

# Regresores cualitativos

## Contents

<b>1</b>	<b>Regresores cualitativos con dos niveles</b>	<b>1</b>
1.1	Variables auxiliares . . . . .	1
1.2	Factores . . . . .	2
1.3	Variables auxiliares 1 . . . . .	3
1.4	Factores 1 . . . . .	4
1.5	Modelo sin ordenada en el origen . . . . .	5
<b>2</b>	<b>Regresores cualitativos con más de dos niveles</b>	<b>6</b>
2.1	Variables auxiliares . . . . .	6
2.2	Factores . . . . .	7
<b>3</b>	<b>Modelo con más de un regresor cualitativo</b>	<b>10</b>
<b>4</b>	<b>Modelo con regresores cualitativos y cuantitativos</b>	<b>10</b>
4.1	Variables auxiliares . . . . .	10
4.2	Factores . . . . .	11
<b>5</b>	<b>Modelo con interacción entre regresores cuantitativos y cualitativos</b>	<b>12</b>
5.1	Variables auxiliares . . . . .	12
5.2	Factores . . . . .	13

## 1 Regresores cualitativos con dos niveles

Las variables cualitativas se representan en R con *factores*. En este caso hay dos variables cualitativas, *mom\_hs* y *mom\_work*. Como no son factores, se van a convertir a factor:

```
d = read.csv("datos/kidiq.csv")
str(d)

## 'data.frame': 434 obs. of 5 variables:
## $ kid_score: int 65 98 85 83 115 98 69 106 102 95 ...
## $ mom_hs : int 1 1 1 1 1 0 1 1 1 1 ...
## $ mom_iq : num 121.1 89.4 115.4 99.4 92.7 ...
## $ mom_work : int 4 4 4 3 4 1 4 3 1 1 ...
## $ mom_age : int 27 25 27 25 27 18 20 23 24 19 ...

d$mom_hs = factor(d$mom_hs, labels = c("no", "si"))
d$mom_work = factor(d$mom_work, labels = c("notrabaja", "trabaja23", "trabaja1_parcial", "trabaja1_comp"))
```

### 1.1 Variables auxiliares

La primera opción para incluir regresores cualitativos en el modelo es crear variables auxiliares con valores cero - uno. En este caso se crea la variable auxiliar *secundaria\_si*:

- *secundaria\_si* = 1, si la madre ha terminado secundaria (*mom\_hs* = si)
- *secundaria\_si* = 0, si la madre no ha terminado secundaria (*mom\_hs* = no)

```
secundaria_si = ifelse(d$mom_hs == "si", 1, 0)
```

El modelo estadístico que vamos a estimar es:

$$kid\_score_i = b_0 + b_1 secundaria\_si_i + e_i$$

```
m = lm(kid_score ~ secundaria_si, data = d)
coef(m)
```

```
## (Intercept) secundaria_si
##      77.54839      11.77126
```

En el fondo tenemos dos modelos, uno para las madres que han terminado secundaria y otro para los que no han terminado:

- Madres sin secundaria terminada (variable secundaria\_si = 0): El modelo correspondiente es

$$kid\_score_i = b_0 + e_i, \quad i \in 1, 2, \dots, n_0$$

donde  $n_0$  es el numero de madres sin secundaria. Si sumamos en ambos lados del modelo se tiene:

$$\sum_{i=1}^{n_0} kid\_score_i = \sum_{i=1}^{n_0} b_0 + \sum_{i=1}^{n_0} e_i = n_0 b_0 \Rightarrow b_0 = \frac{\sum_{i=1}^{n_0} kid\_score_i}{n_0}$$

Es decir, que  $b_0$  representa la puntuación media de los chicos cuya madre no ha terminado secundaria, 77.5483871

```
mean(d$kid_score[d$mom_hs == "no"])
```

```
## [1] 77.54839
```

- Madres con secundaria terminada (variable secundaria\_si = 1): el modelo correspondiente es

$$kid\_score_i = b_0 + b_1 + e_i, \quad i \in 1, 2, \dots, n_1$$

donde  $n_1$  es el numero de madres con secundaria. Si sumamos en ambos lados del modelo se tiene:

$$\sum_{i=1}^{n_1} kid\_score_i = \sum_{i=1}^{n_1} (b_0 + b_1 + e_i) = n_1 (b_0 + b_1) \Rightarrow b_0 + b_1 = \frac{\sum_{i=1}^{n_1} kid\_score_i}{n_1}$$

Es decir, que  $b_0 + b_1$  representa la puntuación media de los chicos cuya madre ha terminado secundaria, 89.3196481. Por tanto,  $b_1$  representa la diferencia entre las puntuaciones medias.

```
mean(d$kid_score[d$mom_hs == "si"])
```

```
## [1] 89.31965
```

## 1.2 Factores

Una manera más elegante de estimar estos modelos en R es utilizar directamente los factores en la formula de `lm()`:

```
m = lm(kid_score ~ mom_hs, data = d)
summary(m)
```

```
##
## Call:
## lm(formula = kid_score ~ mom_hs, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -57.55 -13.32   2.68  14.68  58.45
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   77.548      2.059   37.670 < 2e-16 ***
## mom_hssi      11.771      2.322    5.069 5.96e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.85 on 432 degrees of freedom
## Multiple R-squared:  0.05613,    Adjusted R-squared:  0.05394
## F-statistic: 25.69 on 1 and 432 DF,  p-value: 5.957e-07
```

Internamente, R ha creado la variable auxiliar *mom\_hssi*, que toma los valores

- *mom\_hssi* = 1 si *mom\_hs* = si
- *mom\_hssi* = 0 si *mom\_hs* = no.

R asigna los valores 0 y 1 en función de los niveles del factor:

```
levels(d$mom_hs)
```

```
## [1] "no" "si"
```

```
contrasts(d$mom_hs)
```

```
##      si
## no    0
## si    1
```

### 1.3 Variables auxiliares 1

También se podía haber creado la variable auxiliar *secundaria\_no*:

- *secundaria\_no* = 0, si la madre ha terminado secundaria (*mom\_iq* = si)
- *secundaria\_no* = 1, si la madre no ha terminado secundaria (*mom\_iq* = no)

```
secundaria_no = ifelse(d$mom_hs == "no", 1, 0)
```

El modelo estadístico que vamos a estimar ahora es:

$$kid\_score_i = b_0 + b_1 secundaria\_no_i + e_i$$

```
m = lm(kid_score ~ secundaria_no, data = d)
coef(m)
```

```
##      (Intercept)  secundaria_no
##      89.31965    -11.77126
```

Los dos modelos que tenemos ahora son:

- Madres con secundaria terminada (variable *secundaria\_no* = 0): El modelo correspondiente es

$$kid\_score_i = b_0 + e_i$$

Razonando igual que antes tenemos que  $b_0$  representa la puntuación media de los chicos cuya madre ha terminado secundaria, 89.3196481. Como vemos el valor coincide con lo obtenido antes.

- Madres sin secundaria terminada (variable `secundaria_no = 1`): el modelo correspondiente es

$$kid\_score_i = b_0 + b_1 + e_i$$

Por tanto,  $b_0 + b_1$  representa la puntuación media de los chicos cuya madre no ha terminado secundaria. Sumando se obtiene

```
coef(m)[1] + coef(m)[2]
```

```
## (Intercept)
##      77.54839
```

En este caso,  $b_1$  sigue representando la diferencia entre las puntuaciones medias.

## 1.4 Factores 1

Este nuevo modelo se introduce en `lm()` cambiando el nivel de referencia de la variable factor. Los niveles que tiene actualmente la variable son

```
levels(d$mom_hs)
```

```
## [1] "no" "si"
```

El nivel de referencia es “no”. Los valores que R asigna internamente a cada nivel son

```
contrasts(d$mom_hs)
```

```
##      si
## no    0
## si    1
```

Cambiamos el nivel de referencia:

```
d$mom_hs = relevel(d$mom_hs, ref = "si")
levels(d$mom_hs)
```

```
## [1] "si" "no"
```

Por tanto, los valores que asigna R a los distintos niveles son

```
contrasts(d$mom_hs)
```

```
##      no
## si    0
## no    1
```

Ahora se puede aplicar la función `lm()`:

```
m = lm(kid_score ~ mom_hs, data = d)
summary(m)
```

```
##
## Call:
## lm(formula = kid_score ~ mom_hs, data = d)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -57.55 -13.32   2.68  14.68  58.45
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   89.320      1.075   83.082 < 2e-16 ***
## mom_hsno     -11.771      2.322   -5.069 5.96e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.85 on 432 degrees of freedom
## Multiple R-squared:  0.05613,    Adjusted R-squared:  0.05394
## F-statistic: 25.69 on 1 and 432 DF,  p-value: 5.957e-07
```

Vemos que ahora R ha creado la variable auxiliar *mom\_hsno*, que toma los valores

- *mom\_hsno* = 0 si *mom\_hs* = si
- *mom\_hsno* = 1 si *mom\_hs* = no.

## 1.5 Modelo sin ordenada en el origen

Una tercera opción es utilizar el modelo sin ordenada en el origen:

$$kid\_score = b_1 secundaria\_si + b_2 secundaria\_no + e$$

en el que se utilizan las dos variables auxiliares pero se elimina el parámetro  $b_0$ . Los modelos ahora son:

- madre que si ha terminado secundaria: *secundaria\_si* = 1, *secundaria\_no* = 0

$$kid\_score = b_1 + e$$

- madre que no ha terminado secundaria: *secundaria\_si* = 0, *secundaria\_no* = 1

$$kid\_score = b_2 + e$$

Luego  $b_1$  representa la puntuación media de los chicos cuya madre ha terminado secundaria y  $b_2$  representa la puntuación media de los chicos cuya madre NO ha terminado secundaria.

```
m = lm(kid_score ~ 0 + secundaria_si + secundaria_no, data = d)
summary(m)
```

```
##
## Call:
## lm(formula = kid_score ~ 0 + secundaria_si + secundaria_no, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -57.55 -13.32   2.68  14.68  58.45
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## secundaria_si   89.320      1.075   83.08 <2e-16 ***
## secundaria_no   77.548      2.059   37.67 <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.85 on 432 degrees of freedom
## Multiple R-squared:  0.9506, Adjusted R-squared:  0.9504
## F-statistic: 4161 on 2 and 432 DF,  p-value: < 2.2e-16
```

Con factores:

```
m = lm(kid_score ~ 0 + mom_hs, data = d)
summary(m)
```

```
##
## Call:
## lm(formula = kid_score ~ 0 + mom_hs, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -57.55 -13.32   2.68  14.68  58.45
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## mom_hssi      89.320      1.075   83.08  <2e-16 ***
## mom_h sno     77.548      2.059   37.67  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.85 on 432 degrees of freedom
## Multiple R-squared:  0.9506, Adjusted R-squared:  0.9504
## F-statistic: 4161 on 2 and 432 DF,  p-value: < 2.2e-16
```

## 2 Regresores cualitativos con más de dos niveles

### 2.1 Variables auxiliares

En el caso de tener regresores cualitativos con más de dos niveles:

```
levels(d$mom_work)
```

```
## [1] "notrabaja"      "trabaja23"      "trabaja1_parcial" "trabaja1_completo"
```

Definimos las variables auxiliares:

- notrabaja\_si = 1 si mom\_work = notrabaja
- trabaja23\_si = 1 si mom\_work = trabaja23
- trabaja1\_parcial\_si = 1 si mom\_work = trabaja1\_parcial
- trabaja1\_completo\_si = 1 si mom\_work = trabaja1\_completo

```
notrabaja_si = ifelse(d$mom_work == "notrabaja", 1, 0)
trabaja23_si = ifelse(d$mom_work == "trabaja23", 1, 0)
trabaja1_parcial_si = ifelse(d$mom_work == "trabaja1_parcial", 1, 0)
trabaja1_completo_si = ifelse(d$mom_work == "trabaja1_completo", 1, 0)
```

Como la variable cualitativa tiene **cuatro niveles**, con **tres variables auxiliares** representamos todos los casos. El modelo general es:

$$kid\_score = b_0 + b_1trabaja23\_si + b_2trabaja1\_parcial\_si + b_3trabaja1\_completo\_si + e$$

- El modelo para las madres que no han trabajado es

$$kid\_score = b_0 + e$$

ya que en este caso trabaja23\_si = 0, trabaja1\_parcial\_si = 0 y trabaja1\_completo\_si = 0. Por tanto, la puntuación media de los chicos cuya madre no han trabajado es  $b_0$ .

- El modelo para las madres que trabajaron el segundo o tercer año es:

$$kid\_score = b_0 + b_1 + e$$

por tanto  $b_1$  representa las diferencias entre la puntuación media de los chicos cuya madre no trabaja y los de las madres que trabajaron el segundo o tercer año.

- El modelo para las madres que trabajaron el primer año a tiempo parcial es:

$$kid\_score = b_0 + b_2 + e$$

- Por último, el modelo para las madres que trabajaron el primer año a tiempo completo es:

$$kid\_score = b_0 + b_3 + e$$

En R:

```
m = lm(kid_score ~ trabaja23_si + trabaja1_parcial_si + trabaja1_completo_si, data = d)
summary(m)
```

```
##
## Call:
## lm(formula = kid_score ~ trabaja23_si + trabaja1_parcial_si +
##     trabaja1_completo_si, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -65.85 -12.85   2.79  14.15  50.50
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      82.000     2.305  35.568  <2e-16 ***
## trabaja23_si       3.854     3.095   1.245  0.2137
## trabaja1_parcial_si 11.500     3.553   3.237  0.0013 **
## trabaja1_completo_si  5.210     2.704   1.927  0.0547 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.23 on 430 degrees of freedom
## Multiple R-squared:  0.02444,    Adjusted R-squared:  0.01763
## F-statistic:  3.59 on 3 and 430 DF,  p-value: 0.01377
```

## 2.2 Factores

Utilizando factores se obtienen los mismos resultados:

```
m = lm(kid_score ~ mom_work, data = d)
summary(m)
```

```
##
## Call:
## lm(formula = kid_score ~ mom_work, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -65.85 -12.85   2.79  14.15  50.50
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      82.000      2.305  35.568 <2e-16 ***
## mom_worktrabaja23      3.854      3.095   1.245  0.2137
## mom_worktrabaja1_parcial 11.500      3.553   3.237  0.0013 **
## mom_worktrabaja1_completo  5.210      2.704   1.927  0.0547 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.23 on 430 degrees of freedom
## Multiple R-squared:  0.02444,    Adjusted R-squared:  0.01763
## F-statistic:  3.59 on 3 and 430 DF,  p-value: 0.01377
```

Comprobamos que internamente R crea variables auxiliares según los valores:

```
levels(d$mom_work)
```

```
## [1] "notrabaja"      "trabaja23"      "trabaja1_parcial" "trabaja1_completo"
```

```
contrasts(d$mom_work)
```

```
##              trabaja23 trabaja1_parcial trabaja1_completo
## notrabaja              0              0              0
## trabaja23              1              0              0
## trabaja1_parcial       0              1              0
## trabaja1_completo      0              0              1
```

Podemos hacer otras comparaciones cambiando la variable de referencia:

```
d$mom_work = releval(d$mom_work, ref="trabaja1_parcial")
```

```
levels(d$mom_work)
```

```
## [1] "trabaja1_parcial" "notrabaja"      "trabaja23"      "trabaja1_completo"
```

```
m = lm(kid_score ~ mom_work, data = d)
summary(m)
```

```
##
## Call:
## lm(formula = kid_score ~ mom_work, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -65.85 -12.85   2.79  14.15  50.50
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      93.500      2.703  34.587 <2e-16 ***
## mom_worknotrabaja -11.500      3.553  -3.237  0.0013 **
## mom_worktrabaja23  -7.646      3.402  -2.248  0.0251 *
```



```
## mom_worktrabaja1_completo    -6.290      3.050  -2.062   0.0398 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.23 on 430 degrees of freedom
## Multiple R-squared:  0.02444,    Adjusted R-squared:  0.01763
## F-statistic:  3.59 on 3 and 430 DF,  p-value: 0.01377
```

Como observamos, el nivel de referencia, que en este caso es “trabaja1\_parcial”, no aparece explícitamente en el modelo. Efectivamente, el modelo sería:

$$kid\_score = b_0 + b_1notrabaja\_si + b_2trabaja23\_si + b_3trabaja1\_completo\_si + e$$

El caso de la variable trabaja1\_parcial aparece cuando el resto de variables toma el valor cero. En ese caso el modelo sería:

$$kid\_score = b_0 + e$$

Además de cambiar el nivel de referencia, también se podría reordenar los niveles de la variable factor:

```
d$mom_work1 = factor(d$mom_work, levels=c("trabaja1_completo", "trabaja23", "notrabaja", "trabaja1_parcial"))
levels(d$mom_work1)
```

```
## [1] "trabaja1_completo" "trabaja23"          "notrabaja"          "trabaja1_parcial"
m = lm(kid_score ~ mom_work1, data = d)
summary(m)
```

```
##
## Call:
## lm(formula = kid_score ~ mom_work1, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -65.85  -12.85    2.79   14.15   50.50
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      87.210      1.413   61.723  <2e-16 ***
## mom_work1trabaja23    -1.356      2.502   -0.542   0.5882
## mom_work1notrabaja    -5.210      2.704   -1.927   0.0547 .
## mom_work1trabaja1_parcial  6.290      3.050    2.062   0.0398 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.23 on 430 degrees of freedom
## Multiple R-squared:  0.02444,    Adjusted R-squared:  0.01763
## F-statistic:  3.59 on 3 and 430 DF,  p-value: 0.01377
```

Como vemos de nuevo, el nivel de referencia no aparece explícitamente.

### 3 Modelo con más de un regresor cualitativo

## 4 Modelo con regresores cualitativos y cuantitativos

### 4.1 Variables auxiliares

Lo más frecuente es contar con regresores cualitativos y cuantitativos de manera simultánea. Por ejemplo, vamos a introducir en el modelo el regresor *mom\_iq* que es cuantitativo, y el regresor *mom\_hs* que es cualitativo. Para este último ya tenemos definida la variable auxiliar:

- *secundaria\_si* = 1, si la madre ha terminado secundaria (*mom\_iq* = si)
- *secundaria\_si* = 0, si la madre no ha terminado secundaria (*mom\_iq* = no)

El modelo que vamos a analizar es

$$kid\_score = b_0 + b_1 mom\_iq + b_2 secundaria\_si + e$$

Por tanto:

- si *secundaria\_si* = 0:  $kid\_score = b_0 + b_1 mom\_iq + e$
- si *secundaria\_si* = 1:  $kid\_score = (b_0 + b_2) + b_1 mom\_iq + e$

Tenemos dos rectas, **con la misma pendiente** y distinta  $b_0$ . En R:

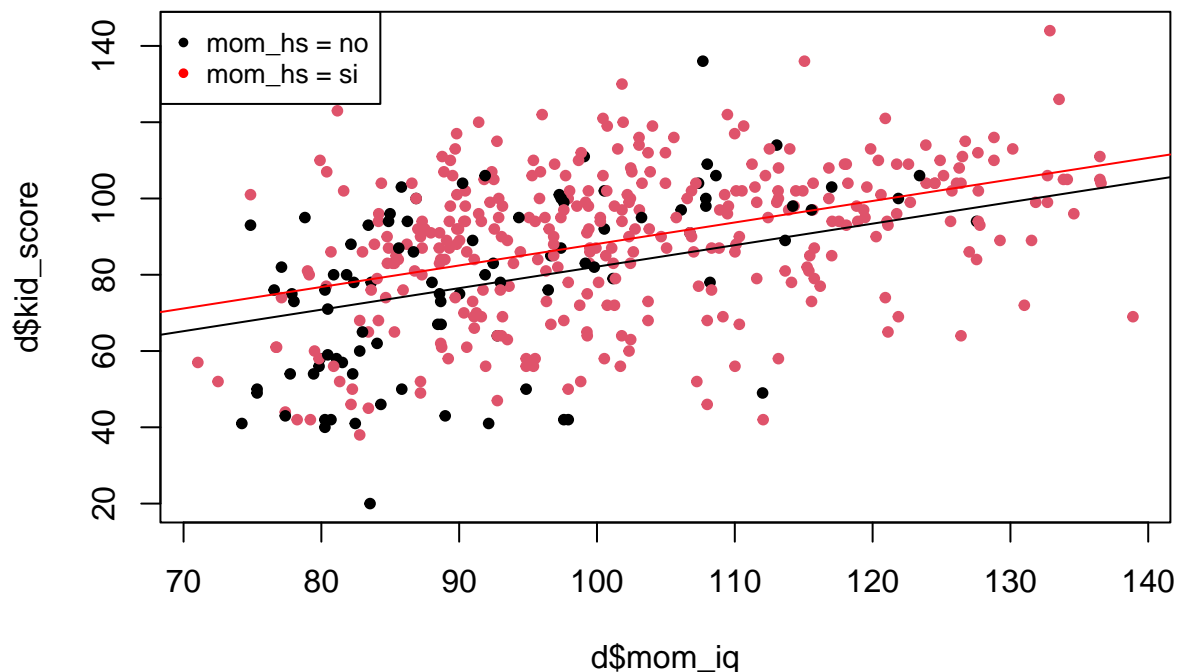
```
m = lm(kid_score ~ mom_iq + secundaria_si, data = d)
summary(m)
```

```
##
## Call:
## lm(formula = kid_score ~ mom_iq + secundaria_si, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -52.873 -12.663   2.404  11.356  49.545
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  25.73154    5.87521   4.380 1.49e-05 ***
## mom_iq        0.56391    0.06057   9.309 < 2e-16 ***
## secundaria_si  5.95012    2.21181   2.690  0.00742 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.14 on 431 degrees of freedom
## Multiple R-squared:  0.2141, Adjusted R-squared:  0.2105
## F-statistic: 58.72 on 2 and 431 DF, p-value: < 2.2e-16
```

Gráficamente:

```
# Para estar seguro del nivel de referencia:
d$mom_hs = releval(d$mom_hs, ref="no")
```

```
plot(d$mom_iq, d$kid_score, col = d$mom_hs, pch = 20)
abline(a = m$coefficients["(Intercept)"], b = m$coefficients["mom_iq"], col = "black")
abline(a = m$coefficients["(Intercept)"] + m$coefficients["secundaria_si"],
      b = m$coefficients["mom_iq"], col = "red")
legend("topleft", legend = c("mom_hs = no", "mom_hs = si"), col = c("black", "red"), pch = 20, cex = 0.8)
```



Si llamamos  $\text{dif\_100}$  a la diferencia entre la puntuación media de un chico cuya madre tiene  $\text{mom\_iq} = 100$  y no ha terminado secundaria y la puntuación media de un chico cuya madre tiene  $\text{mom\_iq} = 100$  y si ha terminado secundaria; y  $\text{dif\_120}$  a la diferencia entre la puntuación media de un chico cuya madre tiene  $\text{mom\_iq} = 120$  y no ha terminado secundaria y la puntuación media de un chico cuya madre tiene  $\text{mom\_iq} = 120$  y si ha terminado secundaria. Entonces,  $\text{dif\_100} = \text{dif\_120} = b_2$ .

## 4.2 Factores

Si utilizamos directamente los factores en el modelo, R automáticamente crea las variables auxiliares necesarias:

```
m = lm(kid_score ~ mom_iq + mom_hs, data = d)
summary(m)
```

```
##
## Call:
## lm(formula = kid_score ~ mom_iq + mom_hs, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -52.873 -12.663   2.404  11.356  49.545
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  25.73154    5.87521   4.380 1.49e-05 ***
## mom_iq        0.56391    0.06057   9.309 < 2e-16 ***
## mom_hssi      5.95012    2.21181   2.690 0.00742 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.14 on 431 degrees of freedom
## Multiple R-squared:  0.2141, Adjusted R-squared:  0.2105
## F-statistic: 58.72 on 2 and 431 DF, p-value: < 2.2e-16
```

El resultado, como no podía ser de otra manera, es el mismo.

## 5 Modelo con interacción entre regresores cuantitativos y cualitativos

### 5.1 Variables auxiliares

En el modelo de la sección anterior se ha modelado el efecto de *mom\_iq* y *mom\_hs* por separado. Sin embargo es posible incluir la interacción de ambas variables, es decir: para las madres que SI terminaron secundaria como influye la variables *mom\_iq*, y para las madres que NO terminaron secundaria, como influye *mom\_iq*. El modelo se escribe así:

$$kid\_score = b_0 + b_1mom\_iq + b_2secundaria\_si + b_3secundaria\_si * mom\_iq + e$$

Como vemos, este modelo incluye dos submodelos:

- si la madre no ha terminado secundaria *secundaria\_si* = 0:  $kid\_score = b_0 + b_1mom\_iq + e$
- si la madre si ha terminado secundaria *secundaria\_si* = 1:  $kid\_score = (b_0 + b_2) + (b_1 + b_3)mom\_iq + e$

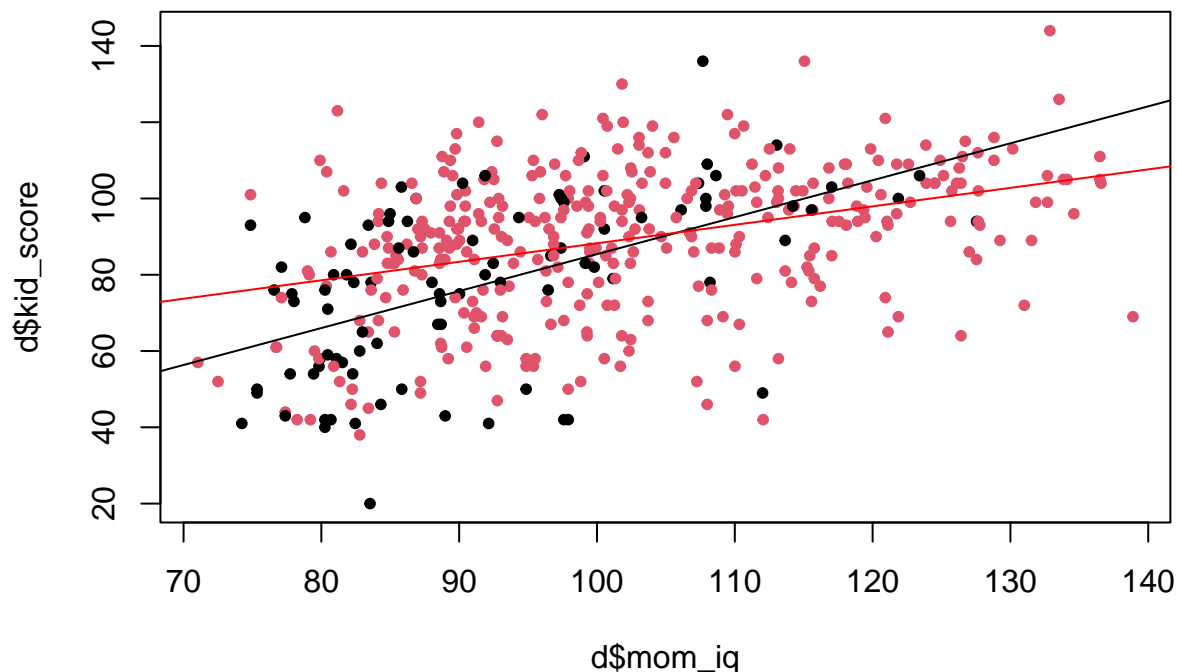
Luego tenemos dos modelos con ordenadas en el origen y pendiente diferentes. En R introducimos la interacción haciendo:

```
m = lm(kid_score ~ mom_iq + secundaria_si + I(mom_iq*secundaria_si), data = d)
summary(m)
```

```
##
## Call:
## lm(formula = kid_score ~ mom_iq + secundaria_si + I(mom_iq *
##      secundaria_si), data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -52.092 -11.332   2.066  11.663  43.880
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -11.4820    13.7580  -0.835  0.404422
## mom_iq           0.9689     0.1483   6.531 1.84e-10 ***
## secundaria_si    51.2682    15.3376   3.343 0.000902 ***
## I(mom_iq * secundaria_si) -0.4843     0.1622  -2.985 0.002994 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.97 on 430 degrees of freedom
## Multiple R-squared:  0.2301, Adjusted R-squared:  0.2247
## F-statistic: 42.84 on 3 and 430 DF,  p-value: < 2.2e-16
```

Gráficamente:

```
plot(d$mom_iq, d$kid_score, col = d$mom_hs, pch = 20)
abline(a = m$coefficients["(Intercept)"], b = m$coefficients["mom_iq"], col = "black")
abline(a = m$coefficients["(Intercept)"] + m$coefficients["secundaria_si"],
      b = m$coefficients["mom_iq"] + m$coefficients["I(mom_iq * secundaria_si)"], col = "red")
```



En este modelo, la diferencia entre puntuaciones medias de chicos no es constante como antes, depende simultáneamente del valor de mom\_iq de su madre y de si terminó o no la secundaria.

## 5.2 Factores

Con factores, la interacción entre variables se incluye con los dos puntos:

```
m = lm(kid_score ~ mom_iq + mom_hs + mom_iq:mom_hs, data = d)
summary(m)
```

```
##
## Call:
## lm(formula = kid_score ~ mom_iq + mom_hs + mom_iq:mom_hs, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -52.092 -11.332   2.066  11.663  43.880
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -11.4820    13.7580  -0.835  0.404422
## mom_iq         0.9689     0.1483   6.531 1.84e-10 ***
## mom_hssi       51.2682    15.3376   3.343 0.000902 ***
## mom_iq:mom_hssi -0.4843     0.1622  -2.985 0.002994 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.97 on 430 degrees of freedom
## Multiple R-squared:  0.2301, Adjusted R-squared:  0.2247
## F-statistic: 42.84 on 3 and 430 DF, p-value: < 2.2e-16
```

Otra alternativa es utilizar el signo de multiplicación, que incluye los regresores por separado y la interacción:

```

m = lm(kid_score ~ mom_iq * mom_hs, data = d)
summary(m)

##
## Call:
## lm(formula = kid_score ~ mom_iq * mom_hs, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -52.092 -11.332   2.066  11.663  43.880
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -11.4820    13.7580  -0.835  0.404422
## mom_iq         0.9689     0.1483   6.531 1.84e-10 ***
## mom_hssi       51.2682    15.3376   3.343 0.000902 ***
## mom_iq:mom_hssi -0.4843     0.1622  -2.985 0.002994 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.97 on 430 degrees of freedom
## Multiple R-squared:  0.2301, Adjusted R-squared:  0.2247
## F-statistic: 42.84 on 3 and 430 DF,  p-value: < 2.2e-16

```