Notas de Clase Regresión Lineal

Capítulo 6: Regresión con Variables Indicadoras

Nelfi González Alvarez
Profesora Asociada

Isabel Cristina Ramírez Guevara
Profesora Asociada

Escuela de Estadística

Universidad Nacional de Colombia, Sede Medellín

UNIVERSIDAD NACIONAL DE COLOMBIA

2021

Capítulo 6

Regresión con variables indicadoras

6.1. Introducción

Considere el caso de una variable predictora X medida en una escala nominal u ordinal (una variable categórica o cualitativa) definida en c categorías. Ej: En un estudio sobre calidad de vida, aparece la variable X estrato socioeconómico del grupo familiar, definido en las categorías 1, 2, 3, 4, 5 y 6. Aquí los valores 1 a 6 no representan información numérica y deben ser considerados como simples etiquetas. Suponga que se desea realizar una regresión lineal entre una variable cuantitativa Y vs. la variable categórica X. Dado la naturaleza cualitativa de ésta última no podemos simplemente formular el modelo de regresión como $Y_i = \beta_0 + \beta_1 X_i + E_i,\ E_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$. ¿Cuál es la estrategia a seguir?

6.2. Modelación con solo un predictor y éste es de naturaleza cualitativa

Para cada categoría de X definimos una variable indicadora, es decir una variable que toma el valor de 1 ó 0 según si la categoría considerada es o no observada en una unidad experimental o de observación: Sean las variables I_j, $j = 1, 2, \cdots, c$, tales que

$$I_j = \begin{cases} 1 & \text{si en la unidad experimental es observada la categoría } j \\ 0 & \text{si en la unidad experimental no es observada la categoría } j. \end{cases} \tag{6.1}$$

Es decir, I_j es la variable indicadora de la categoría j de la variable cualitativa X. Para una misma unidad experimental o de observación sólo una de las variables indicadoras puede tomar el valor de 1, es decir, para la i-ésima observación se cumple que $\sum_{j=1}^{c} I_{ij} = 1$, de aquí que no es necesario considerar las c indicadoras conjuntamente, ya que el valor de cualquiera de ellas en la unidad de observación i, digamos el valor de I_{ic}, puede hallarse como $I_{ic} = 1 - \sum_{j=1}^{c-1} I_{ij}$.
Podemos inicialmente, proponer el siguiente modelo de RLM para modelar la relación de Y vs. X a través del uso de las variables indicadoras,

$$Y_i = \beta_0 + \beta_1 I_{i1} + \beta_2 I_{i2} + \cdots + \beta_c I_{ic} + E_i,\ E_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right) \tag{6.2}$$

Considere de nuevo el caso donde X es el estrato socio económico y sea Y el gasto medio mensual total de la familia. Suponga que se obtuvo una muestra aleatoria simple de dos

familias por cada estrato, para un total de $n = 12$ observaciones, entonces matricialmente, tendríamos el siguiente sistema de ecuaciones,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{E} \Rightarrow \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \\ Y_8 \\ Y_9 \\ Y_{10} \\ Y_{11} \\ Y_{12} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \end{bmatrix} + \begin{bmatrix} E_1 \\ E_2 \\ E_3 \\ E_4 \\ E_5 \\ E_6 \\ E_7 \\ E_8 \\ E_9 \\ E_{10} \\ E_{11} \\ E_{12} \end{bmatrix} \tag{6.3}$$

Pero, en el sistema de ecuaciones tendríamos una dependencia lineal perfecta entre la primera columna de la matriz $\boldsymbol{X}$ y las columnas 2 a 7, lo cual haría que $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ no exista y por tanto, no podriamos estimar el vector de coeficientes. Para solucionar este inconveniente, se tienen tres posibles alternativas:

1. Eliminar el intercepto β_0 de la ecuación del modelo:

$$Y_i = \beta_1 I_{i1} + \beta_2 I_{i2} + \cdots + \beta_c I_{ic} + E_i, \; E_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right) \tag{6.4}$$

   En este caso, *los $\boldsymbol{\beta_j}$ representan la media de Y en la categoría j,* es decir, $\beta_j = E[Y|I_j = 1]$.

2. Eliminar una de las variables indicadoras, por ejemplo, la de la última categoría:

$$Y_i = \beta_0 + \beta_1 I_{i1} + \beta_2 I_{i2} + \cdots + \beta_{c-1} I_{i,c-1} + E_i, \; E_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right) \tag{6.5}$$

   En este caso, *$\boldsymbol{\beta_j}$, $j \neq c$, representa la diferencia de la respuesta media en la categoría j con relación a la media en la categoría c,* es decir, $\beta_j = E[Y|I_j = 1] - E[Y|I_c = 1]$, $j \neq c$.

3. Introducir la restricción $\sum_{j=1}^{c} \beta_j = 0$. El modelo sería el siguiente:

$$Y_i = \beta_0 + \beta_1 I_{i1} + \beta_2 I_{i2} + \cdots + \beta_c I_{ic} + E_i, \text{ sujeto a } \sum_{j=1}^{c} \beta_j = 0, \text{ con } E_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right) \tag{6.6}$$

   Aquí, cada coeficiente *$\boldsymbol{\beta_j}$, $j = 1, 2, \ldots, c$, representa el efecto de la categoría j con respecto a la la media general de la respuesta representada por el intercepto*, es decir, $\beta_j = E[Y|I_j = 1] - \beta_0$.

De estas tres alternativas usaremos la segunda.

6.3. Modelación con un predictor cuantitativo y otro cualitativo

Ahora suponga que se desea modelar por regresión lineal la relación de una variable respuesta cuantitativa Y vs. X_1, siendo X_1 cuantitativa, en presencia de una variable categórica X_2. Es decir, se quiere determinar si la relación lineal entre Y y X_1 depende de la variable categórica X_2. Asumiendo que X_2 es observada en c categorías, podemos considerar las dos siguientes situaciones:

Caso 1. *El efecto promedio de X_1 sobre la respuesta Y cambia según la categoría en que X_2 sea observada.*

Caso 2. *El efecto promedio de X_1 sobre la respuesta Y es el mismo en todas las categorías de X_2 pero la media general de Y no es igual en al menos dos de las categorías.*

En el primer caso es necesario considerar la interacción entre X_1 y X_2 y sólo utilizamos $c-1$ de las posibles variables indicadoras de las categorías de la variable X_2 (para evitar el problema antes descrito sobre la dependencia en las columnas de la matriz de diseño), es decir, postulamos el modelo,

$$Y_i = \beta_0+\beta_1 X_{i1}+\beta_2 I_{i1}+\beta_3 I_{i2}+\cdots+\beta_c I_{i,c-1}+\beta_{1,1} X_{i1}*I_{i1}+\beta_{1,2} X_{i1}*I_{i2}+\cdots+\beta_{1,c-1} X_{i1}*I_{i,c-1}+E_i,$$
$$E_i \overset{\text{iid}}{\sim} N\left(0,\sigma^2\right) \quad (6.7)$$

Observe que la ecuación anterior define c rectas de regresión simple de Y vs. X_1, una en cada categoría de la variable cualitativa X_2, así (ver Figura 6.1(a)):

- Si $I_1 = 1$, entonces el resto de indicadoras son iguales a cero y obtenemos,

$$Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_{1,1})X_1 + E$$

- Si $I_2 = 1$, entonces el resto de indicadoras son iguales a cero y obtenemos,

$$Y = (\beta_0 + \beta_3) + (\beta_1 + \beta_{1,2})X_1 + E$$

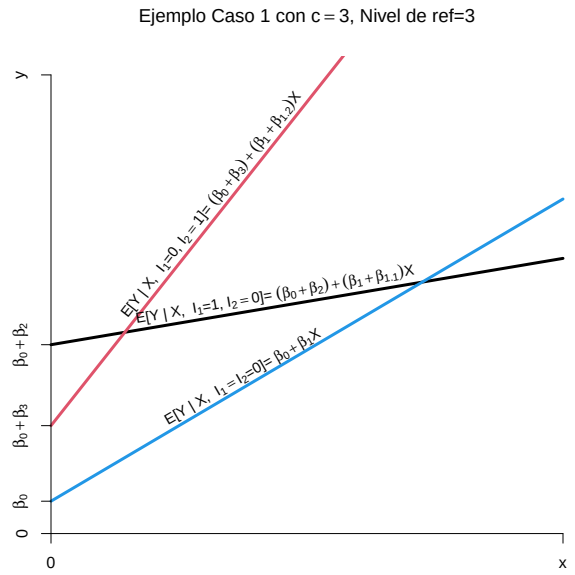
$\vdots$

- Si $I_{c-1} = 1$, entonces el resto de indicadoras son iguales a cero y obtenemos,

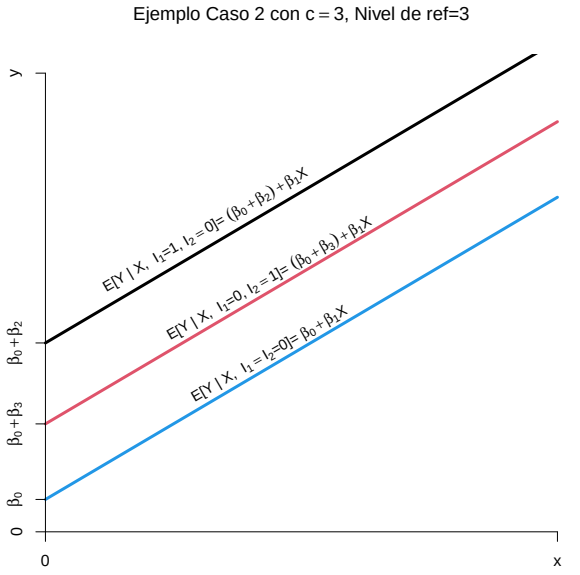
$$Y = (\beta_0 + \beta_c) + (\beta_1 + \beta_{1,c-1})X_1 + E$$

- Finalmente, si $I_1 = I_2 = \cdots I_{c-1} = 0$, necesariamente, la indicadora I_c no incluida en el modelo, debe ser igual a 1, así, cuando todas la indicadoras del modelo son simultáneamente cero, obtenemos la recta de regresión de Y vs. X_1, en la categoría c de la variable categórica X_2,

$$Y = \beta_0 + \beta_1 X_1 + E$$



(a)



(b)

Figura 6.1: (a) Ilustración caso 1, con $c = 3$ y nivel de referencia el 3ro: La relación lineal de Y vs. X cambia con niveles de la variable cualitativa. (b) Ilustración caso 2, con $c = 3$ y nivel de referencia el 3ro: El efecto medio de X sobre Y no cambia con niveles de la variable cualitativa, pero la media de Y no es igual para todos los niveles de la variable cualitativa

En el caso 2, el modelo a considerar es dado por

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 I_{i1} + \beta_3 I_{i2} + \cdots + \beta_c I_{i,c-1} + E_i,\ E_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right) \tag{6.8}$$

donde el efecto promedio de X_1 sobre la respuesta es el mismo, sin importar la categoría en que sea observada X_2, sin embargo, la media de Y no es la misma en todas las categorías, dado que las c ecuaciones resultantes, serían las de c rectas paralelas, que pueden diferir en el intercepto (ver Figura 6.1(b)):,

- Cuando $I_1 = 1$,
$$Y = (\beta_0 + \beta_2) + \beta_1 X_1 + E$$
- Cuando $I_2 = 1$,
$$Y = (\beta_0 + \beta_3) + \beta_1 X_1 + E$$

  $\vdots$

- Cuando $I_{c-1} = 1$,
$$Y = (\beta_0 + \beta_c) + \beta_1 X_1 + E$$
- Cuando $I_1 = I_2 = \cdots I_{c-1} = 0$, es decir, $I_c = 1$, tenemos
$$Y = \beta_0 + \beta_1 X_1 + E$$

6.4. Problema

Un gran almacén realizó un experimento para investigar los efectos de los gastos por publicidad sobre las ventas semanales de sus secciones de ropa para caballeros (A), para niños (B) y para damas (C). Se seleccionaron al azar 5 semanas para observación en cada sección, y un presupuesto para publicidad (X, en cientos de dólares) se asignó a cada una de las secciones. Las ventas semanales (Y, en miles de dólares), los gastos de publicidad en cada uno de las tres secciones en cada una de las cinco semanas del estudio se listan en la Tabla 6.1.

1. Analice en el gráfico de dispersión la relación entre las ventas y los gastos de publicidad según las secciones y globalmente.

2. Tomando como nivel de referencia la Sección C, postule y ajuste un modelo de regresión para estudiar los efectos que las secciones del almacén puedan tener sobre la relación de las ventas versus los gastos de publicidad. Halle las ecuaciones de las rectas ajustadas que relacionan las ventas con la publicidad en cada sección. Tomando como indicadoras de las secciones A, y B, a I_1 e I_2, respectivamente.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 I_{i1} + \beta_3 I_{i2} + \beta_{1,1} X_i I_{i1} + \beta_{1,2} X_i I_{i2} + E_i,\ E_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right)$$

Tabla 6.1: Datos observados

Sección	Publicidad	Ventas semanales
A	5.2	9
A	5.9	10
A	7.7	12
A	7.9	12
A	9.4	14
B	8.2	13
B	9.0	13
B	9.1	12
B	10.5	13
B	10.5	14
C	10.0	18
C	10.3	19
C	12.1	20
C	12.7	21
C	13.6	22

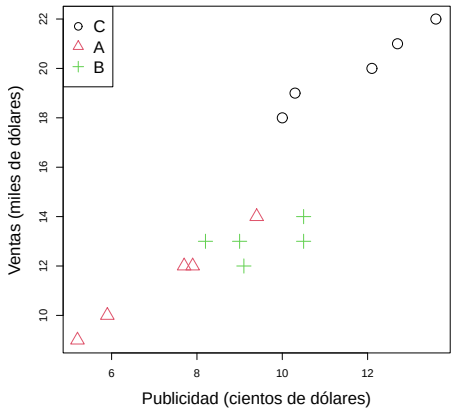


Figura 6.2: Gráfico de dispersión. Observaciones de sección identificadas por color y símbolo

Tabla 6.2: Tabla de parámetros estimados, según modelo general tomando como referencia la Sección C

Parámetro	Estimación	Error Estándar	T_0	$P(\|t_9\| > \|T_0\|)$	L.I(2.5 %)	L.S(97.5 %)
β_0	8.2747	1.7957	4.6081	0.0013	4.2125	12.3368
β_1 (Publicidad)	0.9988	0.1519	6.5751	0.0001	0.6551	1.3424
β_2 (Secc. A)	-5.2429	2.0724	-2.5299	0.0322	-9.9310	-0.5548
β_3 (Secc. B)	1.4888	2.8494	0.5225	0.6139	-4.9570	7.9346
$\beta_{1,1}$ (Publicidad*Secc. A)	0.1603	0.2068	0.7752	0.4581	-0.3075	0.6280
$\beta_{1,2}$ (Publicidad*Secc. B)	-0.6566	0.2780	-2.3621	0.0425	-1.2854	-0.0278
$\widehat{Y}_i = 8.2747 + 0.9988X_i - 5.2429I_{i1} + 1.4888I_{i2} + 0.1603X_iI_{i1} - 0.6566X_iI_{i2}$						
$R^2 = 0.9916$, $R^2_{adj} = 0.9869$						
Resultados para test Anova del modelo: $F_0 = 211.4$, $P(f_{5,9} > F_0) = 4.782 \times 10^{-9}$						

Sección	Modelo	Ecuación ajustada
A ($I_1 = 1, I_2 = 0$)	$Y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_{1,1})X_i + E_i,\ E_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$	$\widehat{Y}_i = 3.0318 + 1.1591X_i$
B ($I_1 = 0, I_2 = 1$)	$Y_i = (\beta_0 + \beta_3) + (\beta_1 + \beta_{1,2})X_i + E_i,\ E_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$	$\widehat{Y}_i = 9.7635 + 0.3422X_i$
C ($I_1 = 0, I_2 = 0$)	$Y_i = \beta_0 + \beta_1X_i + E_i,\ E_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$	$\widehat{Y}_i = 8.2747 + 0.9988X_i$

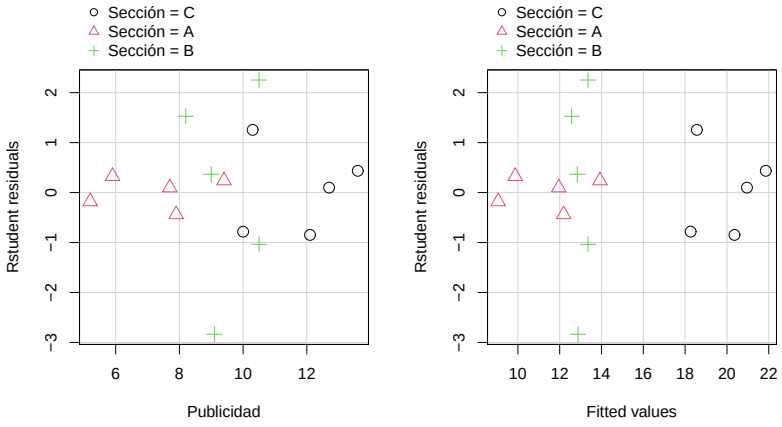
Nota 6.1. En R fue corrido el ajuste de este modelo de la siguiente forma:

```
> #CAMBIANDO NIVEL DE REFERENCIA DE LA VARIABLE CUALITATIVA PARA LA SECCIÓN C
> Sección=relevel(Sección,ref="C")
```

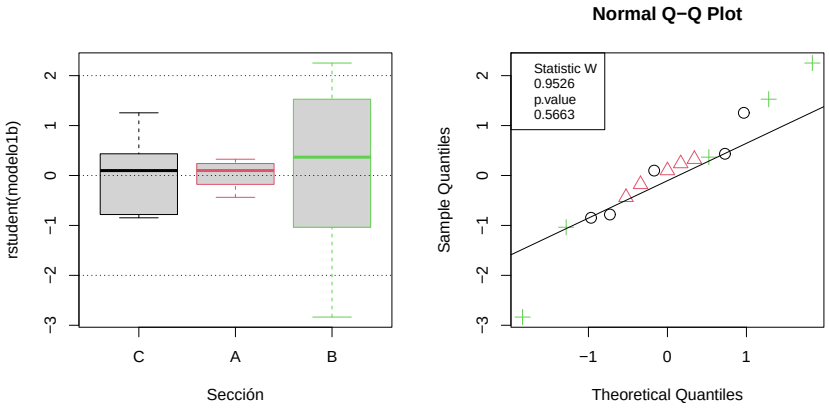
```
> Sección #observe que se informa que los niveles en su orden son C, A, B
 [1] A A A A A B B B B B C C C C C
Levels: C A B

> modelo1=lm(Ventas~Publicidad*Sección)
```

Vemos que no es necesario en principio crear las variables indicadoras, mientras sea la ecuación del modelo como fue dada previametente.

3. Realice análsis de residuales estudentizados ¿se cumplen supuestos de normalidad y varianza constante?



(a)



(b)

Figura 6.3: Gráficos para evaluación supuestos de los errores usando residuos estudentizados (externamente), modelo general

4. *Determine si existe diferencia entre las ordenadas en el origen de las rectas correspondientes a las secciones de caballeros (A) y de damas (C):* Las ordenadas al origen

son los interceptos, luego la igualdad de interceptos en las rectas de Secciones A y C implica que:

$$\beta_0 + \beta_2 = \beta_0 \iff \beta_2 = 0$$

entonces hay que probar $H_0 : \beta_2 = 0$ vs. $H_1 : \beta_2 \neq 0$. Basta hacer prueba t de significancia individual para β_2, ver Tabla 6.2.

5. *Determine si existe diferencia en las pendientes de las rectas correspondientes a las secciones de niños (B) y caballeros (A):* La igualdad de ls pendientes de las recta en secciones A y B implica que:

$$\beta_1 + \beta_{1,1} = \beta_1 + \beta_{1,2} \iff \beta_{1,1} = \beta_{1,2},$$

entonces hay que probar $H_0 : \beta_{1,1} = \beta_{1,2}$ vs. $H_1 : \beta_{1,1} \neq \beta_{1,2}$. Los resultados R pertinentes se muestran a continuación:

```
> linearHypothesis(modelo1,"Publicidad:SecciónA=Publicidad:SecciónB")
Linear hypothesis test
Hypothesis:
Publicidad:SecciónA - Publicidad:SecciónB = 0
Model 1: restricted model
Model 2: Ventas ~ Publicidad * Sección
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     10 3.9993
2      9 1.9960  1    2.0032 9.0326 0.01483 *
---
```

Estos resultados se interpretan como se muestra en la siguiente tabla:

Tabla 6.3: Resultados para prueba en numeral 5

Sumas cuadrados de error			$H_0: \beta_{1,1} = \beta_{1,2}$			
	d.f(SSE's)	SSE's	d.f(SC)	SC	F_0	$\Pr(f_{1,9} > F_0)$
Modelo nulo (MR)	10	3.9993				
Modelo completo (MF)	9	1.9960	1	2.0032	9.0326	0.01483
SC = SSE(MR) − SSE(MF); df(SC) = df[SSE(MR)] − df[SSE(MF)]						
F_0 = [SC ÷ 1] /MSE(MF)						

Nota 6.2. Bajo H_0, el modelo se reduce a (modelo nulo):

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 I_{i1} + \beta_3 I_{i2} + \beta_{1,1} X_i (I_{i1} + I_{i2}) + E_i, \; E_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right),$$

en este modelo los predictores son X, I_1, I_2 y el producto $X(I_1 + I_2)$. Para correrlo en R y ver sus estimaciones, después de ajustar el modelo completo, procedemos como sigue:

```
> #Obtener la matriz de diseño del modelo completo
> MATRIZ.DISEÑOmodelo1=as.data.frame(model.matrix(modelo1))
> names(MATRIZ.DISEÑOmodelo1) #Observe cómo son nombradas las variables en modelo 1
[1] "(Intercept)"         "Publicidad"          "SecciónA"
[4] "SecciónB"            "Publicidad:SecciónA" "Publicidad:SecciónB"
>
> #Modelo reducido en test "Publicidad:SecciónA=Publicidad:SecciónB",
> #o sea la igualdad de pendientes de las rectas en secciones A y B
> MR=lm(Ventas~Publicidad+SecciónA+SecciónB+Publicidad:I(SecciónA+SecciónB),
+       data=MATRIZ.DISEÑOmodelo1)
> summary(MR)
Call:
lm(formula = Ventas ~ Publicidad + SecciónA + SecciónB + Publicidad:I(SecciónA +
    SecciónB), data = MATRIZ.DISEÑOmodelo1)
Residuals:
     Min       1Q   Median       3Q      Max
-0.97906 -0.31086  0.02094  0.29059  1.18617
Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                       8.27466    2.41134   3.432 0.006422 **
Publicidad                        0.99875    0.20398   4.896 0.000626 ***
SecciónA                         -3.67161    2.69292  -1.363 0.202652
SecciónB                         -4.18036    2.86786  -1.458 0.175611
Publicidad:I(SecciónA + SecciónB) -0.05735    0.26008  -0.220 0.829928
---
Residual standard error: 0.6324 on 10 degrees of freedom
Multiple R-squared:  0.9831,    Adjusted R-squared:  0.9763
F-statistic: 145.3 on 4 and 10 DF,  p-value: 8.197e-09
```

Nota 6.3. En la prueba realizada en este numeral, la diferencia entre sumas de cuadrados de residuos de los modelos nulo y completo no corresponde a alguna suma de cuadrados extras.

6. *Si se quiere probar que la recta de ventas vs. publicidad es diferente para cada sección,* plantee la hipótesis a probar, el estadístico de prueba y región crítica al nivel de 0.05, realice la prueba y concluya: Las rectas serán iguales si coinciden sus interceptos y sus pendientes, entonces

$$\text{se requiere que: } \beta_0 + \beta_2 = \beta_0 + \beta_3 = \beta_0, \iff \beta_2 = \beta_3 = 0,$$

$$\text{también que: } \beta_1 + \beta_{1,1} = \beta_1 + \beta_{1,2} = \beta_1, \iff \beta_{1,1} = \beta_{1,2} = 0$$

luego, se debe probar

$H_0 : \beta_2 = \beta_3 = \beta_{1,1} = \beta_{1,2} = 0$ vs.

H_1 : al menos uno de estos parámetros es no nulo: $\beta_2, \beta_3, \beta_{1,1}, \beta_{1,2}$

Los resultados R necesarios son los siguientes:

```
> linearHypothesis(modelo1,c("SecciónA=0","SecciónB=0","Publicidad:SecciónA=0",
                   "Publicidad:SecciónB=0"))
Linear hypothesis test
Hypothesis:
SecciónA = 0
SecciónB = 0
Publicidad:SecciónA = 0
Publicidad:SecciónB = 0
Model 1: restricted model
Model 2: Ventas ~ Publicidad * Sección
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     13 42.541
2      9  1.996  4    40.545 45.705 5.552e-06 ***
---
```

La anterior salida R se interpreta como se muestra en la siguiente tabla:

Tabla 6.4: Resultados para prueba en numeral 6.

Sumas cuadrados de error			$H_0: \beta_2 = \beta_3 = \beta_{1,1} = \beta_{1,2} = 0$			
	d.f(SSE's)	SSE's	d.f(SSRparcial)	SSRparcial	F_0	$\Pr(f_{4,9} > F_0)$
Modelo nulo (MR)	13	42.541				
Modelo completo (MF)	9	1.996	4	40.55	45.71	0.0000
SSRparcial = SSE(MR) − SSE(MF); df(SSRparcial) = df[SSE(MR)] − df[SSE(MF)]						
F_0 = [SSRparcial ÷ 4] /MSE(MF)						

Nota 6.4.

- El modelo nulo en la anterior prueba es simplemente

$$Y_i = \beta_0 + \beta_1 X_i + E_i, \;\; E_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right)$$

- La diferencia entre sumas de cuadrados de residuos de los modelos nulo y completo corresponden en este caso, a la siguiente suma de cuadrados extras:

$$\text{SSRparcial} = \text{SSE(MR)-SSE(MF)} = \text{SSR}\left(I_1, I_2, X * I_1, X * I_2 | X\right)$$

7. *Determine si el cambio promedio en las ventas semanales por unidad de cambio en el presupuesto en publicidad es igual para las secciones de niños y de damas.* Responder a esta pregunta implica probar si las pendientes de las rectas en las secciones B y C son iguales:

$$\beta_1 + \beta_{1,2} = \beta_1 \iff \beta_{1,2} = 0$$

entonces basta hacer test t de la significancia marginal de $\beta_{1,2}$: $H_0 : \beta_{1,2} = 0$ vs. $H_1 : \beta_{1,2} \neq 0$. Ver Tabla 6.2.

8. *Asumiendo válidos los supuestos de los errores y considerando los resultados de las pruebas de significancia ¿puede reducirse modelo? ¿cuál sería ese modelo?* Escriba las ecuaciones de las rectas ajustadas en cada sección, según este nuevo modelo: Bajo supuestos sobre los errores, vemos posible reducción eliminando los parámetros no significativos (ver Tabla 6.2), y así se tendrá el siguiente modelo:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 I_{i1} + \beta_{1,2} X_i I_{i2} + E_i,\ E_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right)$$

Los resultados de su estimación se muestran en la siguiente salida R y en la tabla a continuación de ésta:

```
> MATRIZ.DISEÑOmodelo1=as.data.frame(model.matrix(modelo1))
> names(MATRIZ.DISEÑOmodelo1) #Observe cómo son nombradas las variables en modelo 1
[1] "(Intercept)"         "Publicidad"          "SecciónA"
[4] "SecciónB"            "Publicidad:SecciónA" "Publicidad:SecciónB"
> #AJUSTE MODELO SIN TÉRMINOS NO SIGNIFICATIVOS EN modelo1
> modelo3=lm(Ventas~Publicidad+SecciónA+Publicidad:SecciónB,
+             data=MATRIZ.DISEÑOmodelo1)
> summary(modelo3)
Call:
lm(formula = Ventas ~ Publicidad + SecciónA + Publicidad:SecciónB,
    data = MATRIZ.DISEÑOmodelo1)
Residuals:
     Min       1Q   Median       3Q      Max
-0.78729 -0.26009 -0.01237  0.30372  0.70110
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          7.84909    1.05310   7.453 1.27e-05 ***
Publicidad           1.03648    0.08877  11.676 1.54e-07 ***
SecciónA            -3.93247    0.48839  -8.052 6.14e-06 ***
Publicidad:SecciónB -0.49382    0.03624 -13.626 3.12e-08 ***
---
Residual standard error: 0.4619 on 11 degrees of freedom
Multiple R-squared:  0.9901,    Adjusted R-squared:  0.9874
F-statistic: 365.7 on 3 and 11 DF,  p-value: 2.698e-11
```

Tabla 6.5: Estimación del modelo enunciado en numeral 8, como referencia la Sección C

Parámetro	Estimación	Error Estándar	T_0	$P(\lvert t_9\rvert > \lvert T_0\rvert)$	L.I(2.5 %)	L.S(97.5 %)
β_0	7.85	1.05	7.45	0.00	5.53	10.17
β_1 (Publicidad)	1.04	0.09	11.68	0.00	0.84	1.23
β_2 (Secc. A)	-3.93	0.49	-8.05	0.00	-5.01	-2.86
$\beta_{1,2}$ (Publicidad*Secc. B)	-0.49	0.04	-13.63	0.00	-0.57	-0.41
$\widehat{Y}_i = 7.84909 + 1.03648 X_i - 3.93247 I_{i1} - 0.49382 X_i I_{i2}$						
$R^2 = 0.9901$, $R^2_{adj} = 0.9874$						
Resultados para test Anova del modelo: $F_0 = 365.7$, $P\left(f_{3,11} > F_0\right) = 2.698 \times 10^{-11}$						

De la ecuación del modelo se definen las ec. por secciones y sus ajustes:

Sección	Modelo	Ecuación ajustada
A $(I_1 = 1, I_2 = 0)$	$Y_i = (\beta_0 + \beta_2) + \beta_1 X_i + E_i,\ E_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$	$\widehat{Y}_i = 3.91662 + 1.03648 X_i$
B $(I_1 = 0, I_2 = 1)$	$Y_i = \beta_0 + (\beta_1 + \beta_{1,2}) X_i + E_i,\ E_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$	$\widehat{Y}_i = 7.84909 + 0.54266 X_i$
C $(I_1 = 0, I_2 = 0)$	$Y_i = \beta_0 + \beta_1 X_i + E_i,\ E_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$	$\widehat{Y}_i = 7.84909 + 1.03648 X_i$

9. *Usando como nivel de referencia la Sección C, postule y ajuste un modelo de regresión en donde se considere que en promedio el efecto de los gastos en publicidad sobre las ventas es el mismo para las tres secciones, pero la media de las ventas es diferente.* Escriba las ecuaciones de las rectas ajustadas para cada sección. Teniendo en cuenta los resultados previos ¿Será apropiado este modelo?

   La solución a esta pregunta es un modelo según el caso 2:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 I_{i1} + \beta_3 I_{i2} + E_i,\ E_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

Tabla 6.6: Tabla de parámetros estimados para modelo pedido en numeral 9, tomando como referencia la Sección C

Parámetro	Estimación	Error Estándar	T_0	$P(\lvert t_{11}\rvert > \lvert T_0\rvert)$	L.I(2.5 %)	L.S(97.5 %)
β_0	8.6888	1.4455	6.0109	0.0001	5.5072	11.8703
β_1 (Publicidad)	0.9635	0.1210	7.9657	0.0000	0.6973	1.2297
β_2 (Secc. A)	-4.2451	0.6671	-6.3634	0.0001	-5.7134	-2.7768
β_3 (Secc. B)	-4.8033	0.4714	-10.1901	0.0000	-5.8407	-3.7658
$\widehat{Y}_i = 8.6888 + 0.9635 X_i - 4.2451 I_{i1} - 4.8033 I_{i2}$						
$R^2 = 0.983$, $R^2_{adj} = 0.9784$						
Resultados para test Anova del modelo: $F_0 = 212$, $P(f_{3,11} > F_0) = 5.186 \times 10^{-10}$						

Sección	Modelo	Ecuación ajustada
A $(I_1 = 1, I_2 = 0)$	$Y_i = (\beta_0 + \beta_2) + \beta_1 X_i + E_i,\ E_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$	$\widehat{Y}_i = 4.4437 + 0.9635 X_i$
B $(I_1 = 0, I_2 = 1)$	$Y_i = (\beta_0 + \beta_3) + \beta_1 X_i + E_i,\ E_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$	$\widehat{Y}_i = 3.8855 + 0.9635 X_i$
C $(I_1 = 0, I_2 = 0)$	$Y_i = \beta_0 + \beta_1 X_i + E_i,\ E_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$	$\widehat{Y}_i = 8.6888 + 0.9635 X_i$

Nota 6.5. En R fue corrido el ajuste de este modelo de la siguiente forma:

```
> ###MODELO CON RECTAS DIFERENTES SÓLO EN EL INTERCEPTO###
> #No hay interacción entre predictor cuantitativo y la Sección
> modelo2=lm(Ventas~Publicidad+Sección)
```

Adicionalmente, considere la prueba de si el efecto de los gastos en publicidad sobre las ventas es el mismo para las tres secciones: ***Consideramos de nuevo el modelo inicial con rectas distintas tanto en pendiente como interceptos*** y en ese modelo el efecto medio de los gastos en publicidad sobre las ventas será igual entre las secciones, si

$$\beta_1 + \beta_{1,1} = \beta_1 + \beta_{1,2} = \beta_1 \iff \beta_{1,1} = \beta_{1,2} = 0$$

Luego, se debe realizar el siguiente test

$H_0 : \beta_{1,1} = \beta_{1,2} = 0$ vs.
H_1 : al menos uno de los parámetros $\beta_{1,1}, \beta_{1,2},$ es $\neq 0$
Los resultados R necesarios son los siguientes:

```
> linearHypothesis(modelo1,c("Publicidad:SecciónA=0","Publicidad:SecciónB=0"))
Linear hypothesis test
Hypothesis:
Publicidad:SecciónA = 0
Publicidad:SecciónB = 0
Model 1: restricted model
Model 2: Ventas ~ Publicidad * Sección

  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     11 4.0187
2      9 1.9960  2    2.0227 4.5601 0.04289 *
---
```

La anterior salida R se interpreta de la siguiente forma

Tabla 6.7: Resultados para prueba en numeral 6.

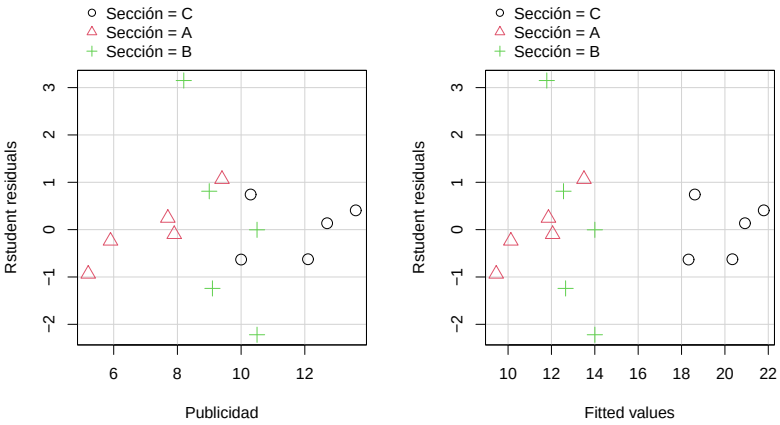
Sumas cuadrados de error			$H_0: \beta_{1,1} = \beta_{1,2} = 0$			
	d.f(SSE's)	SSE's	d.f(SSRparcial)	SSRparcial	F_0	$\Pr(f_{4,9} > F_0)$
Modelo nulo (MR)	11	4.0187				
Modelo completo (MF)	9	1.996	2	2.0227	4.5601	0.04289
SSRparcial = SSE(MR) − SSE(MF); df(SSRparcial) = df[SSE(MR)] − df[SSE(MF)]						
$F_0 = [\text{SSRparcial} \div 4] / \text{MSE(MF)}$						

De esta tabla se concluye que para al menos dos de las secciones hay diferencia entre el efecto medio que los gastos de publicidad tiene sobre las ventas, pues se rechaza H_0.

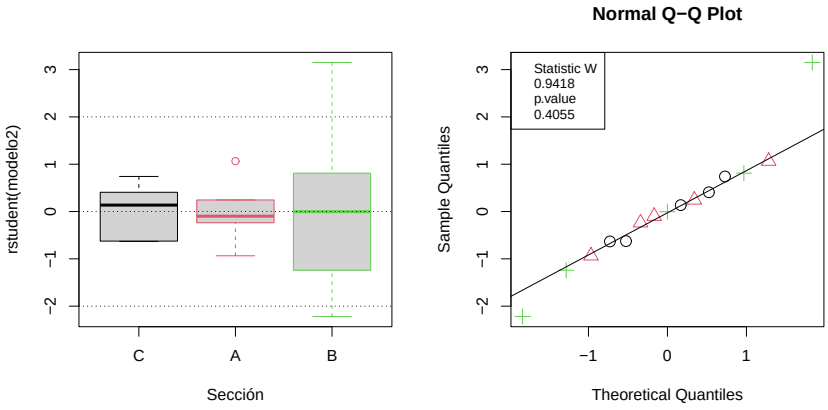
Nota 6.6. En este test, la diferencia entre sumas de cuadrados de residuos corresponde a la siguiente suma de cuadrados extras

$$\text{SSRparcial} = \text{SSE(MR)-SSE(MF)} = \text{SSR}\left(X * I_1, X * I_2 | X, I_1, I_2\right)$$

Considere también los residuos de ajuste y las gráficas de los ajustes del modelo 1: Relación lineal de ventas vs. gastos de publicidad difiere en cada sección (rectas no paralelas) y modelo 2: el efecto de los gastos en publicidad sobre las ventas es el mismo para las tres secciones pero la media de las ventas es diferente (rectas paralelas).

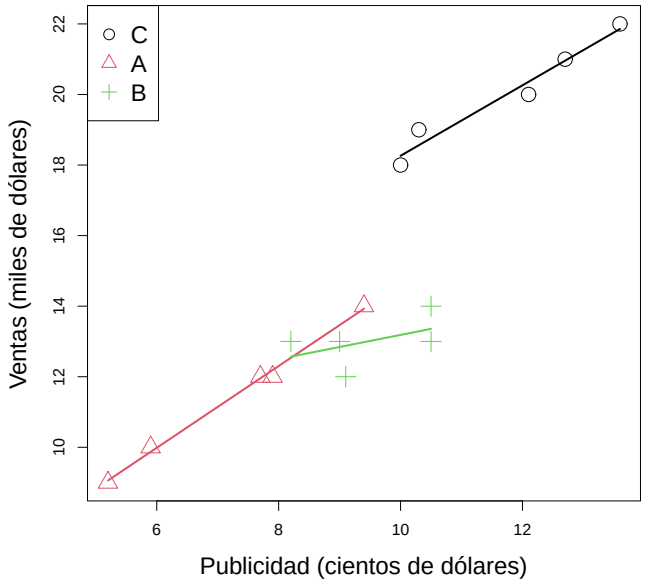


(a)

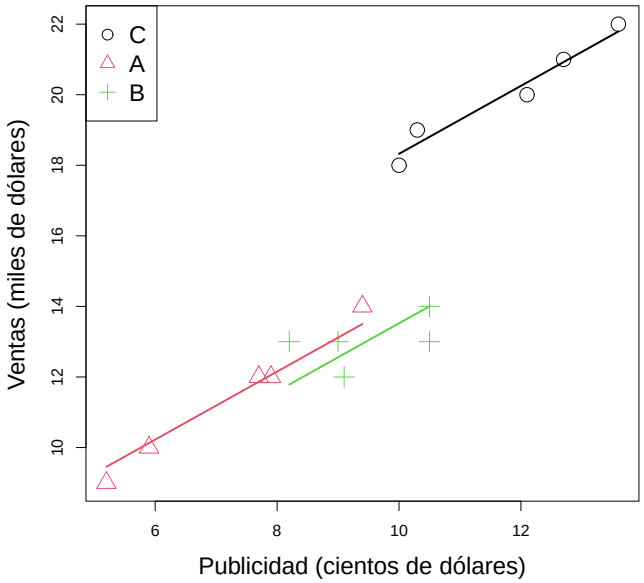


(b)

Figura 6.4: Gráficos para evaluación supuestos de los errores usando residuos estudentizados (externamente), modelo con rectas paralelas



(a)



(b)

Figura 6.5: Gráfico de dispersión con rectas ajustadas (a) en el modelo 1; (b) en el modelo 2

Código R 6.1. *A continuación, el código R completo usado*

```
library(car)
###LECTURA DE LOS DATOS###
datos=data.frame(Sección=factor(rep(c("A","B","C"),each=5)),
                 scan(what=list(Publicidad=0,Ventas=0)))
5.2 9
5.9 10
7.7 12
7.9 12
9.4 14
8.2 13
9.0 13
9.1 12
10.5 13
10.5 14
10.0 18
10.3 19
12.1 20
12.7 21
13.6 22

attach(datos)

#CAMBIANDO NIVEL DE REFERENCIA DE LA VARIABLE CUALITATIVA PARA LA SECCIÓN C###
Sección=relevel(Sección,ref="C")
Sección #observe que se informa que los niveles en su orden son C, A, B

#Gráfico de dispersión con identificación de Secciones
plot(Publicidad,Ventas,pch=as.numeric(Sección),col=as.numeric(Sección),
      xlab="Publicidad (cientos de dólares)",
      ylab="Ventas (miles de dólares)",cex=2,cex.lab=1.5)
legend("topleft",legend=c("C","A","B"),pch=c(1:3),col=c(1:3),cex=1.5)

##MODELO GRAL: RECTAS DIFERENTES. NIVEL DE REF=C###
modelo1=lm(Ventas~Publicidad*Sección)
summary(modelo1)
confint(modelo1)

#ANÁLISIS RESIDUALES EN modelo1
win.graph(width=8.5,height=5)
residualPlots(modelo1,groups=Sección,type="rstudent",linear=F,cex=1.5,pch=1:3,col=1:3)
win.graph(width=9,height=4.7)
layout(cbind(c(1),c(2)))
plot(rstudent(modelo1)~Sección,border=1:3)
abline(h=c(-2,0,2),lty=3)
```

```
test=shapiro.test(rstudent(modelo1))
qqnorm(rstudent(modelo1),pch=as.numeric(Sección),cex=1.5,col=as.numeric(Sección))
qqline(rstudent(modelo1))
legend("topleft",legend=rbind(c("Statistic W","p.value"),
       round(c(test$statistic,test$p.value),digits=4)),cex=0.8)

#PROBANDO HIPÓTESIS LINEALES PARA MODELO AJUSTADO EN modelo1
names(coef(modelo1)) #Observe el nombre de los términos en el modelo1

linearHypothesis(modelo1,"Publicidad:SecciónA=Publicidad:SecciónB")
linearHypothesis(modelo1,c("SecciónA=0","SecciónB=0","Publicidad:SecciónA=0",
                  "Publicidad:SecciónB=0"))
linearHypothesis(modelo1,c("Publicidad:SecciónA=0","Publicidad:SecciónB=0"))

###MODELO CON RECTAS DIFERENTES SÓLO EN EL INTERCEPTO###
#No hay interacción entre predictor cuantitativo y la Sección
modelo2=lm(Ventas~Publicidad+Sección)
summary(modelo2)
confint(modelo2)

#ANÁLISIS RESIDUALES EN modelo2: Rectas paralelas
win.graph(width=8.5,height=5)
residualPlots(modelo2,groups=Sección,type="rstudent",linear=F,cex=1.5,pch=1:3,col=1:3)
win.graph(width=9,height=4.7)
layout(cbind(c(1),c(2)))
plot(rstudent(modelo2)~Sección,border=1:3)
abline(h=c(-2,0,2),lty=3)
test2=shapiro.test(rstudent(modelo2))
qqnorm(rstudent(modelo2),pch=as.numeric(Sección),cex=1.5,col=as.numeric(Sección))
qqline(rstudent(modelo2))
legend("topleft",legend=rbind(c("Statistic W","p.value"),
        round(c(test2$statistic,test2$p.value),digits=4)),cex=0.8)

#GRÁFICOS DE DISPERSIÓN CON RECTAS AJUSTADAS EN MODELOS 1, 2
plot(Publicidad,Ventas,pch=as.numeric(Sección),col=as.numeric(Sección),
     xlab="Publicidad (cientos de dólares)",ylab="Ventas (miles de dólares)",
     cex=2,cex.lab=1.5)
legend("topleft",legend=c("C","A","B"),pch=c(1:3),col=c(1:3),cex=1.5)
lines(Publicidad[Sección=="A"],fitted(modelo1)[Sección=="A"],col=2,lwd=2)
lines(Publicidad[Sección=="B"],fitted(modelo1)[Sección=="B"],col=3,lwd=2)
lines(Publicidad[Sección=="C"],fitted(modelo1)[Sección=="C"],col=1,lwd=2)

win.graph()
plot(Publicidad,Ventas,pch=as.numeric(Sección),col=as.numeric(Sección),
     xlab="Publicidad (cientos de dólares)",ylab="Ventas (miles de dólares)",
     cex=2,cex.lab=1.5)
```

```
legend("topleft",legend=c("C","A","B"),pch=c(1:3),col=c(1:3),cex=1.5)
lines(Publicidad[Sección=="A"],fitted(modelo2)[Sección=="A"],col=2,lwd=2)
lines(Publicidad[Sección=="B"],fitted(modelo2)[Sección=="B"],col=3,lwd=2)
lines(Publicidad[Sección=="C"],fitted(modelo2)[Sección=="C"],col=1,lwd=2)

#AJUSTE MODELO SIN TÉRMINOS SIGNIFICATIVOS EN modelo1
MATRIZ.DISEÑOmodelo1=as.data.frame(model.matrix(modelo1))
names(MATRIZ.DISEÑOmodelo1) #Observe cómo quedan nombradas
                            #las variables del modelo en modelo1
modelo3=lm(Ventas~Publicidad+SecciónA+Publicidad:SecciónB,MATRIZ.DISEÑOmodelo1)
summary(modelo3)
confint(modelo3)

#Modelo reducido en test "Publicidad:SecciónA=Publicidad:SecciónB",
#o sea la igualdad de pendientes de las recta en secciones A y B
MR=lm(Ventas~Publicidad+SecciónA+SecciónB+Publicidad:I(SecciónA+SecciónB),
       data=MATRIZ.DISEÑOmodelo1)
summary(MR)

detach(datos)
```

Bibliografía

Fox, J. and Weisberg, S. (2019). *An R Companion to Applied Regression*, 3rd ed. Sage, Thousand Oaks CA.

Kutner, M. H., Nachtsheim, C. J., Neter, J. and Li. W. (2005). *Applied Linear Statistical Models*, 5th ed. McGraw-Hill Irwing, New York.

Montgomery, D. C., Peck, E. A. and Vining, G. G. (2012). *Introduction to Linear Regression Analysis*, 5th ed. Wiley, New Jersey.