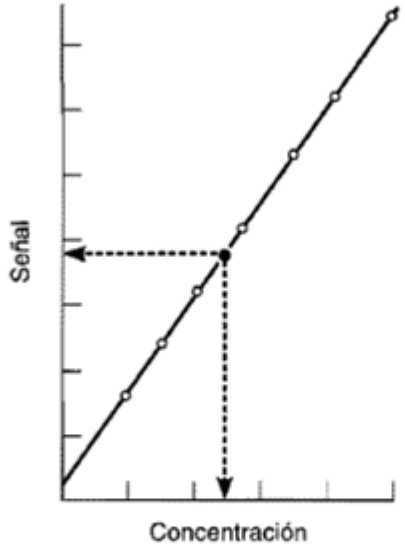
REGRESIÓN LINEAL SIMPLE

Los métodos de regresión se usan para estudiar la relación entre dos variables numéricas. Este tipo de problemas aparecen con frecuencia en el contexto de química analítica cuando se desea realizar el calibrado en análisis instrumental.
El procedimiento habitual es el siguiente: el analista toma una serie de materiales (pueden ser 3 ó 4 ó más aún) en los que conoce la concentración del analito. Estos patrones de calibración se miden con el instrumento analítico en las mismas condiciones en las que se trabajará en los ensayos con el material desconocido. Una vez establecido el gráfico de calibrado puede obtenerse la concentración del analito como se muestra en el siguiente gráfico:



Procedimiento de calibración en análisis instrumental: ○ puntos de calibrado; ● muestra de ensayo.

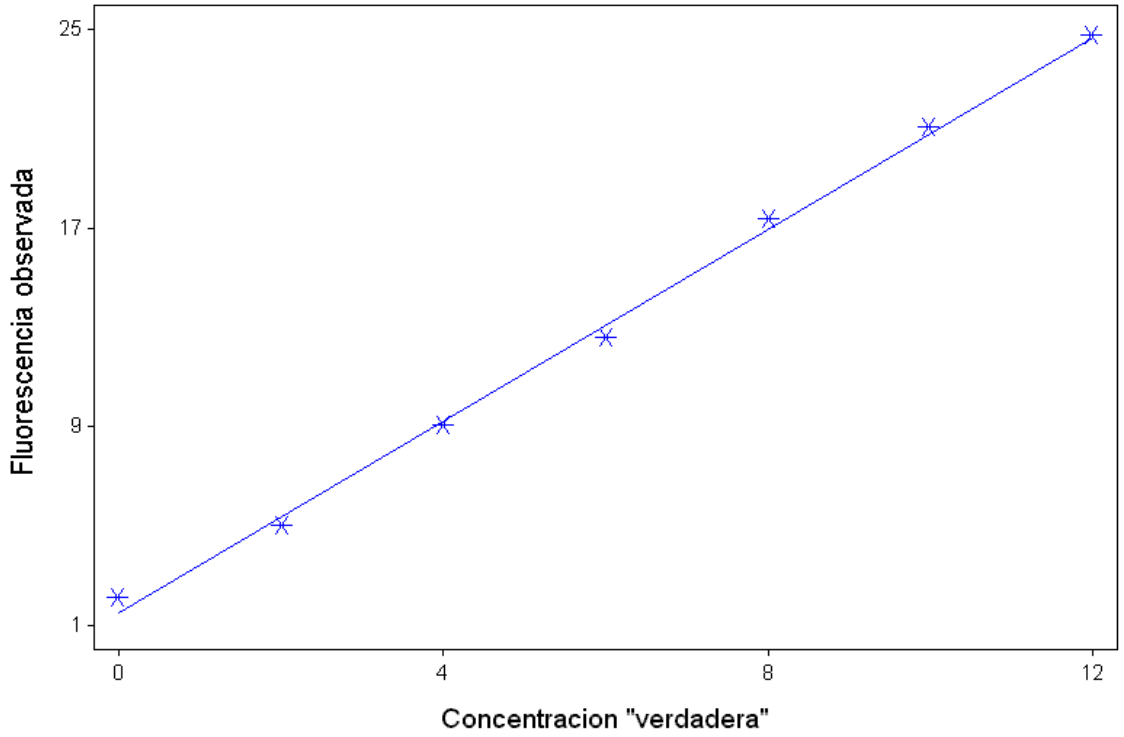
Veamos un ejemplo numérico:

Ejemplo 1: Para calibrar un fluorímetro se han examinado 7 soluciones estándar de fluoresceína (de las que se conoce la concentración medida con mucha precisión) en el fluorímetro. Los siguientes datos corresponden a las *concentraciones* y la *intensidad de fluorescencia* observada en el fluorímetro:

Concentración (pg/ml):	0	2	4	6	8	10	12
Intensidad de fluorescencia:	2.1	5.0	9.0	12.6	17.3	21.0	24.7

En un problema de calibración, queremos, a partir de mediciones hechas en muestras estándar, estudiar la relación entre las mediciones y el “verdadero valor”. Esta relación permitirá en el futuro, medir una muestra desconocida y conocer aproximadamente su valor verdadero.

Lo primero que se hace para estudiar la relación entre dos variables numéricas es un diagrama de dispersión (scatter plot), como el que se presenta a continuación.



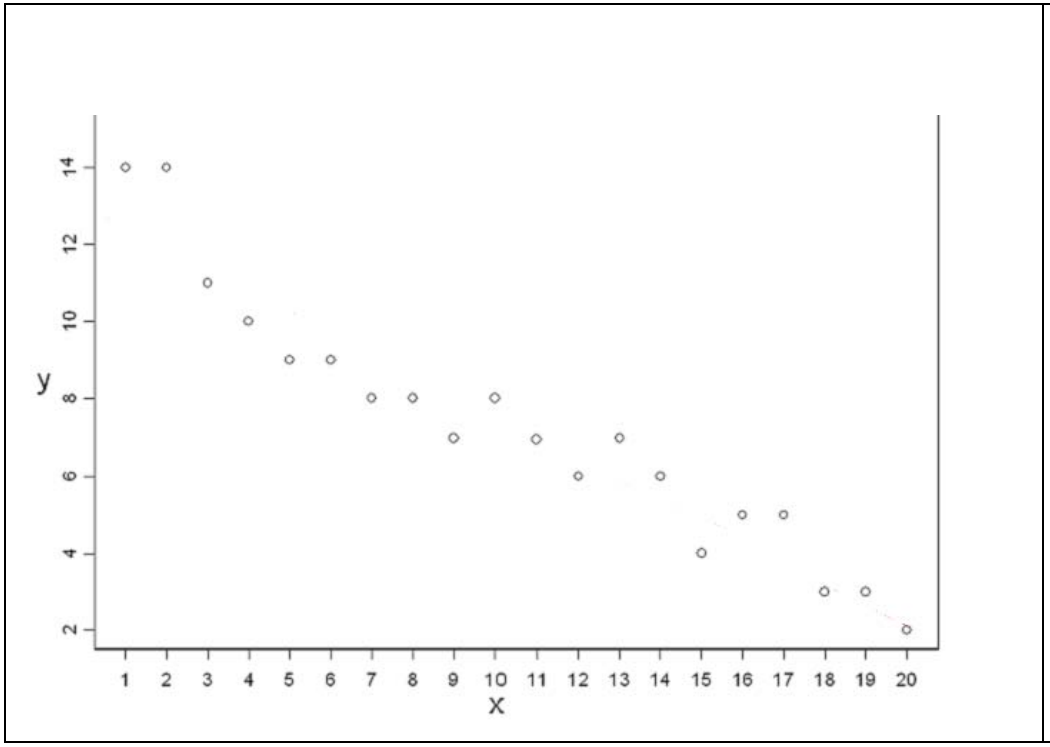
Para ayudar a visualizar la relación, hemos agregado a los puntos del gráfico de dispersión una recta que se llama "recta de regresión" o "recta de cuadrados mínimos". Veremos cómo hallar esta recta.

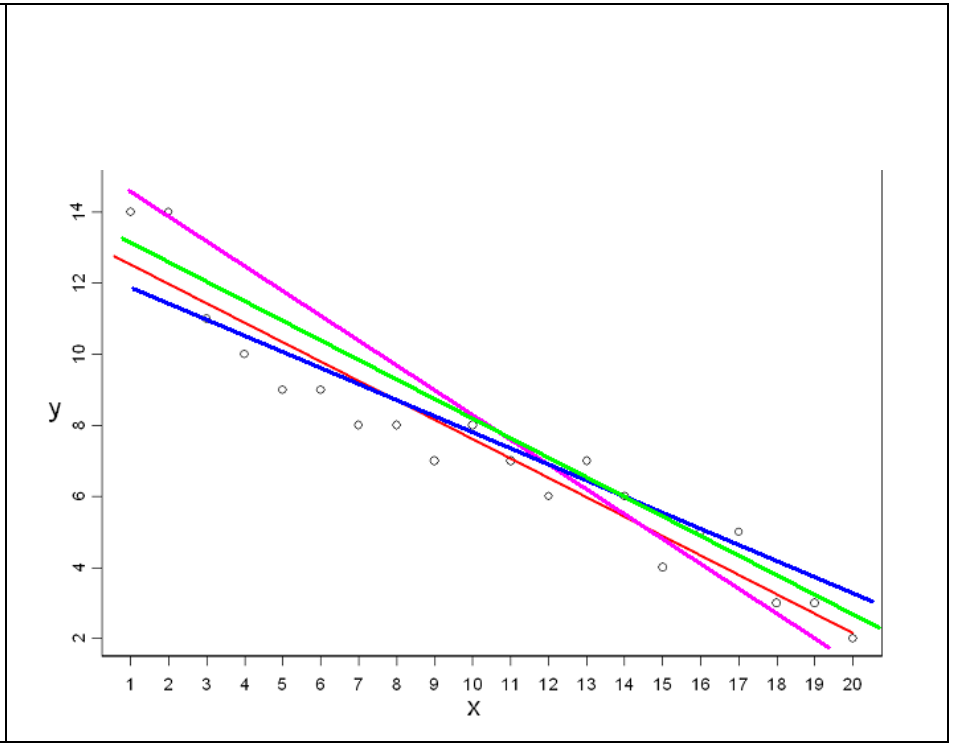
Recordemos que la ecuación de una recta es de la forma

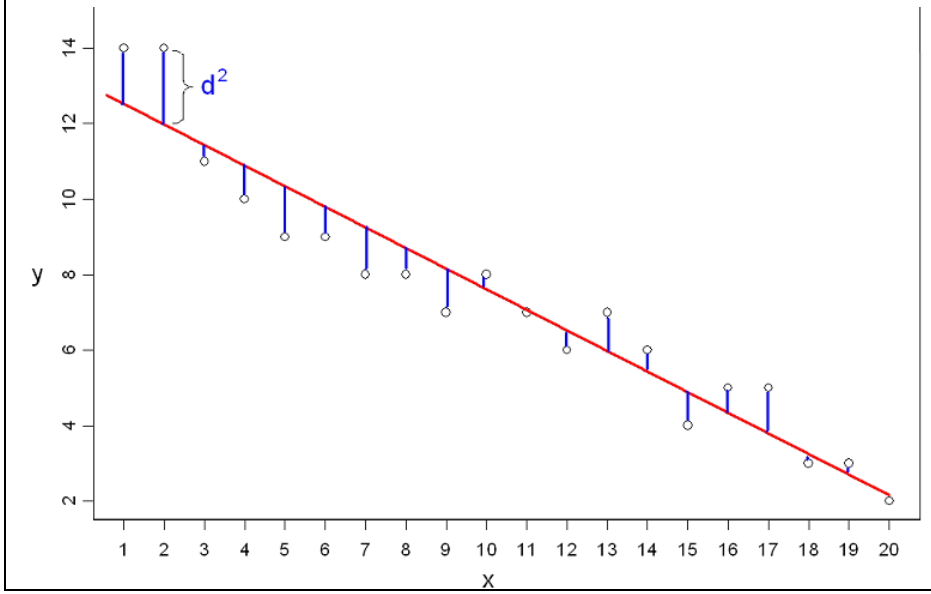
$$y = \alpha + \beta\, x$$

Ordenada al origen **Pendiente**

¿Cómo elegimos la recta que mejor ajusta a los datos?







d_i^2 : distancia del i - ésimo punto a la recta

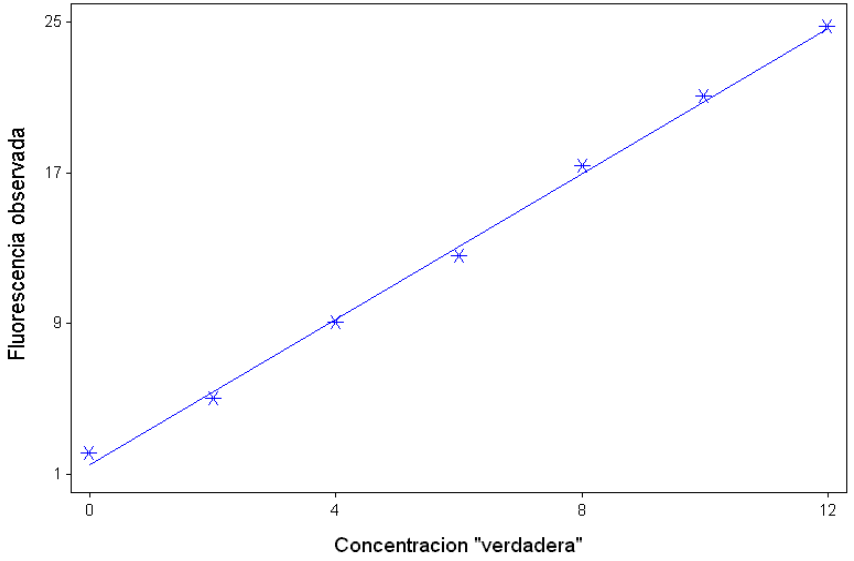
Buscamos la recta que minimiza

$$\left(\frac{d_1^2 + d_2^2 + ... + d_n^2}{n}\right)$$

Recta de cuadrados mínimos.

La recta representada en el gráfico anterior es *la recta de cuadrados mínimos*. Esta es la recta que está "más cerca" de los puntos, en el sentido siguiente: hace mínima la suma de los cuadrados de las distancias de cada punto a la recta, midiendo las distancias verticalmente. O sea, minimiza:

$$\sum_{i=1}^{n} (y_i - (a + bx_i))^2 \qquad (1)$$



PREDICTOR VARIABLES	COEFFICIENT	STD ERROR	STUDENT'S T	P
CONSTANT	**1.51786**	0.29494	5.15	0.0036
CONCENTRA	**1.93036**	0.04090	47.20	0.0000

Observando los "coeficientes" de la salida vemos que la recta de mínimos cuadrados tiene ordenada al origen **1.51786** y pendiente **1.93036**. Si los puntos (como en este ejemplo) están cerca de la recta, podemos decir que

$$y \cong 1.51786 + 1.93036\, X$$

o

$$\text{Fluorescencia} \cong 1.51786 + 1.93036\ \text{Concentración}$$

Por ejemplo, si la concentración de fluoresceína de una muestra fuera 8, la ordenada de acuerdo con esta recta seria

$$1.51786 + 1.93036 * 8 = 16.96.$$

Esto no quiere decir que para la muestras que tengan concentración=8 la intensidad de la fluorescencia es exactamente 16.96 (como se ve en el gráfico, los puntos están muy cerca de la recta, pero no están sobre la recta).

Modelo de regresión lineal

Para hacer inferencia, es decir realizar tests de hipótesis o calcular intervalos de confianza, se necesita suponer un modelo, que llamaremos "*modelo de regresión lineal simple*".

La palabra "simple" se debe a que consideramos una sola variable independiente o predictora (X). Se generaliza en forma natural al caso en que hay más variables independientes y en ese caso se llama "modelo de regresión lineal múltiple".
Las suposiciones del modelo de regresión lineal simple son las siguientes.

MODELO: Se observan pares de valores (x_i, y_i) para i=1,...n, que cumplen:

$$y_i = \alpha + \beta\ x_i + e_i \qquad (2)$$

donde $e_1, e_2, ..., e_n$ son variables aleatorias tales que

1) $E(e_i) = 0$ para todo i
2) $Var(e_i) = \sigma^2$
3) $e_1, e_2, ..., e_n$ son v. a. independientes

Para obtener algunos resultados alcanzan las suposiciones 1) a 3), pero para obtener tests e intervalos es necesario agregar algo más:

4) $e_i \sim$ Normal

Obviamente las suposiciones 1) a 4) se pueden escribir en forma más breve:

$$1) \text{ a } 4) \iff e_i \sim N(0, \sigma^2) \text{ i.i.d.}$$

Observación:

Supongamos que se cumple (2). Hay dos modelos un poco diferentes: el modelo con x_i's fijas y el modelo con x_i's aleatorias.

En el primero los valores x_i's no son variables aleatorias sino que son números fijados por el experimentador. En el segundo tanto x_i como y_i son observaciones de variables aleatorias. Los problemas de calibración son ejemplo con x_i's fijas.

En otras situaciones como podría ser en un problema en el que se desea estudiar la relación entre estatura y perímetro cefálico de recién nacidos, las covariables x_i's son aleatorias.

Justificaremos los resultados sobre estimadores, IC y tests sólo para el modelo con x_i's fijas, que es más simple, pero casi todos estos resultados son los mismos para ambos modelos.

Una forma equivalente de escribir el modelo de regresión lineal simple (en el caso en que las x_i's son números fijos) es la siguiente:

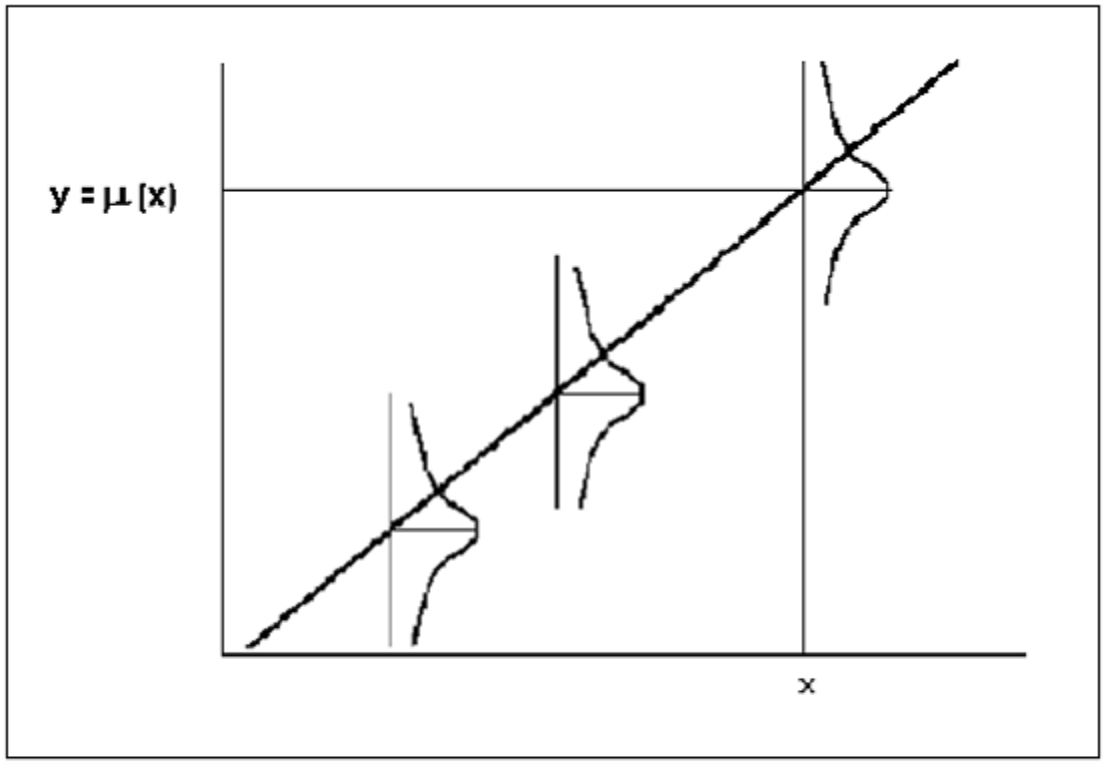
1*) $E(y_i) = \alpha + \beta x_i$ (para i=1,...,n)
2*) $Var(y_i) = \sigma^2$ (para i=1,...,n)
3*) $y_1, y_2, ..., y_n$ son v. a. independientes
4*) $y_i \sim$ Normal

Nuevamente, las suposiciones 1*) a 4*) se pueden escribir en forma más breve:

$$1^*) \text{ a } 4^*) \Leftrightarrow y_i \sim N(\alpha + \beta x_i, \sigma^2) \text{ independientes}$$

Observación: en el modelo con x_i's aleatorias, no hay que hacer ninguna suposición sobre la distribución de las x_i's . Puede ser normal o no.

Como de costumbre, no se espera que las suposiciones del modelo se cumplan exactamente en un problema real, pero al menos que sean aproximadamente válidas. Si están lejos de cumplirse, las conclusiones pueden ser erróneas. Por ejemplo, la presencia de algunos valores de la respuestas y_i atípicos (alejados de la recta, lo que implica que no se cumple la suposición 4) pueden invalidar las conclusiones. En efecto, la recta de cuadrados mínimos, al igual que la media, es sensible a unos pocos valores atípicos.



La figura representa dos variables para las cuales se satisfacen los supuestos de linealidad ($\mu(x) = \alpha + \beta x$, la media de la variable *Y* crece linealmente con *x*), normalidad y homoscedasticidad de los errores .

Estimadores de α y β por el método de cuadrados mínimos

Llamemos $\hat{\alpha}$ y $\hat{\beta}$ a los valores de a y b que minimizan (1) que se llaman "*estimadores de cuadrados mínimimos*" de α y β.

¿Cómo hallamos a y b ?

$$\frac{\partial\left(\sum_{i=1}^{n}(y_i - a - bx_i)^2\right)}{\partial a} = -2\sum_{i=1}^{n}(y_i - a - bx_i) = 0$$

$$\frac{\partial\left(\sum_{i=1}^{n}(y_i - a - bx_i)^2\right)}{\partial b} = -2\sum_{i=1}^{n}x_i(y_i - a - bx_i) = 0$$

Mostraremos resolviendo estas ecuaciones que

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{(\sum_{i=1}^{n} x_i y_i) - n\,\bar{x}\,\bar{y}}{(\sum_{i=1}^{n} x_i^2) - n\,\bar{x}^2} \qquad (3)$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x} \qquad (4)$$

La ecuación (4) nos dice que la recta de mínimos cuadrados pasa por $(\bar{x}, \bar{y})$, ya que

$$\bar{y} = \hat{\alpha} + \hat{\beta}\,\bar{x}$$

Probaremos que estos estimadores dados en (3) y (4) son insesgados bajo la condición 1), es decir $E(\hat{\alpha}) = \alpha$ y $E(\hat{\beta}) = \beta$.

Además, se puede demostrar que estos estimadores son óptimos si se cumplen las suposiciones 1) a 4).

Residuos: Se llaman residuos las diferencias entre los valores observados y las respectivas ordenadas de la recta:

$$\hat{e}_i = y_i - (\hat{\alpha} + \hat{\beta}\, x_i)$$

Valores predichos: Llamamos valores predichos a

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

Estimador de σ^2: σ^2 es la varianza de e_i , es decir σ^2 = Var(e_i). Los e_i son v. a. "no observables". Parece natural que el estimador de σ^2 se base en los residuos $\hat{e}_i$. Se puede demostrar que el estimador

$$s^2 = \frac{\sum_{i=1}^{n} \hat{e}_i^2}{n-2} = \frac{\sum_{i=1}^{n} (y_i - (\hat{\alpha} + \hat{\beta}\, x_i))^2}{n-2} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2} \qquad (5)$$

es un estimador insesgado de σ^2.

Varianza de $\hat{\alpha}$ y : $\hat{\beta}$:

Se puede demostrar que:

$$Var(\hat{\alpha}) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] \qquad Var(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad (6)$$

y además

$$\mathrm{cov}(\bar{y}, \hat{\beta}) = 0 \qquad (7)$$

Los estimadores de Var($\hat{\alpha}$) y Var($\hat{\beta}$) se obtienen reemplazando σ^2 por s^2.

Intervalo de confianza para β

Llamemos

$$ES(\hat{\beta}) = \sqrt{\hat{V}ar(\hat{\beta})} = \sqrt{\frac{s^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Si la suposición 4) de normalidad se cumple, el intervalo

$$\hat{\beta} \pm t_{n-2;\alpha/2} \sqrt{\frac{s^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \qquad (8)$$

es un IC para β con nivel 1-α.

Una medida de cuán buena es X para predecir Y:
el coeficiente de correlación lineal "r" de Pearson.

Este coeficiente puede interpretarse como una medida de cuán cerca están los puntos de una recta. La definición de r^2 es la siguiente:

$$r^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 \quad - \quad \sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (9)$$

Puede observarse que r^2 compara la dispersión de los valores de y con respecto a la recta de cuadrados mínimos con la dispersión de los valores de y con respecto a su media.

r^2 es la proporción de la "variación total" entre los valores de y que se puede explicar prediciéndolos por un recta en función de los valores de x.

Puede demostrarse que

$$r^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Se cumple que

$$0 \leq r^2 \leq 1$$

Significado del valor de r^2

$r^2 = 1$	significa que los puntos están exactamente sobre una recta (*)
r^2 cerca de 1	los puntos están cerca de una recta
r^2 cerca de 0	significa que la recta de cuadrados mínimos es prácticamente horizontal y por lo tanto no hay relación creciente ni decreciente.

(*) En las aplicaciones prácticas es "casi imposible" que r^2 valga exactamente igual a 1.

El coeficiente de correlación r es la raiz de r^2 y se le pone signo negativo si la pendiente de la recta de cuadrados mínimos es negativa (recta decreciente).
Otra expresión equivalente para calcular r es:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} \qquad (10)$$

Siempre es

$$-1 \leq r \leq 1$$

y r cerca de 1 o -1 indicará que los puntos están cerca de una recta creciente o decreciente respectivamente.

Veamos todo esto en el ejemplo.

```
concentra=c(0,2,4,6,8,10,12)
fluo<- c(2.1,5,9,12.6,17.3,21,24.7)
salida<- lm(fluo~concentra)
summary(salida)

Call:
lm(formula = fluo ~ concentra)
Residuals:
1 2 3 4 5 6 7
0.58214 -0.37857 -0.23929 -0.50000 0.33929 0.17857 0.01786

Coefficients:
             Estimate Std. Error   t value    Pr(>|t|)
(Intercept)  1.5179    0.2949       5.146    0.00363 **
concentra    1.9304    0.0409      47.197    8.07e-08 ***
---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4328 on 5 degrees of freedom
Multiple R-squared: 0.9978, Adjusted R-squared: 0.9973
F-statistic:   2228   on   1   and   5   DF,   p-value:   8.066e-08
```

En el ejemplo de la fluorescencia, tenemos que

R-SQUARED 0.9978

y, como la pendiente es positiva, es $r = (0.9978)^{½} = 0.9989$. Ambos muy cerca de 1, son una medida de lo que vemos en el gráfico: los puntos están muy cerca de una recta.

En el caso en que las x_i's son aleatorias, el coeficiente r es un estimador consistente del coeficiente de correlación $\rho(x,y)$.