Estadística: estadística
Grado en Relaciones Laborales | Curso 2019-2020
Tema 3. Técnicas de regresión y correlación.

Alejandro Saavedra Nieves

Recta de regresión: introducción

- En el Tema 1 nos hemos ocupado de la descripción de variables estadísticas unidimensionales.
- Lo habitual es que tendamos a considerar un conjunto amplio de características para describir a cada uno de los individuos de la población, y que estas características puedan presentar relación entre ellas.
- Nos centraremos en el estudio de variables estadísticas bidimensionales.
- Tal y como vimos en el Tema 2, representaremos por (X, Y) la variable bidimensional estudiada, donde X e Y son las variables unidimensionales correspondientes a las primera y segunda características, respectivamente, medidas para cada individuo.

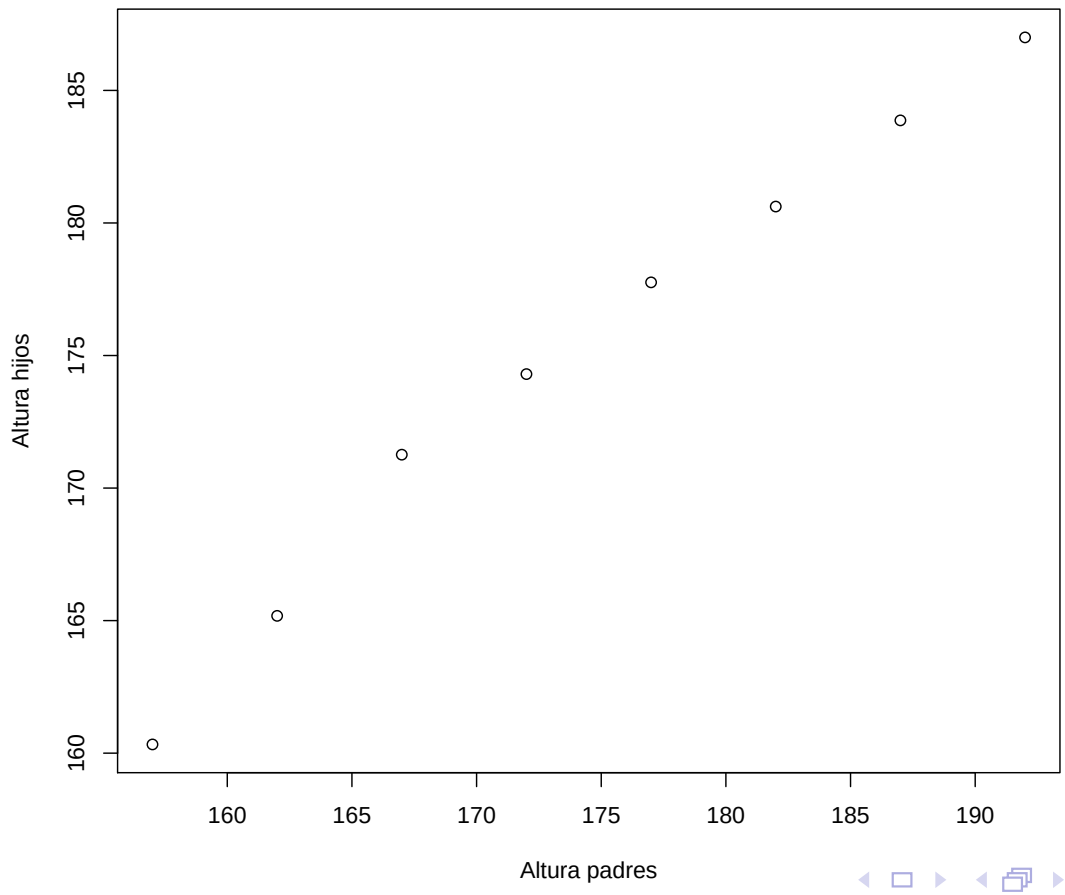
Recta de regresión. Ejemplos

- ¿Existe relación entre la altura en el peso? ¿de qué tipo es esa relación?
- ¿Cómo se relaciona la cantidad de dinero que se ha invertido un laboratorio para anunciar un nuevo fármaco con las cifras de ventas durante el primer mes?
- ¿Está relacionada la altura de un padre con la de su hijo? ¿cómo?
- ¿Está relacionado el Volumen Expiratorio Forzado con la estatura?

Recta de regresión: altura de los hijos vs. altura de los padres

La siguiente tabla detalla los datos (por pares) de la altura de 8 padres (X) y la altura de sus hijos (Y).

	1	2	3	4	5	6	7	8
Altura de los padres (X)	157	162	167	172	177	182	187	192
Altura de los hijos (Y)	160.33	165.18	171.26	174.30	177.76	180.62	183.87	187.00



Recta de regresión: introducción

La recta de regresión determina la relación lineal entre dos variables continuas. Esta recta describe cómo varía la media de una variable en función de los valores de la otra.

El coeficiente de correlación lineal es una medida resumen de la asociación lineal entre dos variables continuas.

Una manera alternativa es indicar la ecuación de la recta que describe la situación de los puntos.

En Ciencias Sociales, se establecen relaciones entre variables en promedio.

Recta de regresión: introducción

Ejemplo Existe relación entre la renta y el gasto de las familias si, por ejemplo, al aumentar la renta de las familias disminuye la proporción de gasto. Esto no implica que todas las familias de mayor renta gasten menos que las de menor renta.

En promedio, las familias de renta más alta destinarán menos a la alimentación.

- Relación positiva: si al aumentar una variable, aumenta en promedio la otra variable.
- Relación negativa: si al aumentar una variable, disminuye en promedio la otra variable.

Parece natural medir la dependencia entre dos variables describiendo cómo varía la variable dependiente en función de la variable independiente.

- Por ejemplo, cómo varía la media de la variable dependiente condicionada a los valores de la independiente.

Ejemplo. Volumen Expiratorio Forzado y estatura

- EL Volumen Expiratorio Forzado (VEF) es una medida de la capacidad pulmonar.
- Se cree que el VEF está relacionado con la estatura.
- Nos interesa estudiar la variable bidimensional (X, Y):
  - X es la estatura de niños de 10 a 15 años de edad.
  - Y es el VEF.
- A continuación se muestra la estatura (en cm.) y el VEF (en l.) de 12 niños en ese rango de edad:

Estatura	134	138	142	146	150	154	158	162	166	170	174	178
VEF	1.7	1.9	2.0	2.1	2.2	2.5	2.7	3.0	3.1	3.4	3.8	3.9

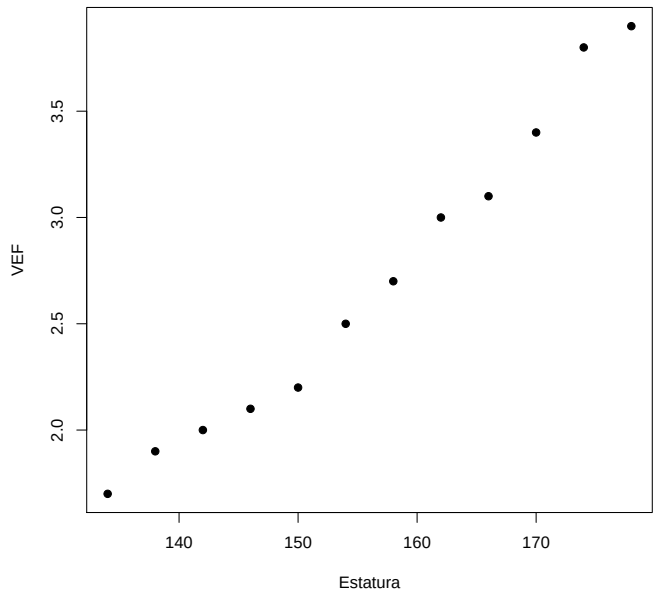
El diagrama de dispersión

- La representación gráfica más útil de dos variables continuas es el **diagrama de dispersión**.
- Consiste en representar en un eje de coordenadas los pares de observaciones (x_i, y_i).
- La nube así dibujada refleja la posible relación entre las variables.
- A mayor relación entre las variables más estrecha y alargada será la nube.

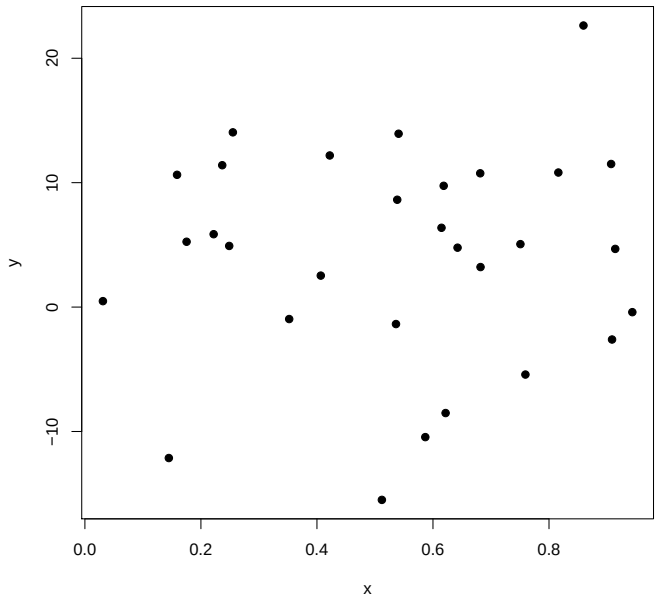
El diagrama de dispersión

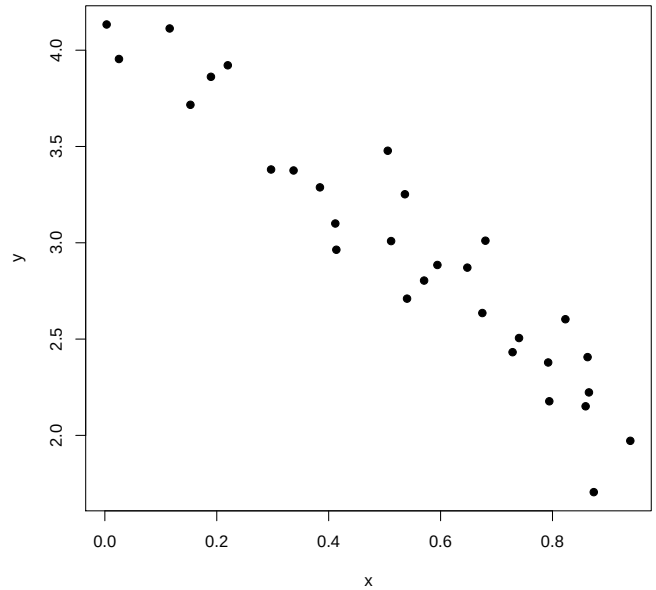
- La representación gráfica más útil de dos variables continuas es el **diagrama de dispersión**.
- Consiste en representar en un eje de coordenadas los pares de observaciones (x_i, y_i).
- La nube así dibujada refleja la posible relación entre las variables.
- A mayor relación entre las variables más estrecha y alargada será la nube.

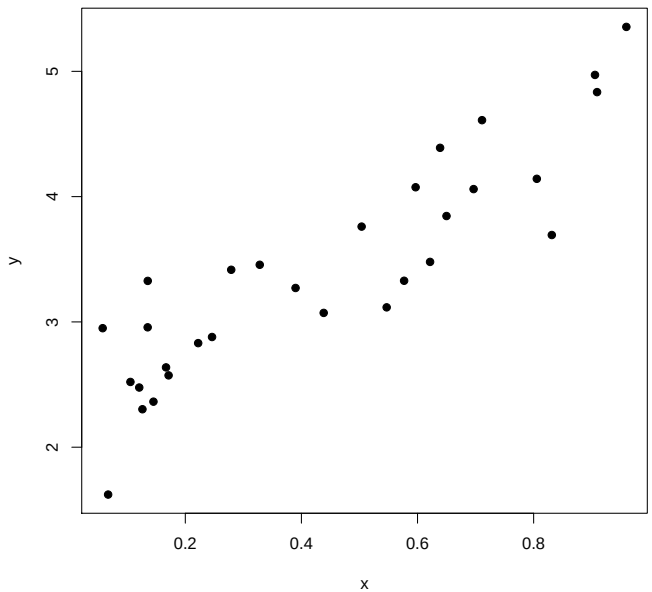
Estatura	134	138	142	146	150	154	158	162	166	170	174	178
VEF	1.7	1.9	2.0	2.1	2.2	2.5	2.7	3.0	3.1	3.4	3.8	3.9

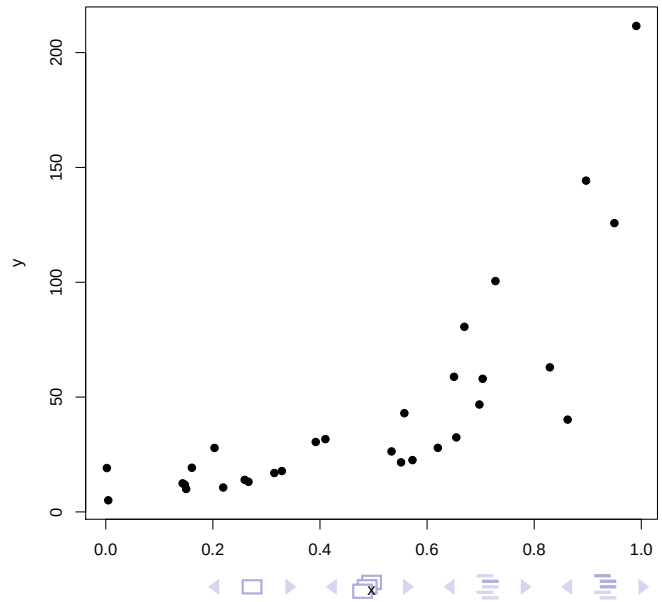


Algunos diagramas de dispersión









Covarianza

- La mayoría de las medidas características estudiadas en el caso unidimensional (como por ejemplo la media) pueden extenderse al caso bidimensional.
- Además, en el contexto bidimensional surgen nuevas medidas que nos permiten cuantificar la dispersión conjunta de dos variables estadísticas.

Covarianza entre X e Y

$$S_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

Covarianza

- La mayoría de las medidas características estudiadas en el caso unidimensional (como por ejemplo la media) pueden extenderse al caso bidimensional.
- Además, en el contexto bidimensional surgen nuevas medidas que nos permiten cuantificar la dispersión conjunta de dos variables estadísticas.

Covarianza entre X e Y

$$S_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

- La covarianza puede interpretarse como una medida de relación lineal entre las variables X e Y.

Covarianza

- La mayoría de las medidas características estudiadas en el caso unidimensional (como por ejemplo la media) pueden extenderse al caso bidimensional.
- Además, en el contexto bidimensional surgen nuevas medidas que nos permiten cuantificar la dispersión conjunta de dos variables estadísticas.

Covarianza entre X e Y

$$S_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

- La covarianza puede interpretarse como una medida de relación lineal entre las variables X e Y.
- La covarianza de (X, Y) es igual a la de (Y, X), es decir, $S_{XY} = S_{YX}$.

Covarianza

- La mayoría de las medidas características estudiadas en el caso unidimensional (como por ejemplo la media) pueden extenderse al caso bidimensional.
- Además, en el contexto bidimensional surgen nuevas medidas que nos permiten cuantificar la dispersión conjunta de dos variables estadísticas.

Covarianza entre X e Y

$$S_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

- La covarianza puede interpretarse como una medida de relación lineal entre las variables X e Y.
- La covarianza de (X, Y) es igual a la de (Y, X), es decir, $S_{XY} = S_{YX}$.
- La covarianza de (X, X) es igual a la varianza de X, es decir $S_{XX} = S_X^2$

Ejemplo. Volumen Expiratorio Forzado y estatura: covarianza

Para los datos del ejemplo sobre el VEF y la estatura se obtiene que:

- La estatura media es $\bar{x} = 156$ centímetros.
- El VEF medio es $\bar{y} = 2.691$ litros.
- La covarianza entre X e Y se calcula como

$$S_{XY} = \frac{(134 - 156) \cdot (1.7 - 2.691) + \ldots + (178 - 156) \cdot (3.9 - 2.691)}{12} = 9.783$$

- El signo de la covarianza nos indica que hay una relación positiva, es decir, a medida que aumenta la estatura aumenta el VEF.

Coeficiente de correlación lineal

- La covarianza cambia si modificamos las unidades de medida de las variables.
- Esto es un inconveniente porque no nos permite comparar la relación entre distintos pares de variables medidas en diferentes unidades.
- La solución es utilizar el **coeficiente de correlación lineal**

Coeficiente de correlación lineal entre X e Y

$$r_{XY} = \frac{S_{XY}}{S_X S_Y}.$$

Coeficiente de correlación lineal

Coeficiente de correlación lineal entre X e Y

$$r_{XY} = \frac{S_{XY}}{S_X S_Y}.$$

Coeficiente de correlación lineal

Coeficiente de correlación lineal entre X e Y

$$r_{XY} = \frac{S_{XY}}{S_X S_Y}.$$

- La correlación lineal toma valores entre -1 y 1 y sirve para investigar la relación lineal entre las variables.

Coeficiente de correlación lineal

Coeficiente de correlación lineal entre X e Y

$$r_{XY} = \frac{S_{XY}}{S_X S_Y}.$$

- La correlación lineal toma valores entre -1 y 1 y sirve para investigar la relación lineal entre las variables.
- Si toma valores cercanos a -1 diremos que hay una relación inversa entre X e Y.

Coeficiente de correlación lineal

Coeficiente de correlación lineal entre X e Y

$$r_{XY} = \frac{S_{XY}}{S_X S_Y}.$$

- La correlación lineal toma valores entre -1 y 1 y sirve para investigar la relación lineal entre las variables.
- Si toma valores cercanos a -1 diremos que hay una relación inversa entre X e Y.
- Si toma valores cercanos a $+1$ diremos que hay una relación directa entre X e Y.

Coeficiente de correlación lineal

Coeficiente de correlación lineal entre X e Y

$$r_{XY} = \frac{S_{XY}}{S_X S_Y}.$$

- La correlación lineal toma valores entre -1 y 1 y sirve para investigar la relación lineal entre las variables.
- Si toma valores cercanos a -1 diremos que hay una relación inversa entre X e Y.
- Si toma valores cercanos a $+1$ diremos que hay una relación directa entre X e Y.
- Si toma valores cercanos a cero diremos que no existe relación lineal entre X e Y.

Ejemplo Volumen Expiratorio Forzado y estatura: correlación

Para los datos del ejemplo sobre el VEF y la estatura se obtiene que:

- La desviación típica de la estatura es $S_x = 13.808$ centímetros.
- La desviación típica del VEF es $S_y = 0.717$ litros.
- El coeficiente de correlación lineal entre X e Y será

$$r_{XY} = \frac{9.783}{13.808 \cdot 0.717} = 0.988$$

- La correlación es próxima a 1 y por lo tanto la relación entre ambas variables es directa.

Modelo de regresión lineal

- El tipo de relación más sencilla que se establece entre un par de variables es la **relación lineal** $Y = \beta_0 + \beta_1 X$
- Sin embargo, este modelo supone que una vez determinados los valores de los parámetros β_0 y β_1 es posible predecir exactamente la respuesta Y dado cualquier valor de la variable de entrada X.
- En la práctica tal precisión casi nunca es alcanzable, de modo que lo máximo que se puede esperar es que la ecuación anterior sea válida sujeta a un error aleatorio, es decir, la relación entre la **variable dependiente** (Y) y la **variable regresora** (X) se articula mediante una **recta de regresión**.

Modelo de regresión lineal simple

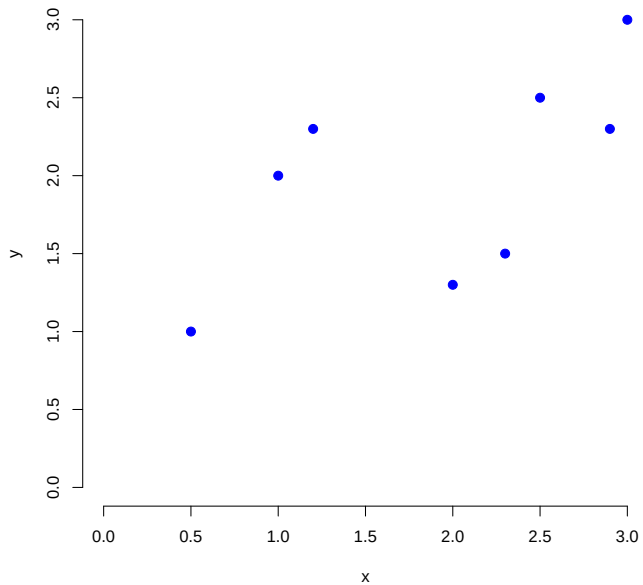
$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

Modelo de regresión lineal

Modelo de regresión lineal simple

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

- Dada una muestra $(x_1,y_1),\ldots,(x_n,y_n)$ de la variable bidimensional (X, Y), ¿Cuál es la recta que mejor ajusta los datos?



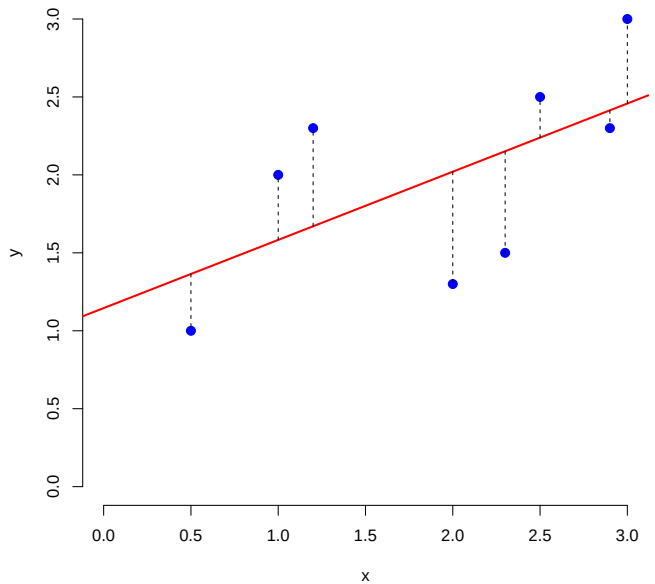
- El objetivo es determinar los valores de los parámetros desconocidos β_0 y β_1 (mediante estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$) de manera que la recta definida ajuste de la mejor forma posible a los datos.

El método de mínimos cuadrados

- El **método de mínimos cuadrados** consiste en encontrar los valores $\hat{\beta}_0$ y $\hat{\beta}_1$ que, dada la muestra de partida, minimizan la suma de los errores al cuadrado.
- Los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ se determinan minimizando las **distancias verticales** entre los puntos observados, y_i, y las ordenadas previstas por la recta para dichos puntos $\hat{y}_i$

El método de mínimos cuadrados

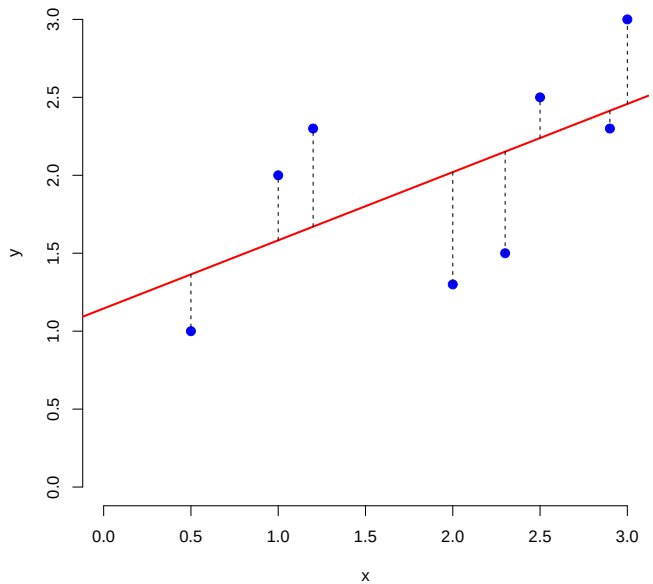
$$M(\beta_0, \beta_1) = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2.$$



El método de mínimos cuadrados

Coeficientes estimados por el método de mínimos cuadrados

$$\hat{\beta}_1 = \frac{S_{XY}}{S_X^2} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$



Recta de regresión de Y sobre X

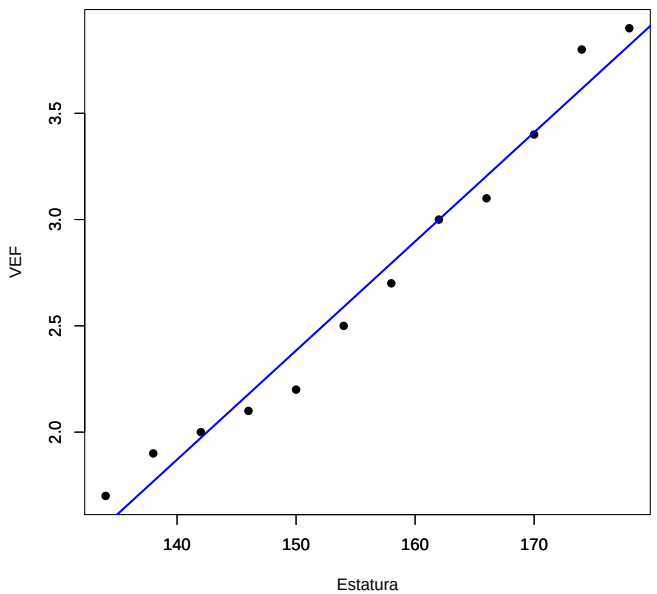
$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

Ejemplo. Volumen Expiratorio Forzado y estatura: recta de regresión

Para los datos del ejemplo sobre el VEF y la estatura se obtiene que:

- $\hat{\beta}_1 = \frac{9.783}{13.808^2} = 0.0513$
- $\hat{\beta}_0 = 2.691 - 156 \cdot 0.0513 = -5.312$
- La recta de regresión será entonces

$$y = \hat{\beta}_0 + \hat{\beta}_1 x = -5.312 + 0.0513x$$



Valores observados y previstos: error en la predicción

- La recta de regresión se construye a partir de las n observaciones de las variables X e Y

$$x_1, \ldots, x_n \text{ y } y_1, \ldots, y_n$$

ya que

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \text{ para todo } i = 1, \ldots, n.$$

- Sin embargo, obtenida la recta de regresión de Y sobre X, podemos predecir las observaciones de Y a partir de las de X. Esto es,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i,$$

para todo $i = 1, \ldots, n$.

- Dados los valores de X inicialmente considerados, los valores que se predicen de Y ($\hat{y}$) no coinciden con los valores inicialmente observados.

- Por lo tanto, existe un error de predicción:

$$\text{residuo} = \text{valor observado - valor de la recta} = y_i - \hat{y}_i$$

Descomposición de la variabilidad

- La variabilidad de toda la muestra se denomina **variabilidad total (VT)**.

$$\text{VT} = \sum_{i=1}^{n} (y_i - \bar{y})^2 .$$

- La variabilidad total se descompone en dos sumandos:
  - La variabilidad explicada (VE).

$$\text{VE} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 .$$

  - La variabilidad no explicada (VNE) por la regresión.

$$\text{VNE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 .$$

Descomposición de la variabilidad

$$\text{VT} = \text{VE} + \text{VNE}.$$

Coeficiente de determinación

- El **coeficiente de determinación** (R^2) se define como la proporción de variabilidad de la variable dependiente que es explicada por la regresión

Coeficiente de determinación

$$R^2 = \frac{\text{VE}}{\text{VT}} = 1 - \frac{\text{VNE}}{\text{VT}}.$$

- En el modelo de regresión lineal simple, el coeficiente de determinación coincide con el cuadrado del coeficiente de correlación.

$$R^2 = r_{XY}^2$$

Ejemplo. Volumen Expiratorio Forzado y estatura: coeficiente de determinación

Para los datos del ejemplo sobre el VEF y la estatura se obtiene que:

- $R^2 = 0.9881^2 = 0.976$
- Con el modelo de regresión lineal simple hallado, la variable X es capaz de explicar el 97.6 % de la variación de Y.