Métodos cuantitativos mediante Regresión y correlación

Contenidos:

- ✓ Dependencia funcional o exacta y dependencia estadística
- ✓ Concepto de regresión
- ✓ Método de mínimos cuadrados
- ✓ Análisis de la bondad de ajuste. Error cuadrático medio, varianza residual y coeficiente de determinación lineal

Independencia - Dependencia

Cuando se estudian dos características simultáneamente sobre una muestra, se puede considerar que **una de ellas influye sobre la otra** de alguna manera. Por ejemplo la altura y el peso o las horas de estudio y la calificación en un examen.

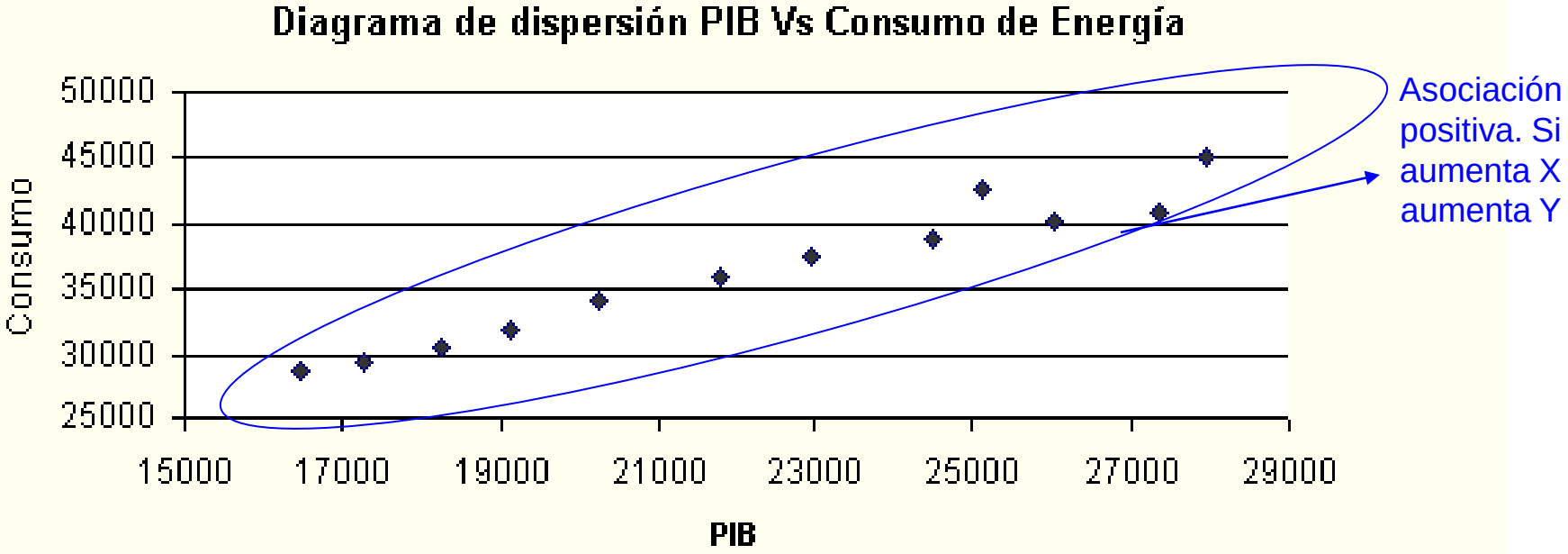
El **objetivo** principal de la regresión es descubrir el modo en que se relacionan.

Dos variables pueden considerarse:

- Variables independientes → No tienen relación (una de ellas no sirve para explicar los movimientos de la otra)
- Dependencia funcional → Y=f(x)
- Dependencia estadística

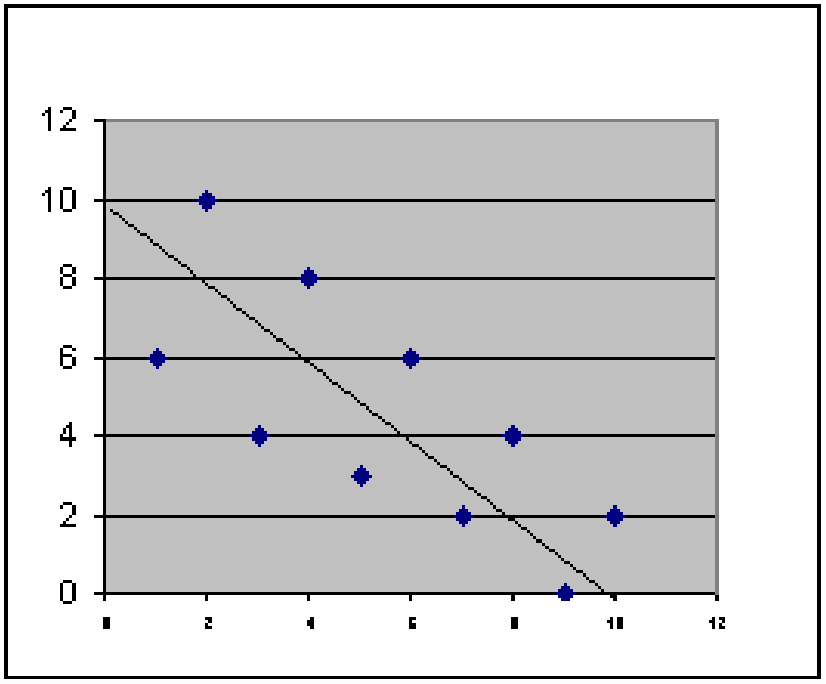
GRÁFICOS DE DISPERSIÓN: Permite ver si hay asociación

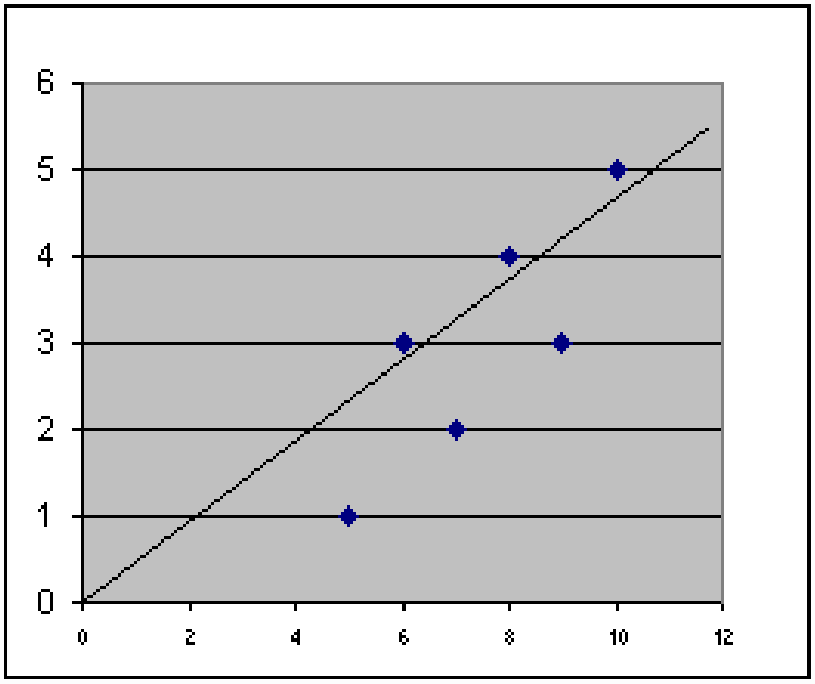
Dadas dos variables X y Y tomadas sobre el mismo elemento de la población, el diagrama de dispersión es simplemente un gráfico de dos dimensiones, donde en un eje (la abscisa) se sitúa una variable, y en el otro eje (la ordenada) se sitúa la otra variable. Si las variables están correlacionadas, el gráfico mostraría algún nivel de correlación (tendencia) entre las dos variables. Si no hay ninguna correlación, el gráfico presentaría una figura sin forma, una nube de puntos dispersos en el gráfico.



GRÁFICOS DE DISPERSIÓN / RECTA DE REGRESIÓN

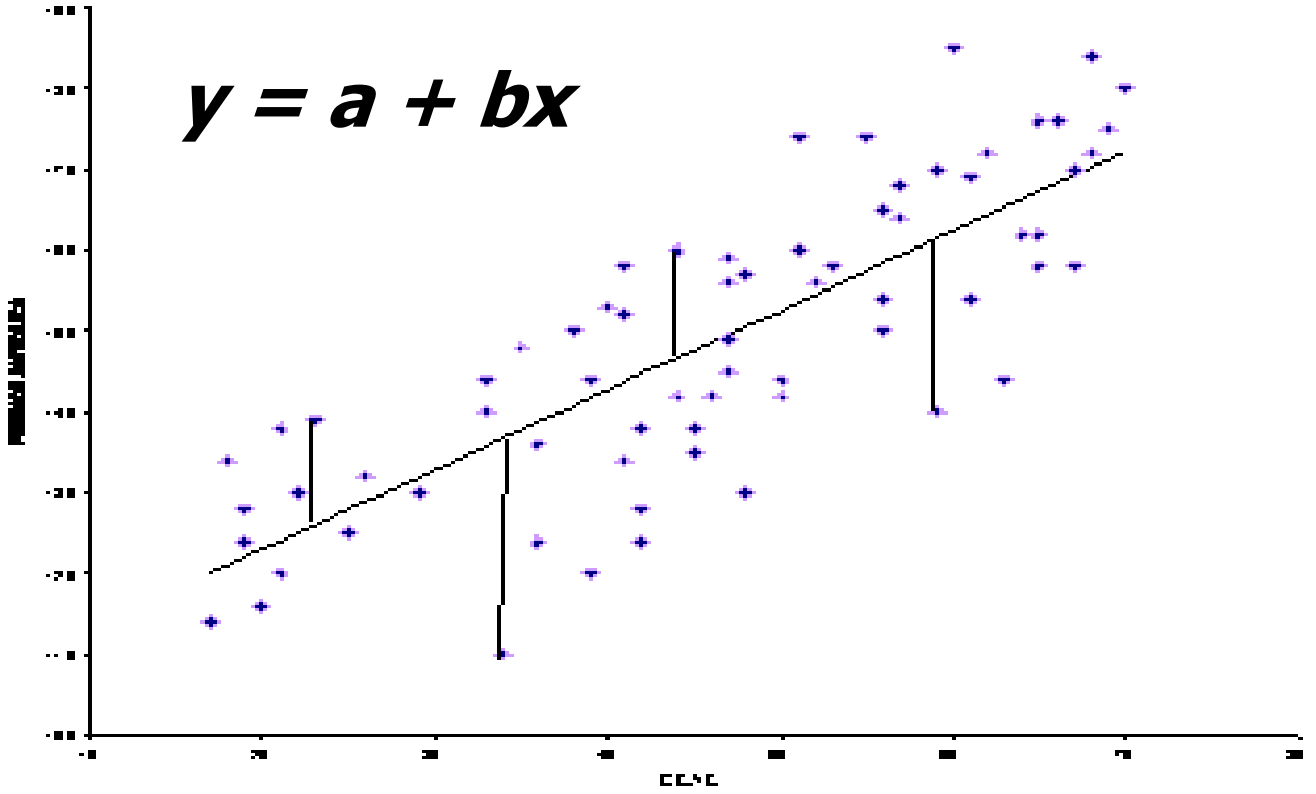
La relación entre dos variables métricas puede ser representada mediante la línea de mejor ajuste a los datos. Esta recta se le denomina recta de regresión, que puede ser negativa o positiva, la primera con tendencia decreciente y la segunda creciente.



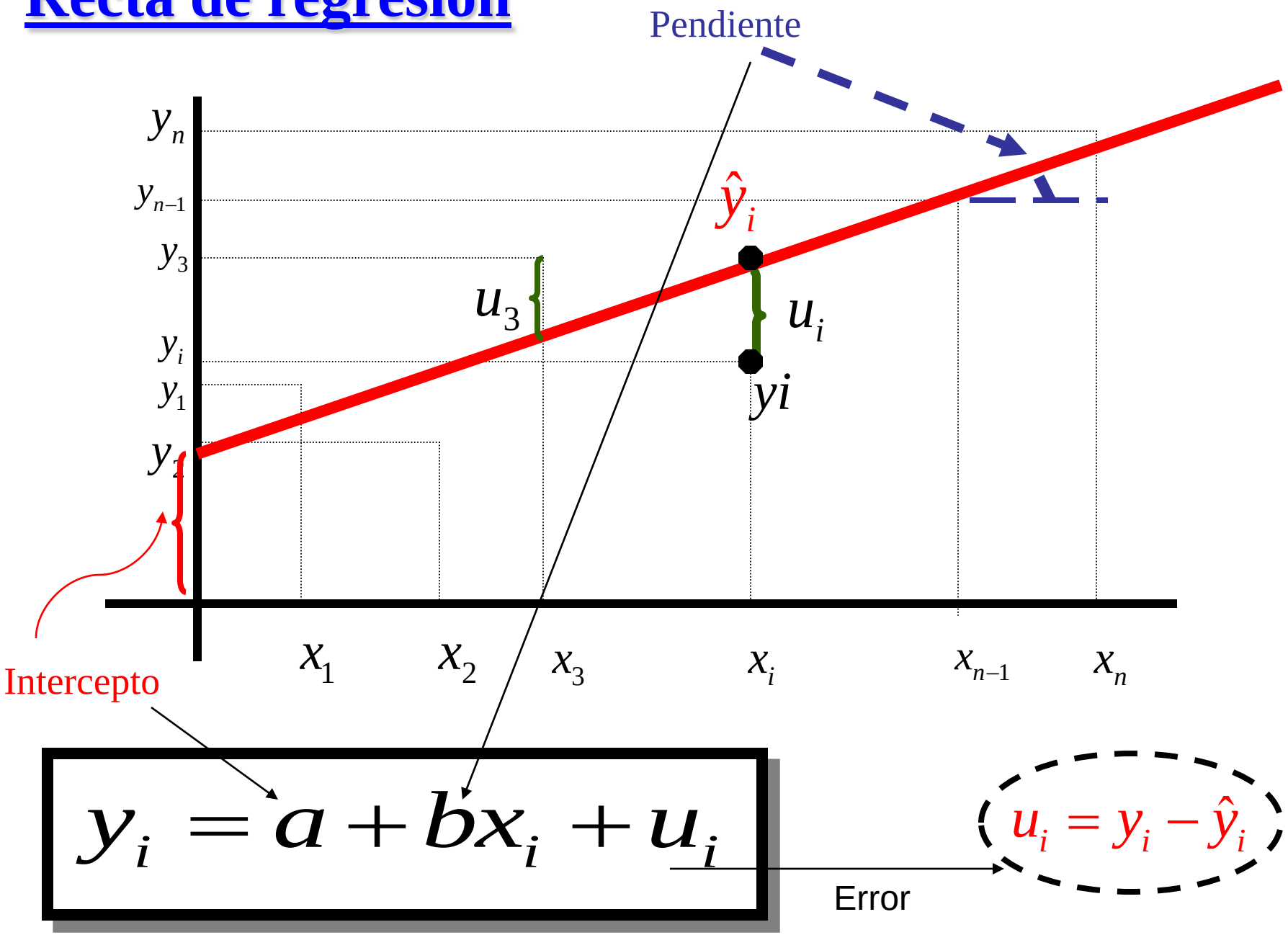


GRÁFICOS DE DISPERSIÓN / RECTA DE REGRESIÓN

Para el cálculo de la recta de regresión se aplica el método de mínimos cuadrados entre dos variables. Esta línea es la que hace mínima la suma de los cuadrados de los residuos, es decir, es aquella recta en la que las diferencias elevadas al cuadrado entre los valores calculados por la ecuación de la recta y los valores reales de la serie, son las menores posibles.



Recta de regresión



Llamemos a "u" perturbación o error, siendo la diferencia que hay entre el valor observado de la variable exógena (y) y el valor estimado que obtendremos a través de la recta de regresión $\hat{y}_i$.

$$\hat{y}_i = a + bx_i$$

La metodología para la obtención de la recta será hacer **MÍNIMA** la **suma** de los **CUADRADOS** de las perturbaciones. ¿Por qué se elevan al cuadrado?

$$u_i^2 = (y_i - \hat{y}_i)^2 \longrightarrow \sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\min_{q,p} \left[\sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left[y_i - (a + b\, x_i) \right]^2 \right]$$

En el modelo de regresión lineal simple la función elegida para aproximar la relación entre las variables es una recta, es decir y=a+bx, donde a,b son los parámetros. A esta recta la llamaremos ***RECTA DE REGRESIÓN DE Y SOBRE X.***

Vamos a deducir su ecuación usando el método de los mínimos cuadrados. Dado un valor de X, tenemos los dos valores de Y, el observado, yi , y el teórico, yi* = a + bxi. **Hemos de minimizar los errores cometidos:**

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

MINIMIZAR

Errores cometidos al aproximar por una recta

El valor que hemos aproximado para "y" con la recta de regresión → y*

$$\left.\begin{aligned}\frac{\partial\Psi}{\partial a}&=-2\sum_i\left(y_i-a-bx_i\right)=0\\ \frac{\partial\Psi}{\partial b}&=-2\sum_i\left(y_i-a-bx_i\right)x_i=0\end{aligned}\right\}\quad\left\{\begin{aligned}\sum_i y_i&=\sum_i a+b\sum_i x_i\\ \sum_i x_i y_i&=a\sum_i x_i+b\sum_i x_i^2\end{aligned}\right.$$

$$-na=-\sum_i y_i+b\sum_i x_i\longrightarrow \underline{\underline{a=\bar{y}-b\bar{x}}}$$

$$\sum_i x_i y_i=(\bar{y}-b\bar{x})\sum_i x_i+b\sum_i x_i^2$$

$$\sum_i x_i y_i=\frac{\sum y_i}{n}\sum_i x_i-b\bar{x}n\bar{x}+b\sum_i x_i^2$$

$$\sum_i x_i y_i-\bar{y}n\bar{x}=b\left(\sum_i x_i^2-n\bar{x}^2\right)$$

$$S_{xy}=bS_x^2\longrightarrow \underline{\underline{b=\frac{S_{xy}}{S_x^2}}}$$

y obtenemos que la recta de regresión de Y sobre X: **y = a + bx** con los valores a y b anteriormente calculados, o bien la siguiente expresión:

$$y-\bar{y}=\frac{S_{xy}}{S_x^2}\left(x-\bar{x}\right)$$

Aplicando el mismo razonamiento llegaríamos a la expresión de la recta de regresión de X sobre Y: **x = a' + b'y** con los valores a' y b' calculados como:

$$b'=\frac{S_{xy}}{S_y^2}\ \ y\ \ a'=\bar{x}-b'\,\bar{y}$$

Por tanto, se podría expresar como:

$$x-\bar{x}=\frac{S_{xy}}{S_y^2}\left(y-\bar{y}\right)$$

Varianza residual: Ayuda a medir la dependencia.

$$\mathbf{VR} = S_u^2 = S_{R_y}^2 = \frac{\sum (y_i - \hat{y}_i)^2}{N}$$

Si es grande, los residuos, por término medio, serán grandes. Dependencia pequeña y viceversa.

Varianza marginal: Es la varianza total de X o de Y. Si dividimos la varianza residual entre esta se elimina el problema de unidades de medida.

$$S_y^2 \qquad S_x^2 \longrightarrow \frac{S_u^2}{S_y^2} = \frac{VR}{VT_y}$$

Ayuda a determinar la asociación pero en sentido inverso. La mejor medida es R.

Coeficiente de correlación general:

$$R = \sqrt{1 - \frac{S_u^2}{S_Y^2}}$$

Haciendo unas transformaciones se demuestra que r(xy) visto en el capítulo 6 sólo es un caso particular de R $\longrightarrow r_{xy} = R$

Elevado al cuadrado obtenemos el **coeficiente de determinación** que sirve como medida del buen ajuste de la recta de regresión R^2

Cuando solo exista una variable explicativa o independiente y una sola dependiente se cumple: $\longrightarrow R^2 = bb' = \frac{S_{xy}}{S_x^2}\frac{S_{xy}}{S_y^2} = \left(\frac{S_{xy}}{S_x S_y}\right)^2 = r_{xy}^2$

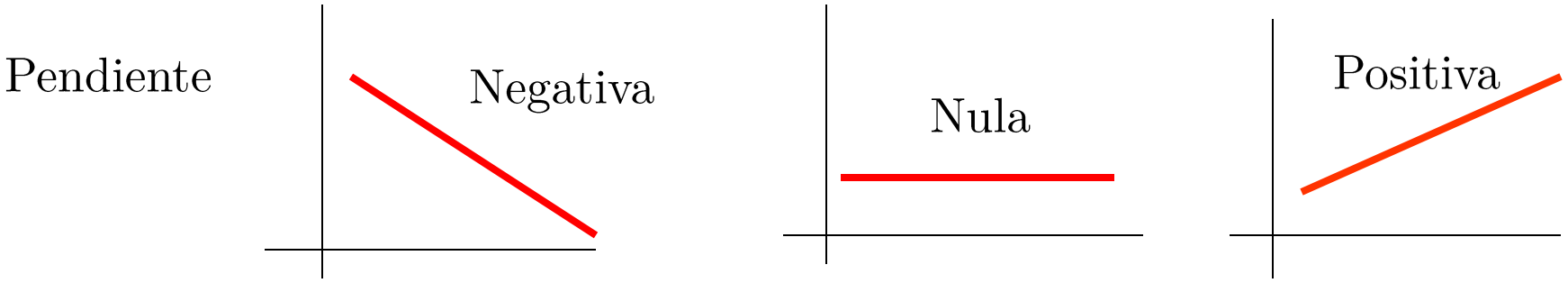
$$-1 \le r \le 1 \quad -1 \le R \le 1 \quad 0 \le r^2 \le 1 \quad 0 \le R^2 \le 1$$

Para el caso de distribuciones bidimensionales: $R = r \Leftrightarrow R^2 = r^2$

Recta de regresión: $\hat{y}_i = \left(\bar{y} - \frac{S_{XY}}{S_X^2} \bar{x} \right) + \frac{S_{XY}}{S_X^2} x_i = \bar{y} + \frac{S_{XY}}{S_X^2} (x_i - \bar{x})$

$$\hat{y}_i = \bar{y} + \frac{S_{XY}}{S_X^2} \frac{S_X}{S_Y} \frac{S_Y}{S_X} (x_i - \bar{x}) = \bar{y} + \frac{S_{XY}}{S_X S_Y} \frac{S_Y}{S_X} (x_i - \bar{x}) = \bar{y} + r \frac{S_Y}{S_X} (x_i - \bar{x})$$

$r = -1$ | $-1 < r < 0$ | $r = 0$ | $0 < r < 1$ | $r = 1$



Minería de datos

- Minería de datos consiste en buscar varias relaciones estadísticas entre variables sin justificativo teórico.

- Por ejemplo, supóngase que se genera una variable dependiente y veinte variables explicativas independientes unas de otras.

- Si se regresa la variable dependiente en función de cada variable explicativa, de forma separada, en promedio el parámetro pendiente será significativo al nivel del 5%.

- Si la minería de datos ocurre, la verdadera significancia será mayor que la significancia nominal.

Estadísiticos de Bondad de ajuste

- Se desea una medición sobre la bondad de ajuste del modelo a los datos.
- El coeficiente de determinación R^2 es el más común. Una manera de definir R^2 es decir la correlación al cuadrado entre y y $\hat{y}$.
- De forma alternativa, se requiere explicar la varianza de la variable endógena, $\bar{y}$, i.e. Por la suma total de cuadrados , STC o *TSS*:

$$STC = \sum_t \left(y_t - \bar{y}\right)^2$$

- Se puede dividir la *STC* en dos partes: 1) la explicada por el modelo (conocido como suma de cuadrados explicada, *SEC*) y la parte que no explica el modelo (*SRC*).

Definiendo R^2

- Esto es, $STC = SEC + SRC$

$$\sum_t (y_t - \bar{y})^2 = \sum_t (\hat{y}_t - \bar{y})^2 + \sum_t \hat{u}_t^2$$

- La bondad de ajuste es:

$$R^2 = \frac{SEC}{STC}$$

- Pero como $STC = SEC + SRC$, se puede escribir:

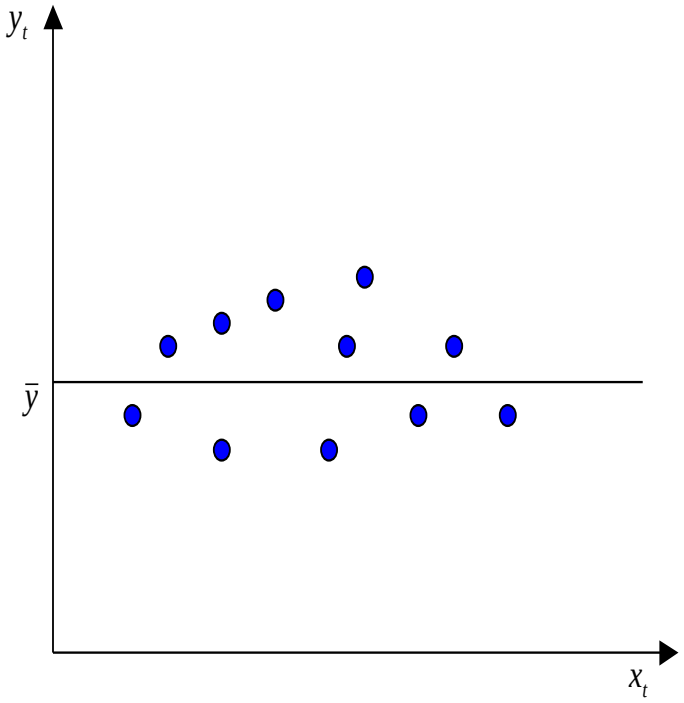
$$R^2 = \frac{SEC}{STC} = \frac{STC - SRC}{STC} = 1 - \frac{SRC}{STC}$$

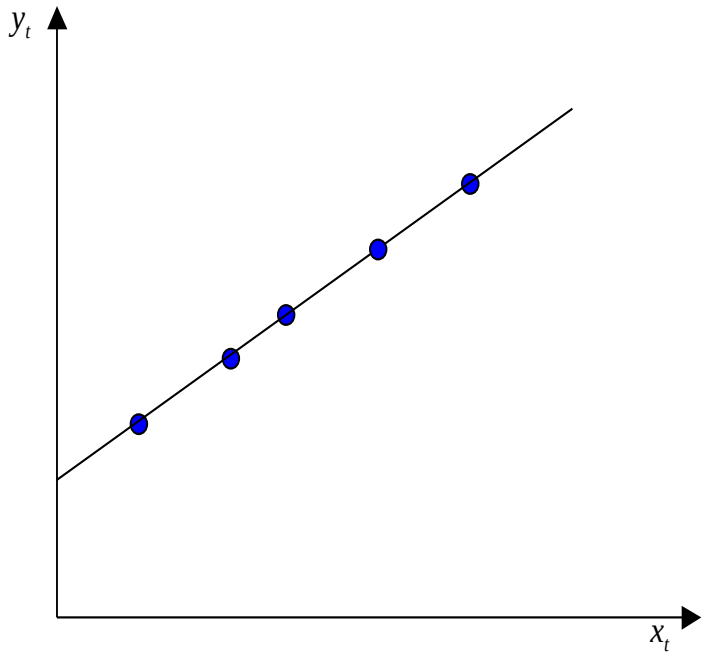
- R^2 siempre estará entre 0 y 1. Considere los dos extremos:

  $SRC = STC$ i.e. $SEC = 0$ así que $R^2 = SEC/STC = 0$

  $SEC = STC$ i.e. $SRC = 0$ así que $R^2 = SEC/STC = 1$

Casos y el límite: $R^2 = 0$ y $R^2 = 1$





Problemas con el R^2 como medida de bondad de ajuste

- Hay un número de problemas:

1. R^2 está definido en términos de la media de y así que el modelo está reparamétrizado (reacomodado) y la variable dependiente siempre cambiará, R^2 cambiará.

2. R^2 nunca cae si se añaden más regresores:

Regresión1: $y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + u_t$

Regresión 2: $y = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \beta_4 x_{4t} + u_t$

R^2 siempre será más alto en la regresión 2 relativa a la regresión1.

3. R^2 es sospecho para series de tiempo mayores a 0.9 en modelos de series de tiempo: puede existir problemas de endogenidad, simultaneidad o operaciones de suma y resta.

R^2 Ajustado

- Para tratar los poblemas previos, se utiliza el R2 ajustado. Se conoce como o $R^{2:}$ ajustado.

$\overline{R}^2$

$$\overline{R}^2 = 1 - \left[\frac{T-1}{T-k} (1 - R^2) \right]$$

- Si se añade un regresor extra, k se incrementa y al menos que R^2 se incremente en la misma proporción, $\overline{R}^2$ caerá.

- Siguen existiendo algunos problemas con este criterio:
  1. Una regla suave.
  2. No hay distribuciones para $\overline{R}^2$ o R2.

Predicción

$$\hat{y}_i = a + b\,x_i = \bar{y} + \frac{S_{XY}}{S_X^2}\left(x_i - \bar{x}\right)$$

El objetivo último de la regresión es la predicción de una variable para un valor determinado de la otra. La predicción de Y para X = x_0 será simplemente el valor obtenido en la recta de regresión de Y sobre X al sustituir el valor de x por x_0. La fiabilidad de esta predicción será tanto mayor cuando mayor sea la correlación entre las variables (es decir mayor sea R2)

Dado un valor de la variable "X" que no ha sido observado, estimar el correspondiente valor de "Y"

Dado x_0 *estimar* $\hat{y}_0$

$$\hat{y}_0 = a' + b\,x_0 = \bar{y} + \frac{S_{XY}}{S_X^2}\left(x_0 - \bar{x}\right)$$