Aprendizaje lineal

Antonio Falcó

Seminario 7

Relación entre dos variables NO independientes

Si X e Y no son independientes entonces conocemos que

$$\mathbb{E}(Y|X = x) = g(x)$$

es una función que depende de los valores observados x, y del mismo modo

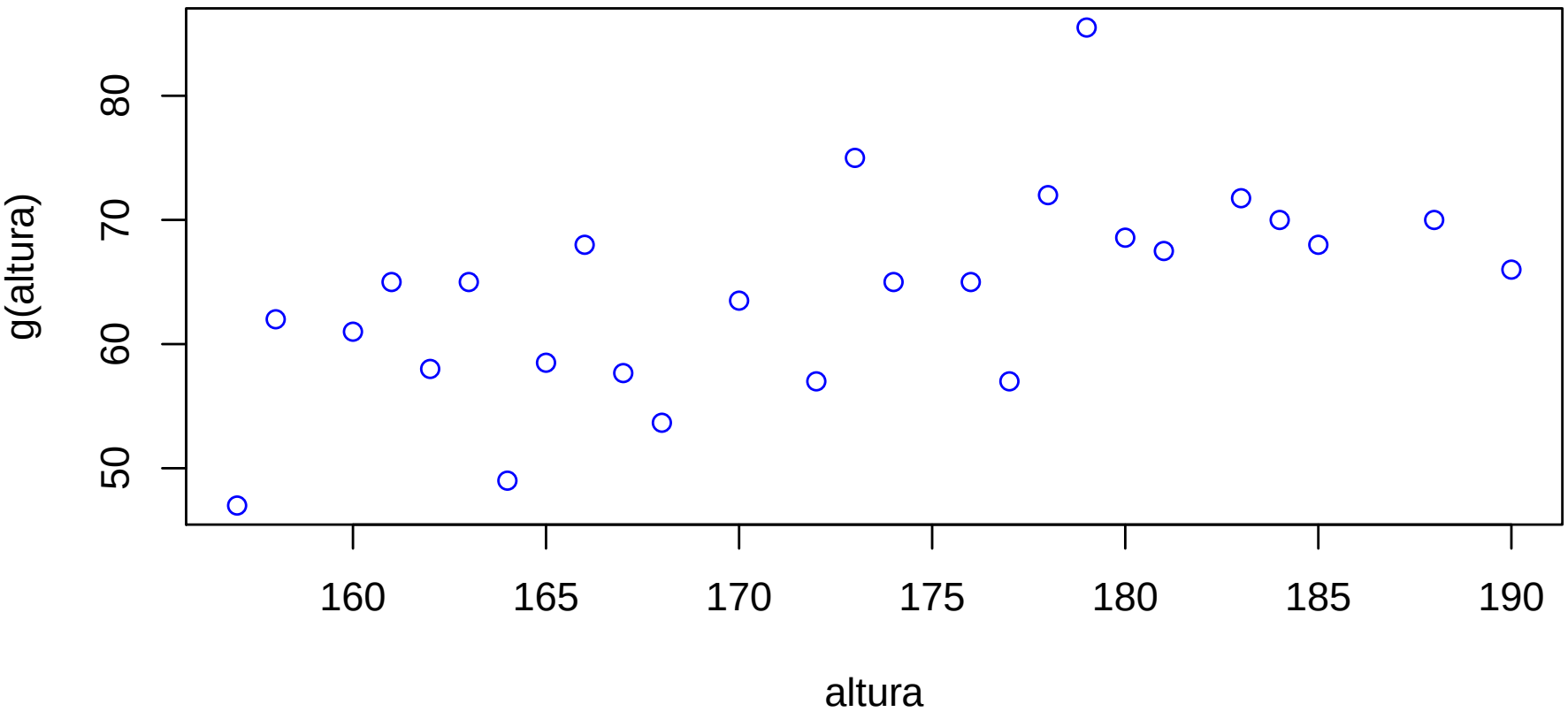
$$\mathbb{E}(X|Y = y) = h(y)$$

es una función que depende de los valores observados y.

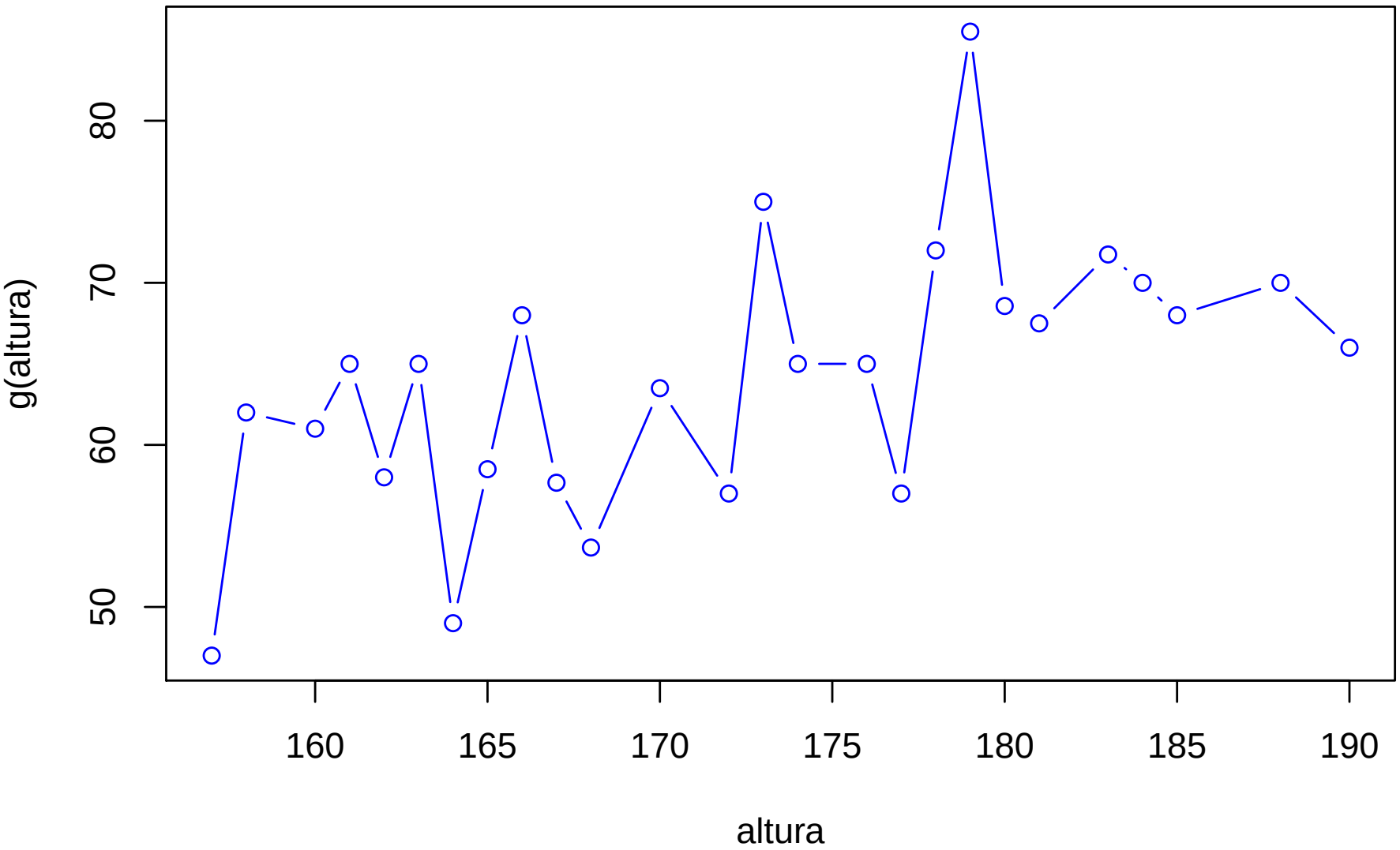
Ejemplo

Consideremos las variables $X = \texttt{altura}$ y $Y = \texttt{peso}$, entonces veamos la función del peso esperado en función de la altura observada:

$$g(x) = \mathbb{E}(\texttt{peso}|\texttt{altura} = x)$$



Podemos interpolar entre los puntos:



El valor esperado condicional como variable aleatoria

Si X e Y no son independientes entonces podemos considerar que

$$\mathbb{E}(Y|X) = g(X)$$

es una variable aleatoria que depende de los valores observados de manera aleatoria de X, y del mismo modo

$$\mathbb{E}(X|Y) = h(Y)$$

es una variable aleatoria que depende de los valores observados de manera aleatoria de Y.

Principio de coherencia

Se puede demostrar matemáticamente que:

- $\mathbb{E}\left(\mathbb{E}(Y|X)\right) = \mathbb{E}\left(g(X)\right) = \mathbb{E}(Y)$,
- $\mathbb{E}\left(\mathbb{E}(X|Y)\right) = \mathbb{E}\left(h(Y)\right) = \mathbb{E}(X)$.

Hipótesis I: $g(X) = \beta X + \alpha$

- Si suponemos (verdadera) la relación lineal entre

$$\texttt{peso} = Y = g(\texttt{altura}) = g(X) = \beta\,\texttt{altura} + \alpha = \beta X + \alpha$$

para unos valores β y α desconocidos, vemos que la relación no es perfecta (en la gráfica anterior los valores que obtenemos no están situados sobre una recta).

- Podemos entonces escribir el valor esperado condicional

$$\mathbb{E}(Y|X) = \sum_{y \in Y(\Omega)} y f_Y(y|X) = g(X) = \beta X + \alpha.$$

es una variable aleatoria que depende de X.

- Empleando el principio de coherencia, obtenemos

$$\mathbb{E}\left(\mathbb{E}(Y|X)\right) = \mathbb{E}\left(Y\right) = \mathbb{E}\left(\beta X + \alpha\right) = \beta\mathbb{E}(X) + \alpha,$$

es decir se cumple:

$$\mathbb{E}\left(Y\right) = \beta\mathbb{E}(X) + \alpha.$$

Hipótesis II: La condición de mínima error cuadrático medio

- Conocemos que por el principio de coherencia

$$\mu_Y = \mathbb{E}(Y) = \beta\mathbb{E}(X) + \alpha = \beta\,\mu_X + \alpha.$$

- La varianza de Y se expresa como

$$\sigma_Y^2 = \mathrm{Var}(Y) = \mathbb{E}((Y - \mu_Y)^2) = \mathbb{E}((Y - (\beta\,\mu_X + \alpha))^2) \geq 0,$$

  y es independiente de los valores de α y β que cumplan $\mu_Y = \beta\,\mu_X + \alpha$.
- Entonces buscamos los valores $\widehat{\beta}$ y $\widehat{\alpha}$ que cumplen el principio de coherencia y que además minimicen el error cuadrático medio, esto es, se tiene que cumplir

$$\mathbb{E}((Y - (\widehat{\beta}\,X + \widehat{\alpha}))^2) \leq \mathbb{E}((Y - (\beta\,X + \alpha))^2)$$

  para cualquier otro valor de β y α.

Calculo de los valores óptimos

- Observemos que

$$\begin{aligned}\mathbb{E}((Y-(\beta X+\alpha))^2) &= \mathbb{E}((Y-\alpha)-\beta X)^2) \\ &= \mathbb{E}((Y-\alpha)^2-2\beta X(Y-\alpha)+X^2\beta^2) \\ &= \mathbb{E}((Y-\alpha)^2)-2\beta\,\mathbb{E}(X(Y-\alpha))+\beta^2\,\mathbb{E}(X^2) \\ &= c-2b\,\beta+a\,\beta^2,\end{aligned}$$

  donde

$$c=\mathbb{E}((Y-\alpha)^2), b=\mathbb{E}(X(Y-\alpha)) \text{ y } a=\mathbb{E}(X^2).$$

- La función a minimizar es un polinomio de segundo grado en la variable β :

$$\mathbb{E}((Y-(\beta X+\alpha))^2)=c-2b\,\beta+a\,\beta^2,$$

- El coeficiente a del término β^2, es $a=\mathbb{E}(X^2)\geq 0$. Entonces el polinomio alcanza su valor mínimo para $\widehat{\beta}$ solución de la ecuación:

$$2b-2\widehat{\beta}a=0\Rightarrow\widehat{\beta}=\frac{b}{a}=\frac{\mathbb{E}(X(Y-\alpha))}{\mathbb{E}(X^2)}.$$

Los valores óptimos para la aproximación lineal

- Conocemos que

$$\widehat{\beta} = \frac{\mathbb{E}(X(Y-\alpha))}{\mathbb{E}(X^2)} = \frac{\mathbb{E}(XY-\alpha X)}{\mathbb{E}(X^2)} = \frac{\mathbb{E}(XY)-\alpha\mathbb{E}(X)}{\mathbb{E}(X^2)} = \frac{\mathbb{E}(XY)-\alpha\mu_X}{\mathbb{E}(X^2)}$$

- Además se tiene que cumplir (principio de coherencia)

$$\mu_Y = \widehat{\beta}\,\mu_X + \widehat{\alpha} \Rightarrow \widehat{\alpha} = \mu_Y - \widehat{\beta}\,\mu_X,$$

es decir

$$\widehat{\beta} = \frac{\mathbb{E}(XY)-(\mu_Y-\widehat{\beta}\,\mu_X)\mu_X}{\mathbb{E}(X^2)}.$$

- Despejando $\widehat{\beta}$ obtenemos los valores óptimos:

$$\widehat{\beta} = \frac{\mathbb{E}(XY)-\mu_X\mu_Y}{\mathbb{E}(X^2)-\mu_X^2} = \frac{\mathbb{E}((X-\mu_X)(Y-\mu_Y))}{\sigma_X^2} \text{ y}$$

$$\widehat{\alpha} = \mu_Y - \frac{\mathbb{E}((X-\mu_X)(Y-\mu_Y))}{\sigma_X^2}\,\mu_X.$$

Ejemplo

En nuestro caso si $X = \texttt{altura}$ e $Y = \texttt{peso}$ obtenemos

$$\widehat{\beta} = 0.585263 \quad \text{y} \quad \widehat{\alpha} = -36.928954,$$

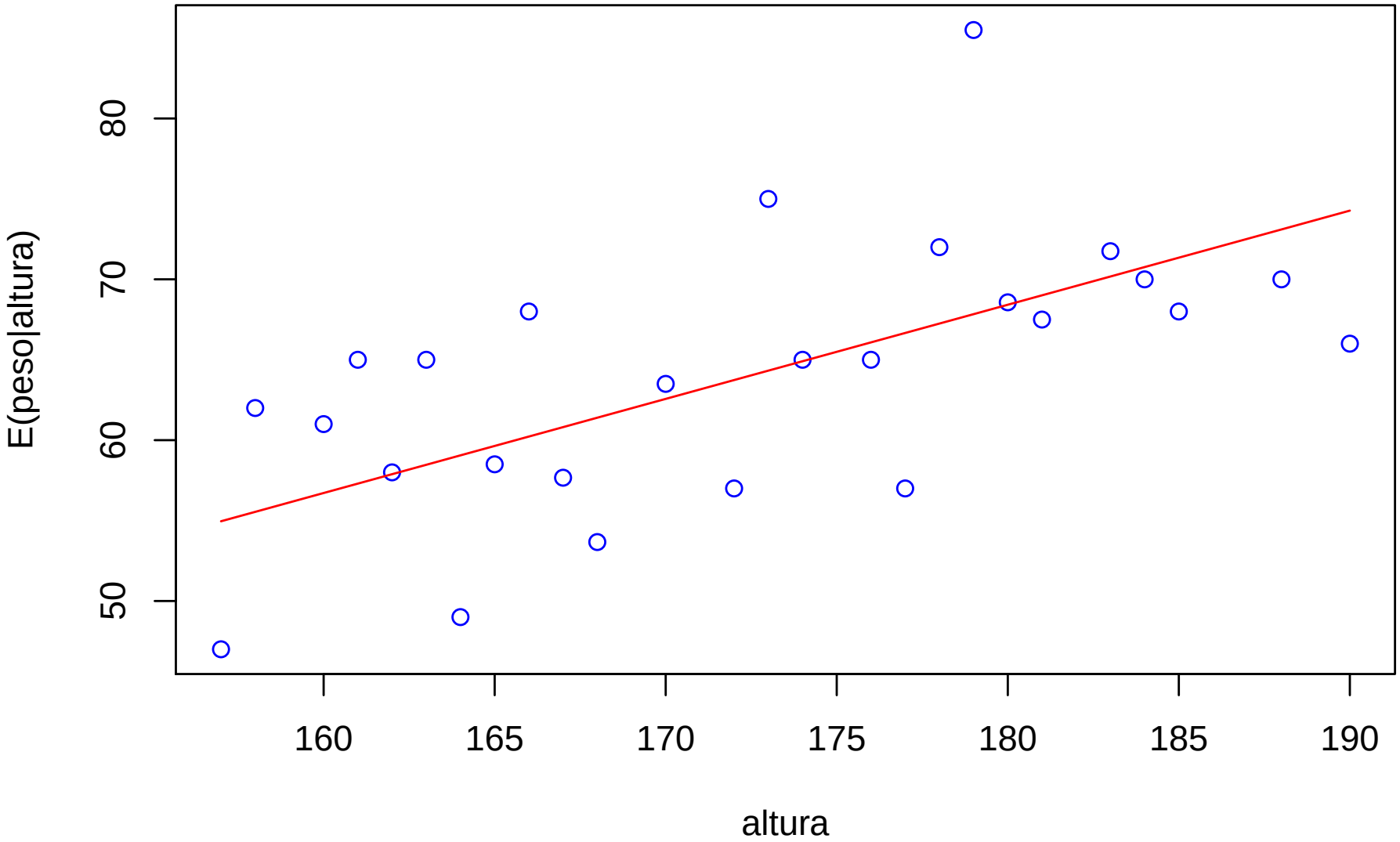
el modelo queda entonces

$$\mathbb{E}(\texttt{peso}|\texttt{altura}) = 0.585263\,\texttt{altura} - 36.928954,$$

de forma que el error cuadrático medio

$$\mathbb{E}((\texttt{peso} - 0.585263\,\texttt{altura} + 36.928954)^2)$$

toma el valor mínimo posible.


E(peso|altura)
50
60
70
80
160
165
170
175
180
185
190
altura

Conclusión

El modelo de regresión lineal establece una relación

$$\mathbb{E}(Y|X) = \widehat{\beta}X + \widehat{\alpha}$$

de forma que el error cuadrático medio $\mathbb{E}((Y - \widehat{\beta}X - \widehat{\alpha})^2)$ es el menor posible, esto es, la desigualdad

$$0 \leq \mathbb{E}((Y - \widehat{\beta}X - \widehat{\alpha})^2) \leq \mathbb{E}((Y - \beta X - \alpha)^2)$$

se verifica para cualquier otro valor de α y β.

El valor del error cuadrático medio

- Observemos que

$$(Y - \widehat{\beta}X - \widehat{\alpha})^2 =$$

$$\left(Y - \frac{\mathbb{E}((X-\mu_X)(Y-\mu_Y))}{\sigma_X^2} X - \mu_Y + \frac{\mathbb{E}((X-\mu_X)(Y-\mu_Y))}{\sigma_X^2}\mu_X\right)^2 =$$

$$\left((Y-\mu_Y) - \frac{\mathbb{E}((X-\mu_X)(Y-\mu_Y))}{\sigma_X^2}(X-\mu_X)\right)^2 =$$

$$(Y-\mu_Y)^2 + \left(\frac{\mathbb{E}((X-\mu_X)(Y-\mu_Y))}{\sigma_X^2}\right)^2 (X-\mu_X)^2 -$$

$$2(Y-\mu_Y)(X-\mu_X)\left(\frac{\mathbb{E}((X-\mu_X)(Y-\mu_Y))}{\sigma_X^2}\right)$$

El valor del error cuadrático medio

- Si tomamos valores esperados, obtenemos que

$$\mathbb{E}((Y-\widehat{\beta}X-\widehat{\alpha})^2)=\sigma_Y^2+\left(\frac{\mathbb{E}((X-\mu_X)(Y-\mu_Y))}{\sigma_X^2}\right)^2\sigma_X^2 - 2\left(\frac{\mathbb{E}((X-\mu_X)(Y-\mu_Y))}{\sigma_X^2}\right)^2\sigma_X^2,$$

es decir, el valor del error cuadrático médio es

$$\mathbb{E}((Y-\widehat{\beta}X-\widehat{\alpha})^2)=\sigma_Y^2-\left(\frac{\mathbb{E}((X-\mu_X)(Y-\mu_Y))}{\sigma_X^2}\right)^2\sigma_X^2\leq\sigma_Y^2,$$

menor que la varianza de Y y es estrictamente menor siempre que cumpla:

$$\mathbb{E}((X-\mu_X)(Y-\mu_Y))\neq 0.$$

Problema metodológico

El modelo establece que

$$\mathbb{E}(\texttt{peso}|\texttt{altura}) = 0.585263\,\texttt{altura} - 36.928954,$$

donde la altura se midel en cm. Si los parámetros $\widehat{\beta}$ y $\widehat{\alpha}$ carecen de unidades, entonces las unidades de $\mathbb{E}(\texttt{peso}|\texttt{altura})$ son también cm. Sin embargo la unidades de la varoable peso son kg. Parece coherente pensar que para que la fórmula para predecir el peso mediante observaciones de la altura sea coherente, los parámetros de regresión $\widehat{\beta}$ y $\widehat{\alpha}$ deberían tener unidades:

$$\widehat{\beta} = 0.585263 \text{ kg/cm}$$

y

$$\widehat{\alpha} = -36.928954 \text{ cm}.$$

Cuestión científica importante

¿Se puede probar que el modelo de regresión no depende de las unidades en las que se miden las variables del modelo?

La independencia de las unidades

Dadas dos variables X e Y **NO** independientes definidas sobre una población experimental $(\Omega, \mathbb{P})$ con valores esperados $\mathbb{E}(X) = \mu_X$, $\mathbb{E}(Y) = \mu_Y$, y varianzas $\text{Var}(X) = \sigma_X^2$ y $\text{Var}(Y) = \sigma_Y^2$, consideramos las variables sin unidades siguientes

$$\frac{X - \mu_X}{\sigma_X} \quad \frac{Y - \mu_Y}{\sigma_Y}$$

entonces planteamos el buscar el valor de γ de forma para que se cumpla

1. $\mathbb{E}\left(\frac{Y-\mu_Y}{\sigma_Y} \middle| X\right) = \gamma \frac{X-\mu_X}{\sigma_X}$, y
2. que minimice el error cuadrático medio

$$\mathbb{E}\left(\left(\frac{Y - \mu_Y}{\sigma_Y} - \gamma \frac{X - \mu_X}{\sigma_X}\right)^2\right)$$

Principio de coherencia

El principio de coherencia se cumple de manera automática:

$$\mathbb{E}\left(\frac{Y-\mu_Y}{\sigma_Y}\right)=\frac{\mathbb{E}(Y)-\mu_Y}{\sigma_Y}=0=\gamma\,\mathbb{E}\left(\frac{X-\mu_X}{\sigma_X}\right)=\gamma\cdot 0.$$

Segunda condición

El valor de γ que minimiza la expresión

$$\mathbb{E}\left(\left(\frac{Y-\mu_Y}{\sigma_Y}-\gamma\frac{X-\mu_X}{\sigma_X}\right)^2\right)$$

nos da un valor de

$$\widehat{\gamma}=\frac{\sigma_X}{\sigma_Y}\frac{\mathbb{E}((X-\mu_x)(Y-\mu_Y))}{\mathrm{Var}(X)}=\frac{\sigma_X}{\sigma_Y}\widehat{\beta}=\frac{\mathbb{E}((X-\mu_x)(Y-\mu_Y))}{\sigma_X\sigma_Y},$$

que coincide con el llamado **coeficiente de correlación** $\rho_{X,Y}$, es decir

$$\widehat{\gamma}=\rho_{X,Y}=\frac{\mathbb{E}((X-\mu_x)(Y-\mu_Y))}{\sigma_X\sigma_Y}=\frac{\sigma_X}{\sigma_Y}\widehat{\beta},$$

y en consecuencia podemos considerar que $\widehat{\beta}$ carece de unidades, ya que $\widehat{\gamma}$ carece de de ellas.

El modelo de regresión lineal

El modelo de regresión lineal es bajo estas hipótesis:

$$\mathbb{E}\left(\frac{Y-\mu_Y}{\sigma_Y}\middle| X\right) = \widehat{\gamma}\,\frac{X-\mu_X}{\sigma_X},$$

que equivale a escribir,

$$\mathbb{E}\,(\,Y-\mu_Y|\,X) = \widehat{\gamma}\,\frac{\sigma_Y}{\sigma_X}\,(X-\mu_X).$$

Finalmente, se tiene

$$\mathbb{E}\,(\,Y|\,X) = \frac{\sigma_Y}{\sigma_X}\,\widehat{\gamma}\,X + \mu_Y - \frac{\sigma_Y}{\sigma_X}\,\widehat{\gamma}\mu_X.$$

Comparativa de modelos

En este caso, el modelo de regresión lineal es equivalente a considerar

$$\mathbb{E}(Y|X) = \frac{\sigma_Y}{\sigma_X}\widehat{\gamma}X + \mu_Y - \frac{\sigma_Y}{\sigma_X}\widehat{\gamma}\mu_X.$$

con

$$\widehat{\gamma} = \frac{\sigma_X}{\sigma_Y}\widehat{\beta},$$

es decir

$$\mathbb{E}(Y|X) = \widehat{\beta}X + (\mu_Y - \widehat{\beta}\mu_X).$$

Recordemos que

$$\mu_Y - \widehat{\beta}\mu_X = \widehat{\alpha}$$

con lo que recuperamos el modelo de regresión lineal original

Conclusión

- Consideremos dos variables X e Y **NO** independientes definidas sobre una población experimental $(\Omega, \mathbb{P})$ con valores esperados $\mathbb{E}(X) = \mu_X$, $\mathbb{E}(Y) = \mu_Y$, y varianzas $\text{Var}(X) = \sigma_X^2$ y $\text{Var}(Y) = \sigma_Y^2$.
- Entonces, la mejor aproximación lineal de

$$\mathbb{E}(Y|X) = \widehat{\beta}\, X + (\mu_Y - \widehat{\beta}\, \mu_X)$$

donde

$$\widehat{\beta} = \frac{\sigma_Y}{\sigma_X}\, \rho_{X,Y}.$$

Observése que el parámetro $\widehat{\beta}$ carece de unidades y el parámetro

$$\widehat{\alpha} = \mu_Y - \widehat{\beta}\, \mu_X,$$

tiene como unidades de $\widehat{\alpha}$ = unidades de Y − unidades de X.

- El modelo hay que interpretarlo entonces como

$$\mathbb{E}((Y - \mu_Y)|X) = \widehat{\beta}\, (X - \mu_X).$$

Observación

Si consideramos el error cuadrático medio de la aproximación anterior

$$\begin{aligned}\mathbb{E}\left(\mathbb{E}(((Y-\mu_Y)-\beta(X-\mu_X))^2|X)\right) &= \mathbb{E}(((Y-\mu_Y)-\beta(X-\mu_X))^2) \\ &= \mathbb{E}((Y-\mu_Y)^2)+\beta^2\mathbb{E}((X-\mu_X)^2) \\ &- 2\beta\mathbb{E}((X-\mu_x)(Y-\mu_Y)) \\ &= \sigma_Y^2+\beta^2\sigma_X^2-2\beta\mathbb{E}((X-\mu_x)(Y-\mu_Y))\end{aligned}$$

es una función de β :

$$\beta^2\sigma_X^2-2\beta\mathbb{E}((X-\mu_x)(Y-\mu_Y))+\sigma_Y^2$$

cuyo mínimo absoluto se alcanza precisamente en

$$\widehat{\beta}=\frac{\mathbb{E}((X-\mu_x)(Y-\mu_Y))}{\sigma_X^2}.$$