Modelo con un regresor

Contents

1 Introducción

Vamos a leer el archivo de datos *kidiq.csv*:

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

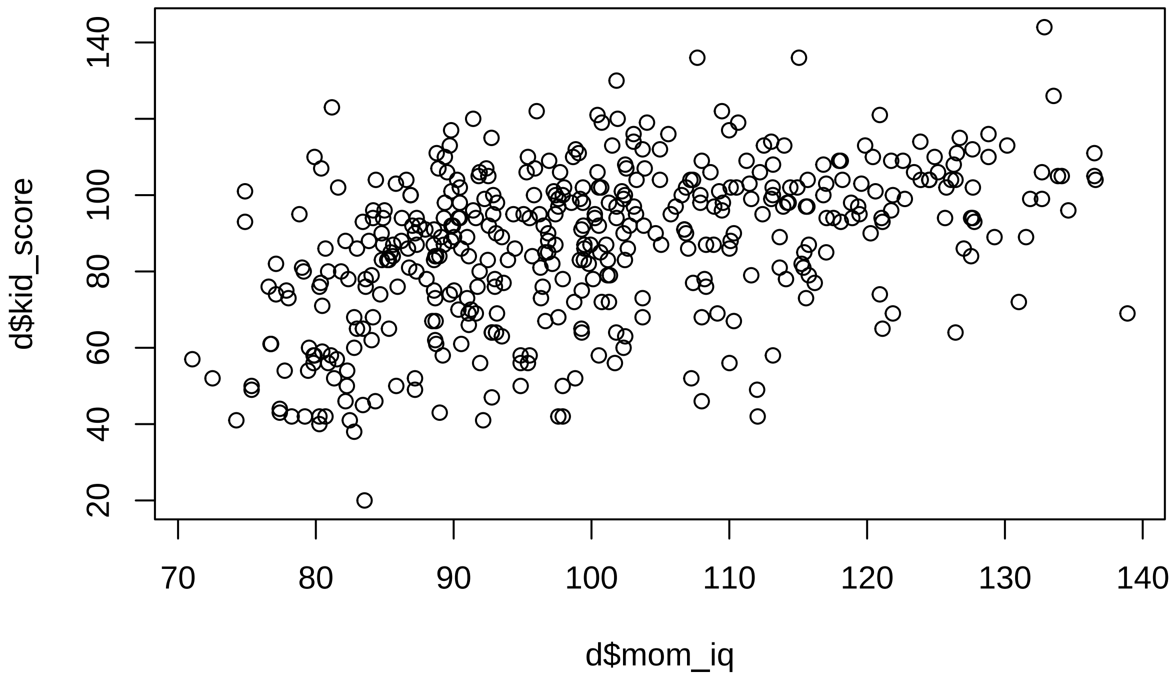
```
## 'data.frame':    434 obs. of  5 variables:
##  $ kid_score: int  65 98 85 83 115 98 69 106 102 95 ...
##  $ mom_hs   : int  1 1 1 1 1 0 1 1 1 1 ...
##  $ mom_iq   : num  121.1 89.4 115.4 99.4 92.7 ...
##  $ mom_work : int  4 4 4 3 4 1 4 3 1 1 ...
##  $ mom_age  : int  27 25 27 25 27 18 20 23 24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

donde se recogen datos de las siguientes variables:

- **kid_score** : puntuacion de un test cognitivo en niños de 3-4 años
- **mom_hs** :
  - mom_hs = 1 : las madres han terminado secundaria (high school)
  - mom_hs = 0 : las madres no terminaron secundaria
- **mom_iq** : puntuación de la madre en otro test cognitivo
- **mom_work** :
  - mom_work = 1 : la madre no trabajó en los primeros tres años del niño
  - mom_work = 2 : la madre trabajó en el segundo o tercer año
  - mom_work = 3 : la madre trabajó a tiempo parcial el primer año
  - mom_work = 4 : la madre trabajó a tiempo completo el primer año
- **mom_age** : edad de la madre

Estamos interesados en estudiar si la puntuación obtenida por los niños (variable *kid_score*) está relacionada con la puntuación obtenida por las madres (*mom_iq*). Primero dibujamos el gráfico de dispersión:

```
plot(d$mom_iq, d$kid_score)
```



Como se observa, en términos generales cuando mayor es la puntuación obtenida por las madres mayor es la puntuación de los niños.

2 Ecuación del modelo

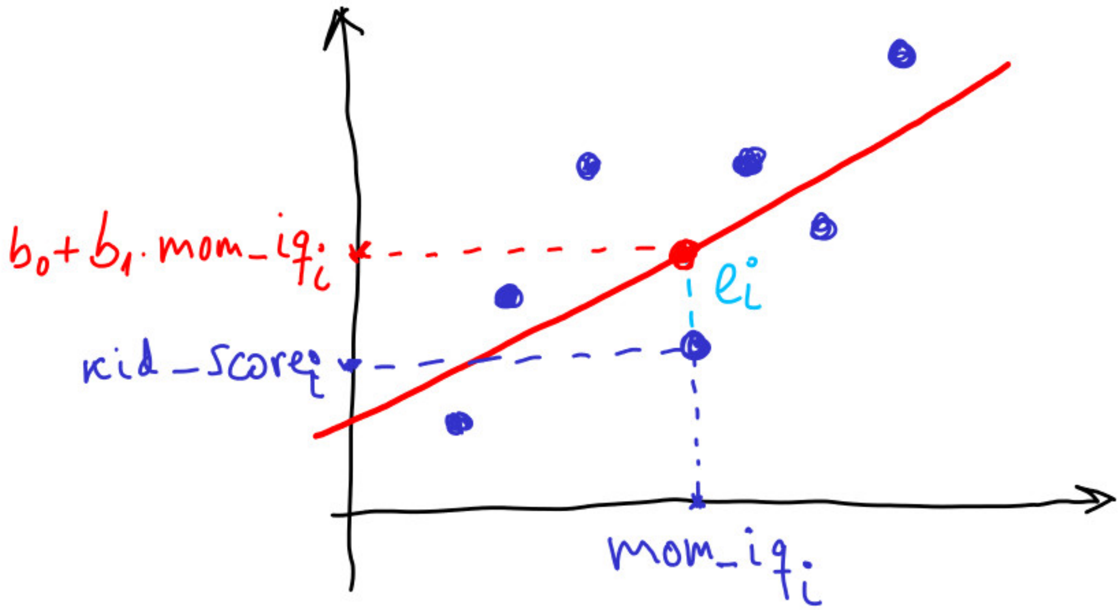
El modelo más sencillo que relaciona ambas variables es el modelo lineal:

$$kid\_score_i = b_0 + b_1 mom\_iq_i,\ i = 1, 2, \cdots, n$$

Es la ecuación de una recta. Sin embargo, es imposible calcular una recta que pase por todos los puntos del gráfico. En este caso se tiene que utilizar el modelo:

$$kid\_score_i = b_0 + b_1 mom\_iq_i + e_i,\ i = 1, 2, \cdots, n$$

es decir, se incluye el término e_i que modela la diferencia entre el valor observado en kid_score_i y el valor que toma la recta en ese punto ($b_0 + b_1 mom\_iq_i$).



Estos términos se denominan **residuos**, y se definen como:

$$e_i = kid\_score_i - (b_0 + b_1 mom\_iq_i),\ i = 1, 2, \cdots, n$$

3 Notación matricial del modelo

El modelo anterior se denomina **modelo de regresión lineal con un regresor**. De forma genérica se puede escribir así:

$$kid\_score_i = b_0 + b_1 mom\_iq_i + e_i,\ i = 1, 2, \cdots, n$$

Si escribimos la ecuación para todos los datos disponibles:

$$i = 1 \Rightarrow kid\_score_1 = b_0 + b_1 mom\_iq_1 + e_1$$

$$i = 2 \Rightarrow kid\_score_2 = b_0 + b_1 mom\_iq_2 + e_2$$

$$\cdots$$

$$i = n \Rightarrow kid\_score_n = b_0 + b_1 mom\_iq_n + e_n$$

Agrupando:

$$\begin{bmatrix} kid\_score_1 \\ kid\_score_2 \\ \cdots \\ kid\_score_n \end{bmatrix} = \begin{bmatrix} 1 & mom\_iq_1 \\ 1 & mom\_iq_2 \\ \cdots & \cdots \\ 1 & mom\_iq_n \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \cdots \\ e_n \end{bmatrix}$$

Finalmente, en notación matricial:

$$y = XB + e$$

donde B es el vector de parámetros:

$$B = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$$

4 Estimación del modelo usando mínimos cuadrados

El modelo propuesto depende de dos parámetros, b_0 y b_1, que son desconocidos. Existen diferentes métodos para calcular dichos parámetros, entre ellos, el método de mínimos cuadrados. Este método consiste en calcular el valor del vector B que minimiza la suma de los residuos al cuadrado (RSS, *residuals sum of squares*):

$$RSS = \sum e_i^2 = e^T e = (y - XB)^T (y - XB) = RSS(B)$$

Desarrollando el producto:

$$RSS(B) = y^T y - y^T XB - B^T X^T y + B^T X^T XB$$

Para calcular el mínimo se deriva respecto a B y se iguala a cero (ver Apendice)

$$\frac{dRSS(B)}{dB} = -X^T y - X^T y + (X^T X + X^T X)B = 0$$

$$B = (X^T X)^{-1} X^T y$$

5 Datos, modelo y residuos

Los datos disponibles son

$$\{kid\_score_i,\ mom\_iq_i\},\ i = 1, \cdots, n$$

Esos datos los modelamos utilizando la ecuación:

$$kid\_score_i = b_0 + b_1 mom\_iq_i + e_i,\ i = 1, 2, \cdots, n$$

Es decir, para una madre dada mom_iq_i, dividimos la puntuación de su hijo kid_score_i en dos partes: la parte que corresponde a la recta $b_0 + b_1 mom\_iq_i$ y los residuos e_i. La parte correspondiente a la recta se puede representar matricialmente como:

$$\hat{y} = XB$$

donde $\hat{y} = [\hat{y}_1\ \hat{y}_2\ \cdots\ \hat{y}_n]^T$. Por tanto los residuos se pueden calcular como

$$e_i = y_i - \hat{y}_i,\ i = 1, 2, \ldots, n$$

o en forma matricial

$$e = y - \hat{y}$$

6 Aplicacion a los datos del ejemplo

- Matrices del modelo

```
y = matrix(d$kid_score, ncol = 1)
head(y)
```

```
##      [,1]
## [1,]   65
## [2,]   98
## [3,]   85
## [4,]   83
## [5,]  115
## [6,]   98
```

```
n = nrow(d)
X = cbind(rep(1,n), d$mom_iq)
head(X)
```

```
##      [,1]      [,2]
## [1,]    1 121.11753
## [2,]    1  89.36188
## [3,]    1 115.44316
## [4,]    1  99.44964
## [5,]    1  92.74571
## [6,]    1 107.90184
```

- Estimacion

```
Xt_X = t(X) %*% X
Xt_y = t(X) %*% y
( B = solve(Xt_X) %*% Xt_y )
```

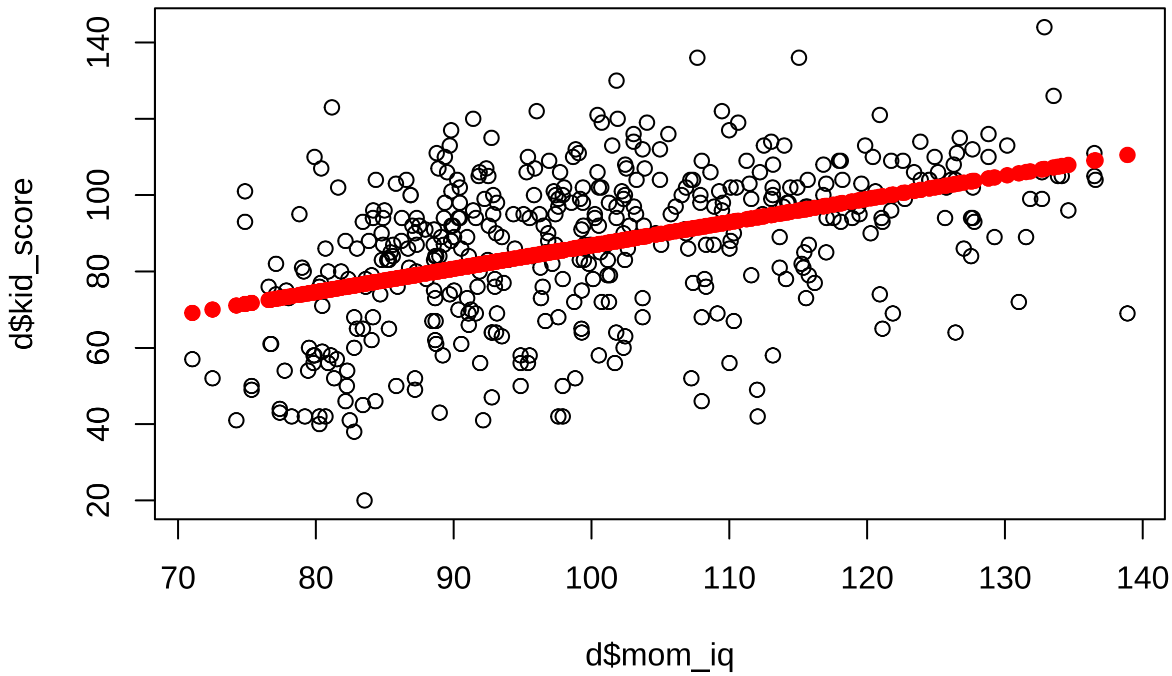
```
##            [,1]
## [1,] 25.7997778
## [2,]  0.6099746
```

- valores de la recta

```
y_e = X %*% B
```

Estos valores se pueden representar

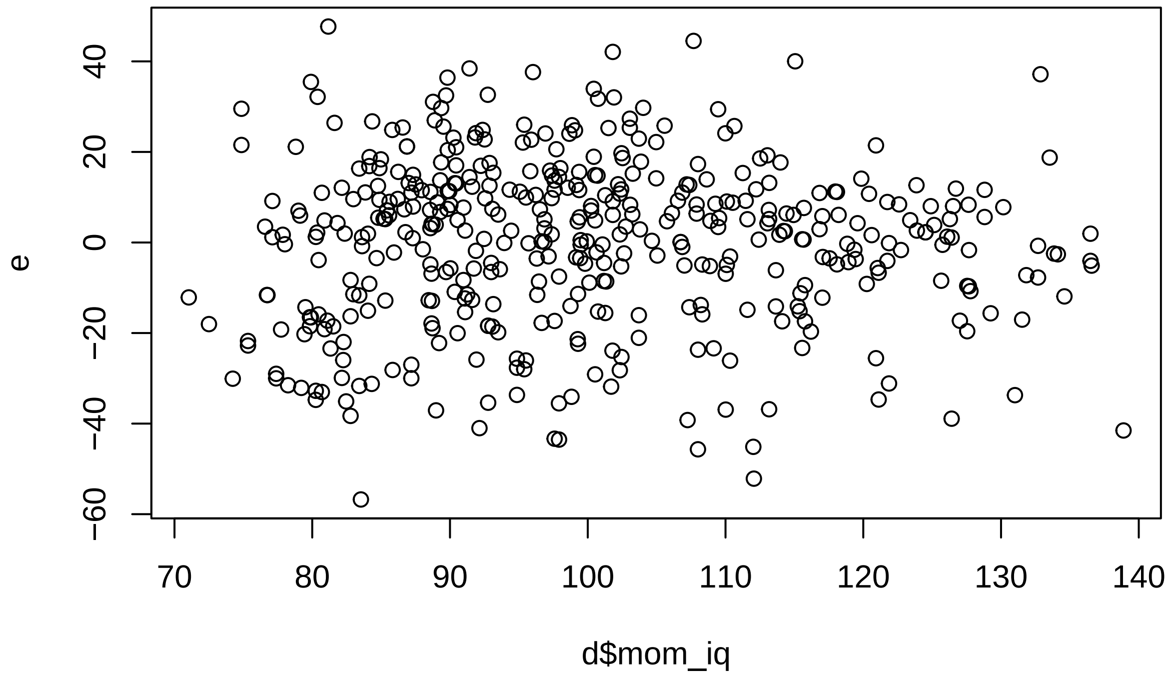
```
plot(d$mom_iq, d$kid_score)
points(d$mom_iq, y_e, col = "red", pch = 19)
```



Finalmente, los residuos se calculan haciendo

```
e = y - y_e
```

```
plot(d$mom_iq, e)
```



7 Bondad del modelo ajustado

Es conveniente medir como de bueno es el ajuste del modelo. Una posibilidad es usar la suma de los residuos al cuadrado o RSS:

```r
(RSS = sum(e^2))
```

```
## [1] 144137.3
```

Pero esta variable depende de las unidades de x e y. Por tanto es difícil saber si un RSS alto indica que el modelo es bueno o malo. Lo ideal es utilizar variables adimensionales. La manera mas usual es utilizar el coeficiente de determinación o R^2:

$$R^2 = 1 - \frac{RSS}{TSS}$$

donde TSS es la suma total de cuadrados

$$TSS = \sum (y_i - \bar{y})^2$$

```r
(TSS = sum((y-mean(y))^2))
```

```
## [1] 180386.2
```

```r
(R2 = 1 - RSS/TSS)
```

```
## [1] 0.2009512
```

El coeficiente R^2 toma valores entre cero y uno. Si $R^2 \approx 1 \Rightarrow RSS \ll TSS$, es decir, los residuos son muy pequeños en comparación a los datos, luego el modelo se ajusta muy bien a los datos. Cuando $R^2 \approx 0$, los residuos son muy grandes y el modelo no se ajusta bien a los datos.

La suma total de cuadrados de y está relacionado con su varianza, ya que

$$s_y^2 = \frac{\sum (y_i - \bar{y})^2}{n-1} \Rightarrow TSS = (n-1)s_y^2$$

```r
(n-1)*var(y)
```

```
##          [,1]
## [1,] 180386.2
```