

Regresores cualitativos

Contents

1	Regresores cualitativos con dos niveles	1
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
2	Regresores cualitativos con más de dos niveles	6
2.1	Variables auxiliares	6
2.2	Factores	8
3	Modelo con más de un regresor cualitativo	10
4	Modelo con interacción entre regresores cuantitativos y cualitativos	10
4.1	Variables auxiliares	10
4.2	Factores	11

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*, que ya son factores:

```
load("datos/kidiq.Rdata")
str(d)

## 'data.frame': 434 obs. of 5 variables:
## $ kid_score: int 65 98 85 83 115 98 69 106 102 95 ...
## $ mom_hs : Factor w/ 2 levels "no","si": 2 2 2 2 2 1 2 2 2 2 ...
## $ mom_iq : num 121.1 89.4 115.4 99.4 92.7 ...
## $ mom_work : Factor w/ 4 levels "notrabaja","trabaja23",...: 4 4 4 3 4 1 4 3 1 1 ...
## $ mom_age : int 27 25 27 25 27 18 20 23 24 19 ...
```

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:

$$kid_score_i = \beta_0 + \beta_1 mom_iq_i + \beta_2 secundaria_si_i + u_i$$

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

```
##      (Intercept)      mom_iq secundaria_si
##      25.731538      0.563906      5.950117
```

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

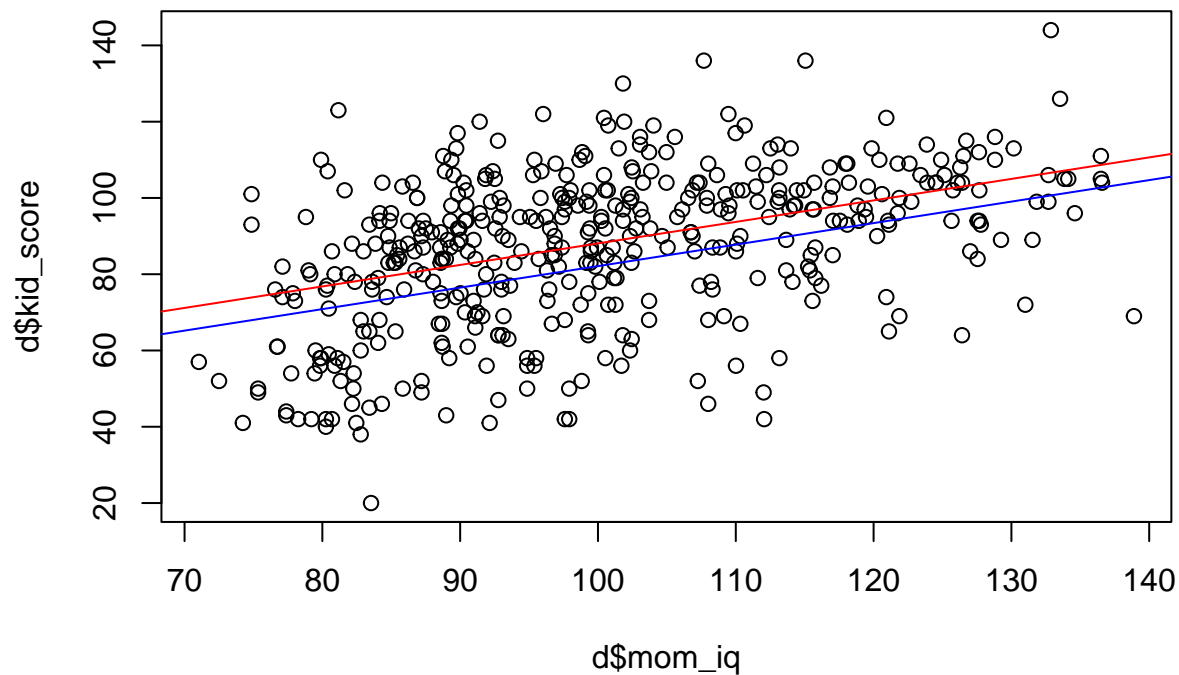
$$kid_score_i = \beta_0 + \beta_1 mom_iq_i + u_i$$

- Madres con secundaria terminada (variable `secundaria_si` = 1): el modelo correspondiente es

$$kid_score_i = (\beta_0 + \beta_2) + \beta_1 mom_iq_i + u_i$$

Es decir, son dos rectas paralelas de pendiente β_1 y separadas una distancia igual a β_2 .

```
plot(d$mom_iq, d$kid_score)
# modelo para secundaria_si = 0
abline(a = m$coef[1], b = m$coef[2], col = "blue")
# modelo para secundaria_si = 1
abline(a = m$coef[1] + m$coef[3], b = m$coef[2], col = "red")
```



1.2 Factores

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

```
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
```

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:

$$kid_score_i = \beta_0^* + \beta_1^* mom_iq_i + \beta_2^* secundaria_no_i + u_i$$

donde se ha enfatizado que, en principio, los parámetros son diferentes ya que las variables son diferentes. En R:

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

```
##      (Intercept)      mom_iq secundaria_no
##      31.681655      0.563906      -5.950117
```

Los dos modelos que tenemos ahora son:

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

$$kid_score_i = (\beta_0^* + \beta_2^*) + \beta_1^* mom_iq_i + u_i$$

- Madres con secundaria terminada (variable `secundaria_no = 0`): El modelo correspondiente es

$$kid_score_i = \beta_0^* + \beta_1^* mom_iq_i + u_i$$

Igualando los modelos creados con las dos variables auxiliares, se tiene que cumplir que:

$$\beta_1 = \beta_1^*$$

$$\beta_0 = \beta_0^* + \beta_2^*$$

$$\beta_0^* = \beta_0 + \beta_2$$

Eliminando β_0^* se tiene que

$$\beta_2 = -\beta_2^*$$

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$mom_hs1 = relevel(d$mom_hs, ref = "si")
levels(d$mom_hs1)
```

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

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

```
contrasts(d$mom_hs1)
```

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

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

```

m = lm(kid_score ~ mom_iq + mom_hs1, data = d)
summary(m)

##
## Call:
## lm(formula = kid_score ~ mom_iq + mom_hs1, 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) 31.68166    6.26881   5.054 6.41e-07 ***
## mom_iq       0.56391    0.06057   9.309 < 2e-16 ***
## mom_hs1no   -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

```

Vemos que ahora R ha creado la variable auxiliar *mom_h sno*, que toma los valores

- *mom_h sno* = 0 si *mom_hs* = si
- *mom_h sno* = 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:

$$kid_score = \beta_1 mom_iq + \beta_2 secundaria_si + \beta_3 secundaria_no + u$$

en el que se utilizan las dos variables auxiliares pero se elimina el parámetro β_0 . Los modelos ahora son:

- madre que si ha terminado secundaria: *secundaria_si* = 1, *secundaria_no* = 0

$$kid_score = \beta_2 + \beta_1 mom_iq + u$$

- madre que no ha terminado secundaria: *secundaria_si* = 0, *secundaria_no* = 1

$$kid_score = \beta_3 + \beta_1 mom_iq + u$$

De nuevo tenemos dos rectas paralelas con pendiente β_1 y separadas una distancia igual a $\beta_2 - \beta_3$.

```

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

```

```

##
## Call:
## lm(formula = kid_score ~ 0 + mom_iq + secundaria_si + secundaria_no,
##      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|)
## mom_iq          0.56391    0.06057   9.309 < 2e-16 ***
## secundaria_si  31.68166    6.26881   5.054 6.41e-07 ***
## secundaria_no  25.73154    5.87521   4.380 1.49e-05 ***
## ---
## 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:  0.9589, Adjusted R-squared:  0.9586
## F-statistic: 3353 on 3 and 431 DF,  p-value: < 2.2e-16
```

Con factores:

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

```
##
## Call:
## lm(formula = kid_score ~ 0 + 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|)
## mom_iq          0.56391    0.06057   9.309 < 2e-16 ***
## mom_hsno  25.73154    5.87521   4.380 1.49e-05 ***
## mom_hssi  31.68166    6.26881   5.054 6.41e-07 ***
## ---
## 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:  0.9589, Adjusted R-squared:  0.9586
## F-statistic: 3353 on 3 and 431 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" "trabaja1p" "trabaja1c"
```

Se definen las variables auxiliares:

```
notrabaja_si = ifelse(d$mom_work == "notrabaja", 1, 0)
trabaja23_si = ifelse(d$mom_work == "trabaja23", 1, 0)
trabaja1p_si = ifelse(d$mom_work == "trabaja1p", 1, 0)
trabaja1c_si = ifelse(d$mom_work == "trabaja1c", 1, 0)
```

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

$$kid_score = \beta_0 + \beta_1 mom_iq + \beta_2 trabaja23_si + \beta_3 trabaja1p_si + \beta_4 trabaja1c_si + u$$

- El modelo para las madres que no han trabajado es

$$kid_score = \beta_0 + \beta_1 mom_iq + u$$

ya que en este caso $trabaja23_si = 0$, $trabaja1p_si = 0$ y $trabaja1c_si = 0$.

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

$$kid_score = (\beta_0 + \beta_2) + \beta_1 mom_iq + u$$

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

$$kid_score = (\beta_0 + \beta_3) + \beta_1 mom_iq + u$$

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

$$kid_score = (\beta_0 + \beta_4) + \beta_1 mom_iq + u$$

En R:

```
m = lm(kid_score ~ mom_iq + trabaja23_si + trabaja1p_si + trabaja1c_si, data = d)
summary(m)
```

```
##
## Call:
## lm(formula = kid_score ~ mom_iq + trabaja23_si + trabaja1p_si +
##      trabaja1c_si, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -57.796 -12.103   1.892  12.019  50.582
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  24.14226    6.14276   3.930 9.89e-05 ***
## mom_iq       0.59478    0.05942  10.009 < 2e-16 ***
## trabaja23_si  3.97026    2.78980   1.423  0.1554
## trabaja1p_si  6.60140    3.23986   2.038  0.0422 *
## trabaja1c_si  3.06392    2.44682   1.252  0.2112
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.24 on 429 degrees of freedom
## Multiple R-squared:  0.2091, Adjusted R-squared:  0.2018
## F-statistic: 28.36 on 4 and 429 DF,  p-value: < 2.2e-16
```

- Es decir, tenemos cuatro rectas paralelas con pendiente β_1 .
- β_2 representa la distancia entre el modelo para las madres que trabajaron el segundo o tercer año y el modelo para las madres que no han trabajado (la referencia en este caso).

- β_3 representa la distancia entre el modelo para las madres que trabajaron el primer año a tiempo parcial y el modelo para las madres que no han trabajado.
- β_4 representa la distancia entre el modelo para las madres que trabajaron el primer año a tiempo completo y el modelo para las madres que no han trabajado.

2.2 Factores

Utilizando factores se obtienen los mismos resultados:

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

```
##
## Call:
## lm(formula = kid_score ~ mom_iq + mom_work, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -57.796 -12.103   1.892  12.019  50.582
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    24.14226     6.14276   3.930 9.89e-05 ***
## mom_iq          0.59478     0.05942  10.009 < 2e-16 ***
## mom_worktrabaja23 3.97026     2.78980   1.423  0.1554
## mom_worktrabaja1p 6.60140     3.23986   2.038  0.0422 *
## mom_worktrabaja1c 3.06392     2.44682   1.252  0.2112
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.24 on 429 degrees of freedom
## Multiple R-squared:  0.2091, Adjusted R-squared:  0.2018
## F-statistic: 28.36 on 4 and 429 DF,  p-value: < 2.2e-16
```

Comprobamos que internamente R crea variables auxiliares según los valores:

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

```
## [1] "notrabaja" "trabaja23" "trabaja1p" "trabaja1c"
```

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

```
##           trabaja23 trabaja1p trabaja1c
## notrabaja         0         0         0
## trabaja23         1         0         0
## trabaja1p         0         1         0
## trabaja1c         0         0         1
```

Podemos hacer otras comparaciones cambiando la variable de referencia:

```
d$mom_work = relevel(d$mom_work, ref="trabaja1p")
levels(d$mom_work)
```

```
## [1] "trabaja1p" "notrabaja" "trabaja23" "trabaja1c"
```

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

```
##
```

```
## Call:
## lm(formula = kid_score ~ mom_iq + mom_work, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -57.796 -12.103   1.892  12.019  50.582
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    30.74366     6.72687   4.570 6.38e-06 ***
## mom_iq          0.59478     0.05942  10.009 < 2e-16 ***
## mom_worknotrabaja -6.60140     3.23986  -2.038  0.0422 *
## mom_worktrabaja23 -2.63114     3.10699  -0.847  0.3976
## mom_worktrabaja1c -3.53747     2.76336  -1.280  0.2012
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.24 on 429 degrees of freedom
## Multiple R-squared:  0.2091, Adjusted R-squared:  0.2018
## F-statistic: 28.36 on 4 and 429 DF,  p-value: < 2.2e-16
```

Como observamos, el nivel de referencia, que en este caso es “trabaja1p”, no aparece explícitamente en el modelo. Efectivamente, el modelo sería:

$$kid_score = \beta_0^* + \beta_1^*mom_iq + \beta_2^*notrabaja_si + \beta_3^*trabaja23_si + \beta_4^*trabaja1c_si + u$$

El caso de la variable trabaja1p aparece cuando el resto de variables toma el valor cero. En ese caso el modelo sería:

$$kid_score = \beta_0^* + \beta_1^*mom_iq + u$$

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("trabaja1c", "trabaja23", "notrabaja", "trabaja1p"))
levels(d$mom_work1)
```

```
## [1] "trabaja1c" "trabaja23" "notrabaja" "trabaja1p"
```

```
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_work1trabaja1p   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 interacción entre regresores cuantitativos y cualitativos

4.1 Variables auxiliares

En los modelos anteriores se ha modelado las variables cuantitativas y cualitativas por separado, obteniendo rectas paralelas. También es posible incluir la interacción de ambas variables, es decir, el comportamiento de una variable influye en la otra variable. El modelo se escribe así:

$$kid_score = \beta_0 + \beta_1 mom_iq + \beta_2 secundaria_si + \beta_3 secundaria_si * mom_iq + u$$

Como vemos, este modelo incluye dos submodelos:

- si la madre no ha terminado secundaria $secundaria_si = 0$: $kid_score = \beta_0 + \beta_1 mom_iq + e$
- si la madre si ha terminado secundaria $secundaria_si = 1$: $kid_score = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) mom_iq + e$

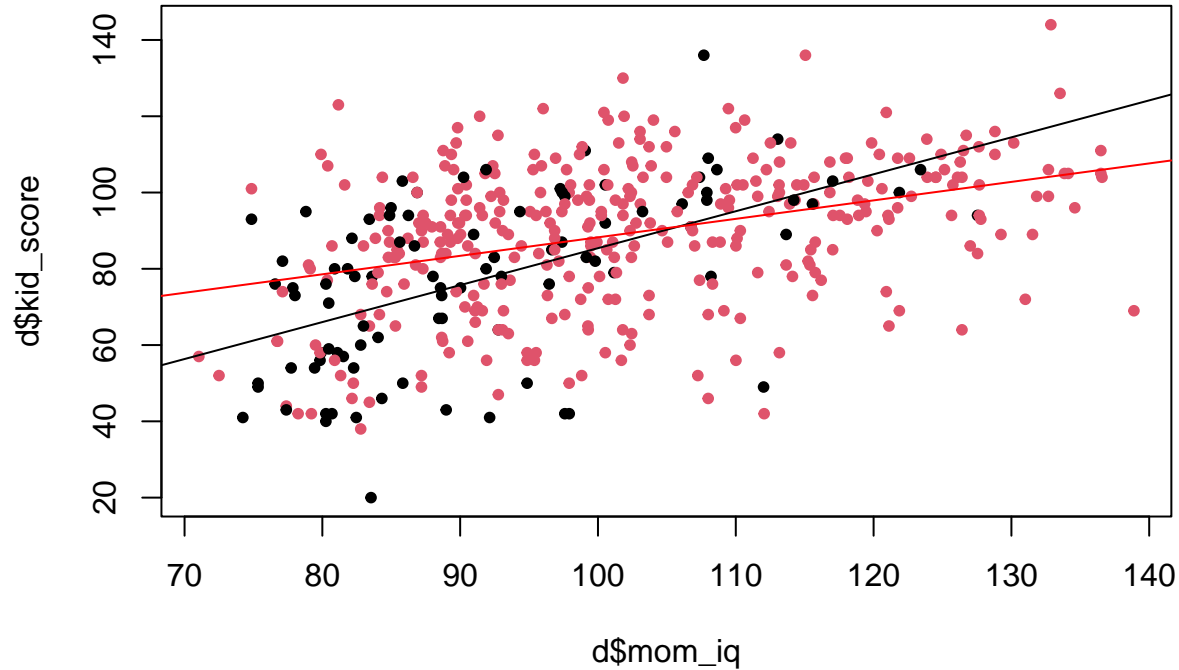
Luego tenemos dos modelos con ordenadas en el origen y pendiente diferentes. En R introducimos la interacción haciendo:

```
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.

4.2 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
```