -0.09999    0.29906  -0.334  0.74579
xx:trat.f3  -0.35753    0.29785  -1.200  0.26064
---
Residual standard error: 1.871 on 9 degrees of freedom
Multiple R-squared:  0.9512,Adjusted R-squared:  0.9241
F-statistic: 35.11 on 5 and 9 DF,  p-value: 1.216e-05
```

¿Qué hipótesis nos interesa testear? ¿Qué hipótesis se testean en esta salida?

Comparación del modelo completo y el aditivo

```
> anova(ajuste2,ajuste3)
Analysis of Variance Table

Model 1: yy ~ xx + trat.f
Model 2: yy ~ xx * trat.f
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     11 38.571
2      9 31.521  2    7.0505 1.0065 0.4032
```