

# Aplicaciones del modelo de regresión lineal: análisis de relaciones entre variables

## Contents

<b>1</b>	<b>Datos</b>	<b>1</b>
<b>2</b>	<b>Un regresor cualitativo</b>	<b>1</b>
<b>3</b>	<b>Un regresor cuantitativo</b>	<b>3</b>
<b>4</b>	<b>Logaritmos y porcentajes</b>	<b>4</b>
<b>5</b>	<b>Un regresor cualitativo y otro cuantitativo</b>	<b>5</b>
5.1	Sin interacción . . . . .	5
5.2	Con interacción . . . . .	7
<b>6</b>	<b>Dos regresores cuantitativos</b>	<b>8</b>

## 1 Datos

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

## 2 Un regresor cualitativo

Estimamos el modelo

$$kid\_score_i = \beta_0 + \beta_1 mom\_hssi_i + u_i$$

donde  $mom\_hssi$  es una variable auxiliar con valores 0,1:

- $mom\_hs = si \Rightarrow mom\_hssi = 1$
- $mom\_hs = no \Rightarrow mom\_hssi = 0$

```
m1 = lm(kid_score ~ mom_hs, data = d)
summary(m1)

##
## 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
```

Tenemos dos modelos

- $\text{mom\_hssi} = 0$ :

$$\text{kid\_score}_i = \beta_0 + u_i$$

Eliminamos el término  $u_i$  tomando esperanzas:

$$E[\text{kid\_score}_i] = \beta_0$$

Es decir,  $\beta_0$  es la media de las puntuaciones de los chicos cuyas madres no han terminado el bachillerato.

```
# lo comprobamos en R
mean(d$kid_score[d$mom_hs=="no"])
```

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

- $\text{mom\_hssi} = 1$ :

$$\text{kid\_score}_i = \beta_0 + \beta_1 + u_i$$

$$E[\text{kid\_score}_i] = \beta_0 + \beta_1$$

Luego  $\beta_1$  es la diferencia entre la media de las puntuaciones de los chicos cuya madre han terminado y las que no han terminado bachillerato.

```
# en R
mean(d$kid_score[d$mom_hs=="si"]) - mean(d$kid_score[d$mom_hs=="no"])
```

```
## [1] 11.77126
```

Estas conclusiones ya las obtuvimos en los primeros temas para los valores estimados con mínimos cuadrados. Sin embargo, ahora utilizamos modelos con parámetros y podemos utilizar la inferencia para comprobar si esa diferencia es fruto del azar o no. Por ejemplo, el contraste:

$$H_0 : \beta_1 = 0 \quad H_1 : \beta_1 \neq 0$$

Mirando el pvalor correspondiente, se rechaza  $H_0$ , luego los hijos de madres con bachillerato tienen una puntuación mayor que los hijos de madres sin bachillerato (una puntuación 11.77 puntos superior en promedio).

También lo podemos hacer con los intervalos de confianza:

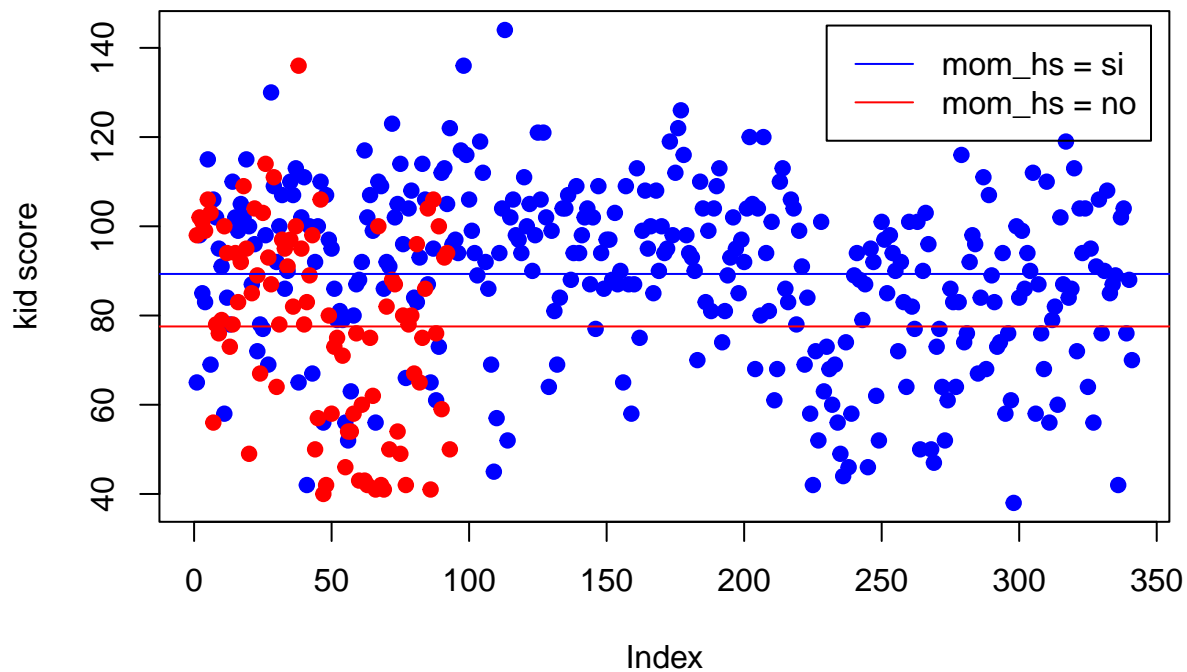
```
confint(m1)
```

```
##           2.5 %   97.5 %
## (Intercept) 73.502246 81.59453
## mom_hssi    7.206598 16.33592
```

El intervalo para  $\beta_1$  es el rango de valores posibles para dicho parámetro, y entre ellos no está el cero.

Gráficamente:

```
plot(d$kid_score[d$mom_hs=="si"], col = "blue", pch = 19, ylab = "kid score")
points(d$kid_score[d$mom_hs=="no"], col = "red", pch = 19)
abline(h=m1$coeff[1], col = "red")
abline(h=m1$coeff[1]+m1$coef[2], col = "blue")
legend(230,145, legend = c("mom_hs = si", "mom_hs = no"), col = c("blue", "red"), lty = c(1,1))
```



### 3 Un regresor cuantitativo

```
m2 = lm(kid_score ~ mom_iq, data = d)
summary(m2)
```

```
##
## Call:
## lm(formula = kid_score ~ mom_iq, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -56.753 -12.074   2.217  11.710  47.691
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  25.79978    5.91741   4.36 1.63e-05 ***
## mom_iq        0.60997    0.05852  10.42 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 18.27 on 432 degrees of freedom
## Multiple R-squared:  0.201, Adjusted R-squared:  0.1991
## F-statistic: 108.6 on 1 and 432 DF,  p-value: < 2.2e-16
```

- Interpretación de  $\beta_1$ : Se interpreta como el aumento de la puntuación media cuando incrementamos en una unidad el IQ de las madres. Efectivamente, sean la madre-hijo 1 y la madre-hijo 2. Los modelos para ambos son:

$$E[kid\_score_1] = \beta_0 + \beta_1 mom\_iq_1 \quad E[kid\_score_2] = \beta_0 + \beta_1 mom\_iq_2$$

Restando:

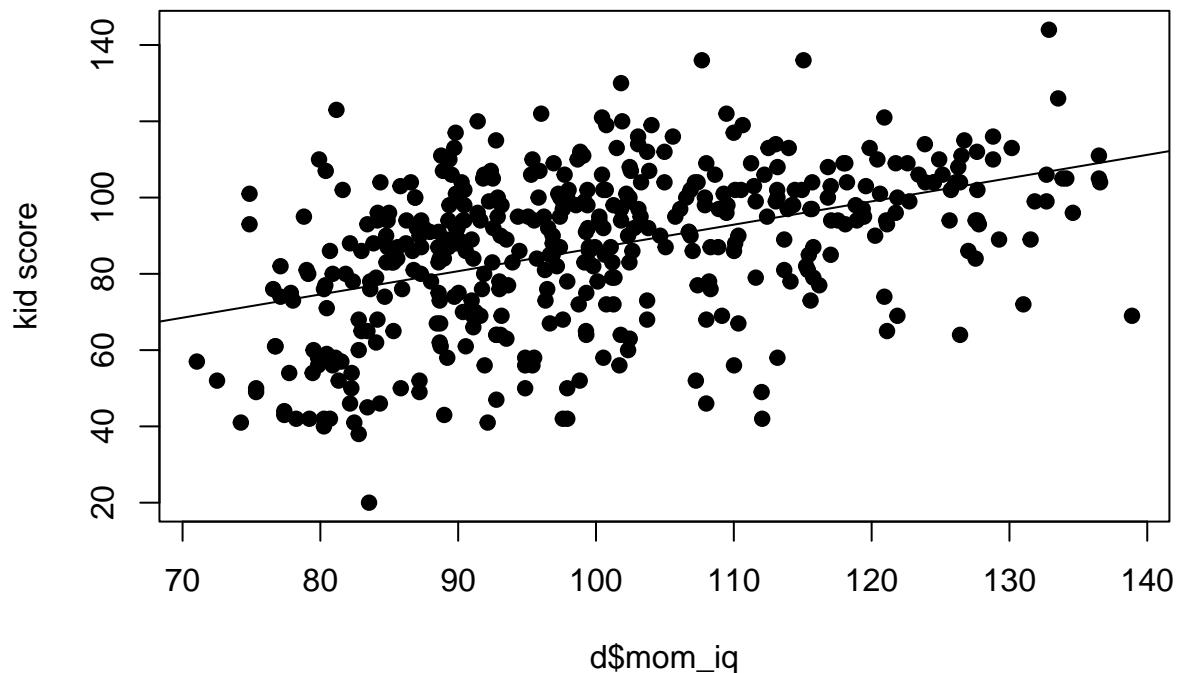
$$E[kid\_score_1] - E[kid\_score_2] = \beta_1(mom\_iq_1 - mom\_iq_2)$$

Luego si  $(mom\_iq_1 - mom\_iq_2 = 1)$ , entonces  $\beta_1 = E[kid\_score_1] - E[kid\_score_2]$ . El pvalor para  $\beta_1$  es muy pequeño, luego  $\beta_1$  es significativo.

- Interpretación de  $\beta_0$ : Se interpreta como la puntuación que obtendría un chico cuya madre tiene IQ=0. En este caso, no tiene mucho sentido interpretar este parámetro. Según el pvalor, es estadísticamente significativo.

Gráficamente:

```
plot(d$mom_iq, d$kid_score, pch = 19, ylab = "kid score")
abline(m2)
```



## 4 Logaritmos y porcentajes

Supongamos que tenemos el modelo:

$$\log(E[y_i]) = \beta_0 + \beta_1 x_i$$

Tomando diferenciales:

$$\frac{dE[y_i]}{E[y_i]} = \beta_1 dx_i \Rightarrow \frac{\Delta E[y_i]}{E[y_i]} \approx \beta_1 \Delta x_i$$

Es decir, un incremento de una unidad de x produce un incremento del  $\beta_1\%$  de y.

```
m3 = lm(log(kid_score) ~ mom_iq, data = d)
summary(m3)

##
## Call:
## lm(formula = log(kid_score) ~ mom_iq, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.30612 -0.13453  0.04982  0.16243  0.52888
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.6474338   0.0787286   46.33  <2e-16 ***
## mom_iq        0.0078342   0.0007786   10.06  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.243 on 432 degrees of freedom
## Multiple R-squared:  0.1899, Adjusted R-squared:  0.188
## F-statistic: 101.2 on 1 and 432 DF,  p-value: < 2.2e-16
```

Luego un incremento de 1 del IQ de las madres produce un incremento del 0.81% de la puntuación de los hijos.

## 5 Un regresor cualitativo y otro cuantitativo

### 5.1 Sin interacción

```
m4 = lm(kid_score ~ mom_iq + mom_hs, data = d)
summary(m4)

##
## 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 modelo es:

$$E[kid\_score_i] = \beta_0 + \beta_1 mom\_iq_i + \beta_2 mom\_hssi_i$$

Que en realidad son dos modelos con distinta ordenada en el origen y distinta pendiente:

- Si  $mom\_hssi = 0$ :

$$E[kid\_score_i] = \beta_0 + \beta_1 mom\_iq_i$$

- Si  $mom\_hssi = 1$ :

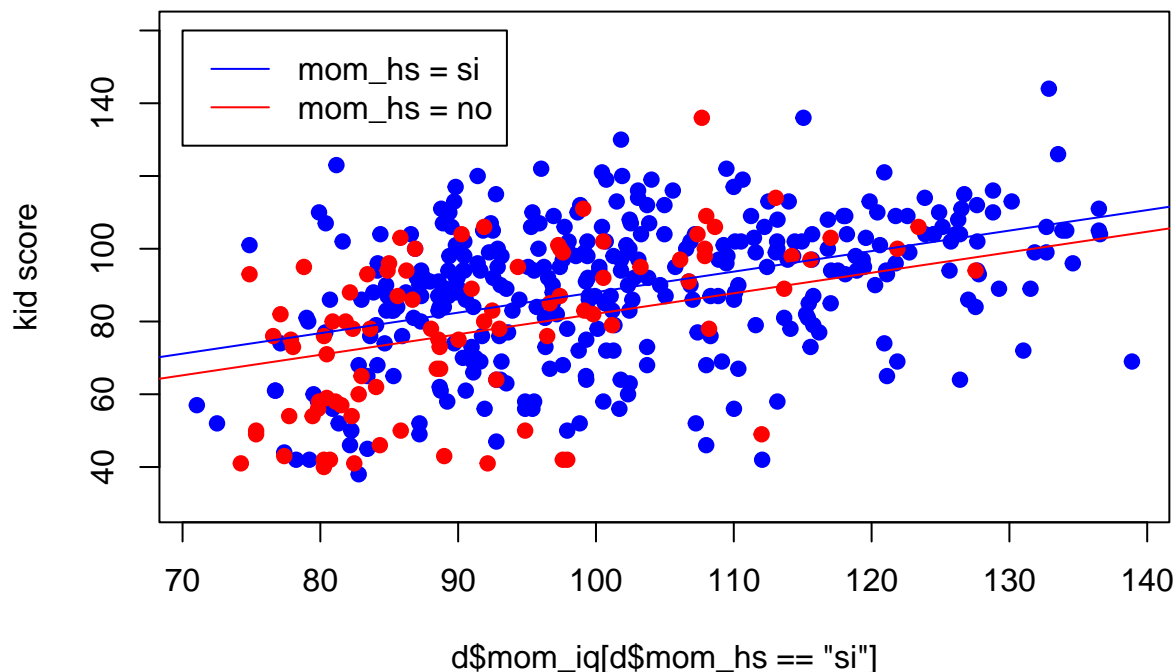
$$E[kid\_score_i] = (\beta_0 + \beta_2) + \beta_1 mom\_iq_i$$

Por tanto:

- $\beta_0$ : puntuación media de un chico cuya madre no ha terminado bachillerato y tiene un IQ=0
- $\beta_1$ : si comparamos chicos con el mismo valor de  $mom\_hssi$ , un incremento de un punto en  $mom\_iq$  conlleva un aumento medio de  $\beta_1$  para  $kid\_score$ . Ese incremento es significativo.
- $\beta_2$ : para dos madres con el mismo IQ, una terminó el bachillerato y la otra no, la puntuación media de los chicos se diferencia en 5.95. Esa diferencia es estadísticamente significativa.

Gráficamente:

```
plot(d$mom_iq[d$mom_hs=="si"], d$kid_score[d$mom_hs=="si"], col = "blue", pch = 19, ylab = "kid score",
points(d$mom_iq[d$mom_hs=="no"], d$kid_score[d$mom_hs=="no"], col = "red", pch = 19)
abline(a = m4$coeff[1], b = m4$coeff[2], col = "red")
abline(a = m4$coeff[1] + m4$coeff[3], b = m4$coeff[2], col = "blue")
legend(70,160, legend = c("mom_hs = si","mom_hs = no"), col = c("blue","red"), lty = c(1,1))
```



## 5.2 Con interacción

```
m5 = lm(kid_score ~ mom_iq * mom_hs, data = d)
summary(m5)

##
## 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
```

El modelo es:

$$E[kid\_score_i] = \beta_0 + \beta_1 mom\_iq_i + \beta_2 mom\_hssi_i + \beta_3 mom\_hssi_i * mom\_iq_i$$

Que en realidad son dos modelos con distinta ordenada en el origen y distinta pendiente:

- Si  $mom\_hssi = 0$ :

$$E[kid\_score_i] = \beta_0 + \beta_1 mom\_iq_i$$

- Si  $mom\_hssi = 1$ :

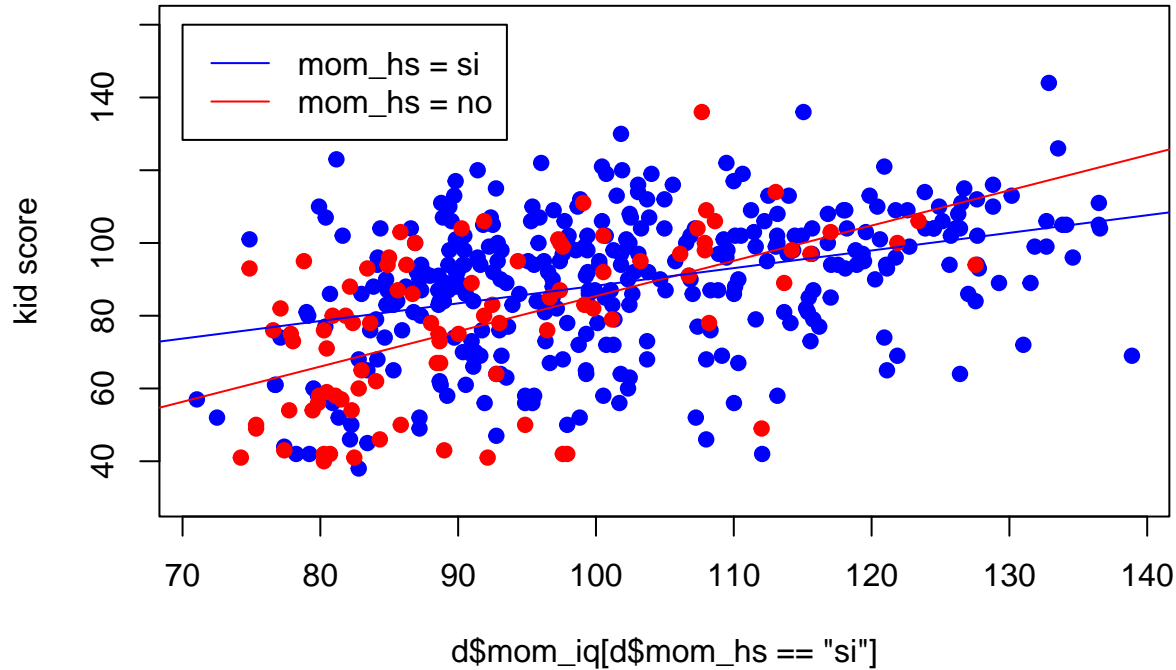
$$E[kid\_score_i] = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) mom\_iq_i$$

Por tanto:

- La puntuación del test para chicos cuya madre no completó el bachillerato y tienen  $IQ = 0$  es -11.48 en promedio. Mirando el pvalor,  $\beta_0 = 0$ .
- La puntuación del test para los chicos cuya madre no completó el bachillerato aumenta 0.97 unidades cuando el IQ de la madre aumenta una unidad. Mirando el pvalor,  $\beta_1 \neq 0$ .
- La puntuación del test para chicos cuya madre completó el bachillerato y tienen  $IQ = 0$  es (-11.48 + 51.27). Mirando el pvalor,  $\beta_2 \neq 0$ , la ordenada en el origen no es la misma para ambos grupos.
- La puntuación del test para los chicos cuya madre completó el bachillerato aumenta (0.97 - 0.48) unidades cuando el IQ de la madre aumenta una unidad. Mirando el pvalor,  $\beta_3 \neq 0$ , pendiente no es la misma para ambos grupos.

Gráficamente:

```
plot(d$mom_iq[d$mom_hs=="si"], d$kid_score[d$mom_hs=="si"], col = "blue", pch = 19, ylab = "kid score",
points(d$mom_iq[d$mom_hs=="no"], d$kid_score[d$mom_hs=="no"], col = "red", pch = 19)
abline(a = m5$coeff[1], b = m5$coeff[2], col = "red")
abline(a = m5$coeff[1] + m5$coeff[3], b = m5$coeff[2] + m5$coeff[4], col = "blue")
legend(70,160, legend = c("mom_hs = si","mom_hs = no"), col = c("blue","red"), lty = c(1,1))
```



## 6 Dos regresores cuantitativos

```
m6 = lm(kid_score ~ mom_iq + mom_age, data = d)
summary(m6)
```

```
##
## Call:
## lm(formula = kid_score ~ mom_iq + mom_age, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -56.941 -12.493   2.257  11.614  46.711
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.59625    9.08397   1.937  0.0534 .
## mom_iq        0.60357    0.05874  10.275 <2e-16 ***
## mom_age       0.38813    0.32620   1.190  0.2348
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.26 on 431 degrees of freedom
## Multiple R-squared:  0.2036, Adjusted R-squared:  0.1999
## F-statistic: 55.08 on 2 and 431 DF, p-value: < 2.2e-16
```



- Interpretación de  $\beta_1$ : Se interpreta como el aumento de la puntuación media cuando incrementamos en una unidad el IQ de las madres y mantenemos constante la edad de las madres. Efectivamente, sean la madre-hijo 1 y la madre-hijo 2. Los modelos para ambos son:

$$E[kid\_score_1] = \beta_0 + \beta_1 mom\_iq_1 + \beta_2 mom\_age_1 \quad E[kid\_score_2] = \beta_0 + \beta_1 mom\_iq_2 + \beta_2 mom\_age_2$$

Restando:

$$E[kid\_score_1] - E[kid\_score_2] = \beta_1(mom\_iq_1 - mom\_iq_2) + \beta_2(mom\_age_1 - mom\_age_2)$$

Luego si  $(mom\_iq_1 - mom\_iq_2) = 1$  y  $(mom\_age_1 - mom\_age_2) = 0$ , entonces  $\beta_1 = E[kid\_score_1] - E[kid\_score_2]$ . El pvalor para  $\beta_1$  es muy pequeño, luego  $\beta_1$  es significativo.

- Interpretación de  $\beta_2$ : Se interpreta como el aumento de la puntuación media cuando incrementamos en una unidad la edad de las madres y mantenemos constante el IQ de las madres. Procediendo igual que antes, tenemos:

$$E[kid\_score_1] - E[kid\_score_2] = \beta_1(mom\_iq_1 - mom\_iq_2) + \beta_2(mom\_age_1 - mom\_age_2)$$

$$\Rightarrow \Delta E[kid\_score] = \beta_1 \Delta mom\_iq_1 + \beta_2 \Delta mom\_age_1$$

Luego si  $\Delta mom\_iq = 0$  y  $\Delta mom\_age = 1$ , entonces  $\beta_2 = \Delta E[kid\_score]$  El pvalor para  $\beta_2$  es mayor que 0.05, luego  $\beta_2$  no es significativo.