Tema 1. El modelo de regresión lineal simple

1. Introducción

Los métodos de la Matemática que estudian los **fenómenos deterministas** relacionan una variable dependiente con diversas variables independientes:

$$y = g(x_1, x_2, ..., x_k)$$

→ El problema se reduce a resolver un sistema lineal, una ecuación diferencial, un sistema no lineal…

Las Ciencias Experimentales han revelado la **poca fiabilidad de las relaciones deterministas**. En tales Ciencias el azar, la aleatoriedad, la variabilidad individual, las variables no controladas,… justifican el planteo de la ecuación fundamental

"observación" = "modelo" + "error aleatorio"

modelo ← su estructura queda fijada por el experimentador teniendo en cuenta las condiciones de su experimento.

error aleatorio ← el experimentador debe tener en cuenta la desviación que existe entre *lo que observa* y *lo que espera observar* según el modelo.

En los Modelos de Regresión:

Modelo: función lineal de unos parámetros.

$$\underbrace{y_i}_{\text{observación}} = \underbrace{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}}_{\text{modelo}} + \underbrace{u_i}_{\text{error aleatorio}}$$

Objetivo: A partir de una *muestra de tamaño* n de una cierta *población*, determinar los parámetros para poder:

1) hacer extensible el modelo a toda la población,
2) poder predecir nuevos valores de la variable dependiente.

Ejemplo 1. Densidad del tráfico.

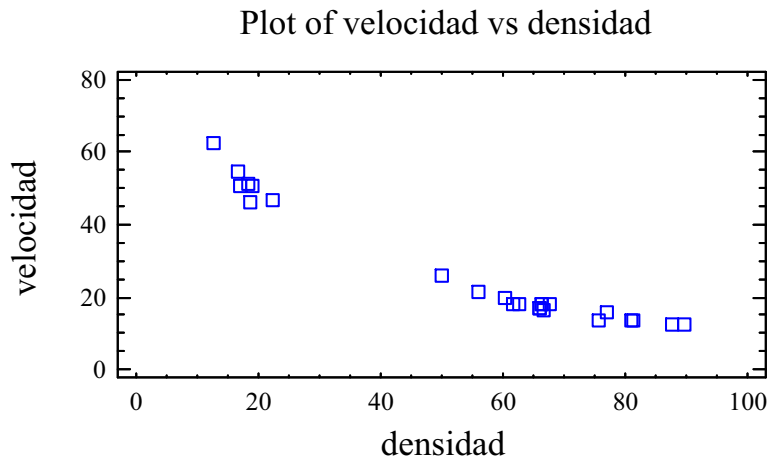
Sabemos que cuántos más coches circulan por una carretera, menor es la velocidad del tráfico.

El estudio de este problema tiene como objetivo la mejora del transporte y la reducción del tiempo de viaje.

Densidad=”número de vehículos por km”
Velocidad=”velocidad del vehículo en km/h”

Como la congestión afecta a la velocidad, estamos interesados en determinar el efecto de la densidad en la velocidad.

Ejemplo 1. Densidad del tráfico.



Modelos que podemos plantear:
(Llamamos Y a la velocidad y X a la densidad)

1. Y= a + b X +"error"
2. Y= a + b X + c X^2 + "error"

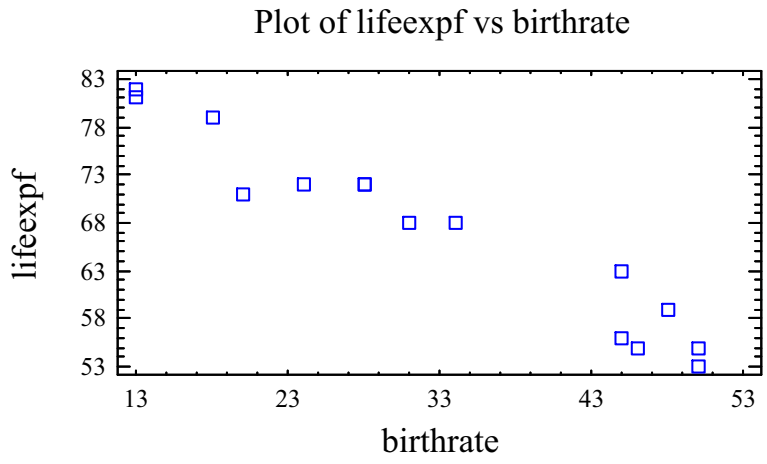
Ejemplo 2. Esperanza de vida.

¿Existe algún tipo de relación entre la esperanza de vida de la mujer y el número de nacimientos cada 1000 habitantes?

lifeexpf= "esperanza de vida de las mujeres"
birthrate= "número de nacimientos cada 1000 habitantes" (datos de 1992)

Country	lifeexpf	birthrate
Somalia	55	46
Tanzania	55	50
Zambia	59	48
Zaire	56	45
Algeria	68	31
Namibia	63	45
Burkina Faso	53	50
Cuba	79	18
Equador	72	28
North Korea	72	24
Mongolia	68	34
Thailand	71	20
Turkey	72	28
France	82	13
Netherlands	81	13

Ejemplo 2. Esperanza de vida.



Modelo que podemos plantear:
(Llamamos Y a la esperanza de vida y X a los nacimientos)

Y= a + b X +"error"

1.1 Covarianza

En la práctica muchas relaciones que analizaremos serán débiles y no bastará ver su gráfico, sino que además tendremos que medir la **magnitud** o el **grado de relación lineal** entre estas variables.

Dos de las medidas más utilizadas para datos bivariantes que sirven para cuantificar el grado de relación lineal son:

- La covarianza
- El coeficiente de correlación lineal de Pearson

Dados n pares de valores $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, que corresponden a la observación de dos variables X e Y sobre n individuos, se define la **covarianza muestral entre** X **e** Y como:

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

donde

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \,, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Ejercicio: Demostrar que

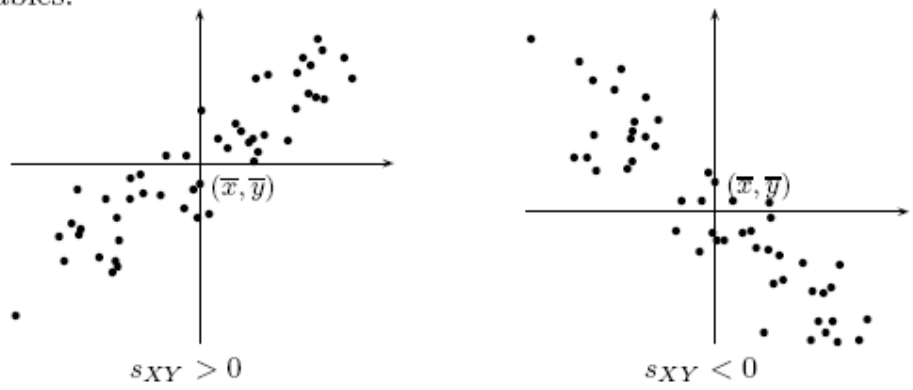
$$s_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\ \bar{y}$$

Interpretación geométrica de la covarianza

Consideremos una nube de puntos formada por los n pares de datos $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. El centro de gravedad de esta nube de puntos, que es $(\bar{x}, \bar{y})$, divide a la nube en cuatro cuadrantes.

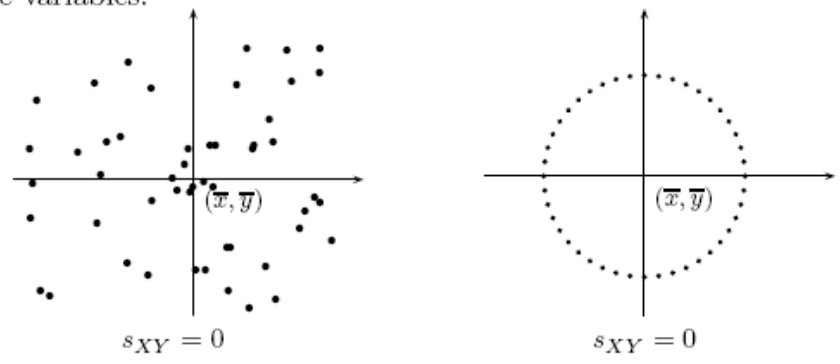
Los puntos que se encuentran en el primer y tercer cuadrante contribuyen positivamente a s_{XY}, mientras que los puntos que están en el segundo y cuarto cuadrantes, lo hacen negativamente.

Figure 1: Interpretación geométrica de la covarianza. Relación lineal entre variables.



- Si hay más puntos en el tercer y primer cuadrantes, entonces $s_{XY} \geq 0$, lo que puede interpretarse como que la variable Y tiende a aumentar cuando lo hace X.
- Si hay más puntos en el segundo y cuarto cuadrantes, entonces $s_{XY} \leq 0$, lo que puede interpretarse como que la variable Y tiende a disminuir cuando X aumenta.

Figure 2: Interpretación geométrica de la covarianza. Relación no lineal entre variables.



- Si los puntos se reparten con igual intensidad alrededor del centro de gravedad, entonces se tendrá que $s_{XY} = 0$, lo que indicará que **no hay relación lineal** entre las variables.

  Atención: Esto no significa que no pueda existir ***otro tipo de relación*** entre ambas variables.

La covarianza presenta los siguientes inconvenientes:

1) depende de las unidades de medida de las variables,
2) no está acotada ni superior ni inferiormente.

Proposición: Si $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ es una muestra de tamaño n de la variable bidimensional (X,Y), y a, b son dos valores reales, entonces:

$$s_{X,\,a+bY} = b\, s_{XY}$$

Demostración (pizarra)

1.2 Correlación

El **coeficiente de correlación lineal de Pearson** es una medida adimensional de la variación conjunta de dos variables.

Se define como:

$$r_{X,Y} = \frac{s_{XY}}{s_X s_Y}$$

donde

$$s_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2, \quad s_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Propiedades:

- Es una medida acotada: $-1 \leq r_{X,Y} \leq 1$.
- No se ve afectado por traslaciones ni por cambios de escala (del mismo signo), es decir: $r_{X,a+bY} = \mathrm{r}_{XY}, \ \forall b > 0$.
- Su signo coincide con el signo de $s_{X,Y}$ e indica el tipo de dependencia entre X e Y.
- Si $|\, r_{X,Y} \,| = 1 \Rightarrow$ correlación máxima entre X e Y.
- Si $r_{X,Y} = 0 \Rightarrow$ no existe relación lineal entre X e Y, y se dice que las variables están incorreladas.

Ejercicio 1. Con los datos del ejemplo 2 (esperanza de vida), calcula la covarianza y el coeficiente de correlación lineal.

Country	lifeexpf (***y***)	birthrate (***x***)
Somalia	55	46
Tanzania	55	50
Zambia	59	48
Zaire	56	45
Algeria	68	31
Namibia	63	45
Burkina Faso	53	50
Cuba	79	18
Equador	72	28
North Korea	72	24
Mongolia	68	34
Thailand	71	20
Turkey	72	28
France	82	13
Netherlands	81	13

Solución Ejercicio 1.

	y	x	xy	x^2	y^2
Somalia	55	46	2530	2116	3025
Tanzania	55	50	2750	2500	3025
Zambia	59	48	2832	2304	3481
Zaire	56	45	2520	2025	3136
Algeria	68	31	2108	961	4624
Namibia	63	45	2835	2025	3969
Burkina Faso	53	50	2650	2500	2809
Cuba	79	18	1422	324	6241
Equador	72	28	2016	784	5184
North Korea	72	24	1728	576	5184
Mongolia	68	34	2312	1156	4624
Thailand	71	20	1420	400	5041
Turkey	72	28	2016	784	5184
France	82	13	1066	169	6724
Netherlands	81	13	1053	169	6561
total	1006	493	31258	18793	68812
medias	**67,07**	**32,87**	**2083,87**	**1252,87**	**4587,47**
varianzas	**89,53**	**172,65**		**covarianza**	**-120,39**
				correlación	**-0,97**

Ejercicio 2. Calcular el coeficiente de correlación para los datos de la tabla siguiente:

x	***y***
2	5
3	7
4	8
5	13
6	14

Cómo cambiará el coeficiente de correlación si:

a) sumamos 5 a la variable ***x***.
b) sumamos 5 a ambas variables.
c) intercambiamos los valores de ***x*** por los de ***y***.

Ejercicio 3. ¿Cuál sería el coeficiente de correlación entre las edades de los cónyuges si las mujeres **siempre** se casaran con hombres 2 años mayores que ellas?

2. Hipótesis básicas

Los factores que influyen en la **variable respuesta** y pueden dividirse en:

1) Un primer grupo que contiene a una variable x, **independiente**, **no aleatoria** y **conocida** al observar y.
2) Un segundo grupo de múltiples factores que afectan a y, cada uno en pequeñas cantidades, que se denomina **perturbación aleatoria** o **error aleatorio**.

$$y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\textit{primer grupo}} + \underbrace{u_i}_{\textit{segundo grupo}}$$

y_i, u_i son variables aleatorias,
x_i es una variable predeterminada con valores conocidos,
β_0, β_1 son parámetros desconocidos.

Modelo general	$y_i = g(x_{1i}, x_{2i}, ..., x_{ni})$
Hipótesis inicial	$y_i = g_1(x_{1i}) + g_2(x_{2i}, ..., x_{ni})$
Aproximación	$y_i = \beta_0 + \beta_1 x_i + u_i$

Se establecen las siguientes **hipótesis para la perturbación aleatoria**:

a) Tiene esperanza nula:
$$E(u_i) = 0$$
b) Tiene varianza constante (**homocedasticidad**) y no depende de x:
$$\text{var}(u_i) = \sigma^2$$
c) Como consecuencia del TLC tiene distribución normal:
$$u_i \sim N(0, \sigma^2)$$
d) Las perturbaciones son independientes (dos a dos):
$$E(u_i u_j) = E(u_i)E(u_j), \quad \forall i \neq j$$

Pregunta: ¿cuánto vale la covarianza entre dos perturbaciones cualesquiera?

Observaciones:

1) Las condiciones a), b), c) y d) se resumen diciendo que $u_1, u_2, \ldots, u_n$ son n v.a. i.i.d. con ley normal $N(0, \sigma^2)$.
2) Las condiciones a), b), d) se denominan **condiciones de Gauss-Markov**.

Estas cuatro condiciones inducen la siguiente estructura sobre la variable respuesta y:

a') La esperanza de y depende linealmente de x:

$$E(y_i) = E(\underbrace{\beta_0 + \beta_1 x_i}_{\textbf{constante}} + u_i) = \beta_0 + \beta_1 x_i + \underbrace{E(u_i)}_{=0}$$

β_0 es el valor medio de y cuando x vale 0,
β_1 es el incremento que experimenta la media de y cuando x aumenta en una unidad.

b') La varianza de y es constante:

$$\text{var}(y_i) = \text{var}(\underbrace{\beta_0 + \beta_1 x_i}_{\textbf{constante}} + u_i) = \text{var}(u_i) = \sigma^2$$

σ^2 es la varianza del modelo.

c') Para cada valor de x, la respuesta y tiene ley normal:

$$y_{|x=x_i} \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

Pregunta: ¿Son las y_i idénticamente distribuidas?

d') Las y_i son independientes dos a dos.

Consideraciones importantes:

- La hipótesis principal del modelo es que **la media de la ley de la respuesta**, para x fija, **varía linealmente con** x. Esta hipótesis debe comprobarse siempre, pues **condiciona toda la construcción del modelo**.
- La utilidad del modelo lineal radica en que muchas relaciones no lineales puede convertirse en lineales **transformando las variables adecuadamente**. Una relación lineal debe considerarse como una **aproximación simple**, en un rango de valores limitado, de una relación más compleja. Por tanto, será necesario tener presente:
  1) El rango de valores dentro del cual se va a trabajar,
  2) El peligro de extrapolar una relación fuera de ese rango.
- ¿**Cuándo no se cumplirán las condiciones de Gauss-Markov**?

  La condición a) no será cierta si existen observaciones tomadas en condiciones heterogéneas con el resto. Esto puede comprobarse mediante un análisis de residuos del modelo, y es importante porque **una única observación atípica puede tener gran influencia en la estimación**.

  La condición b) no se cumplirá si la variabilidad de y depende de la media de y (**heterocedasticidad**).

  La condición d) es esperable en situaciones estáticas (todas las observaciones corresponden al mismo período temporal), pero casi nunca en **situaciones dinámicas**, donde la variable respuesta se mide a lo largo del tiempo.

3. Estimación por el método de los mínimos cuadrados

3.1 Estimación de los parámetros β_1 y β_0

Dadas n observaciones $(x_1,y_1), \ldots, (x_n,y_n)$ el método de los mínimos cuadrados (MMC) selecciona como estimación de la recta poblacional

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

a la recta de regresión

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

donde

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2},\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

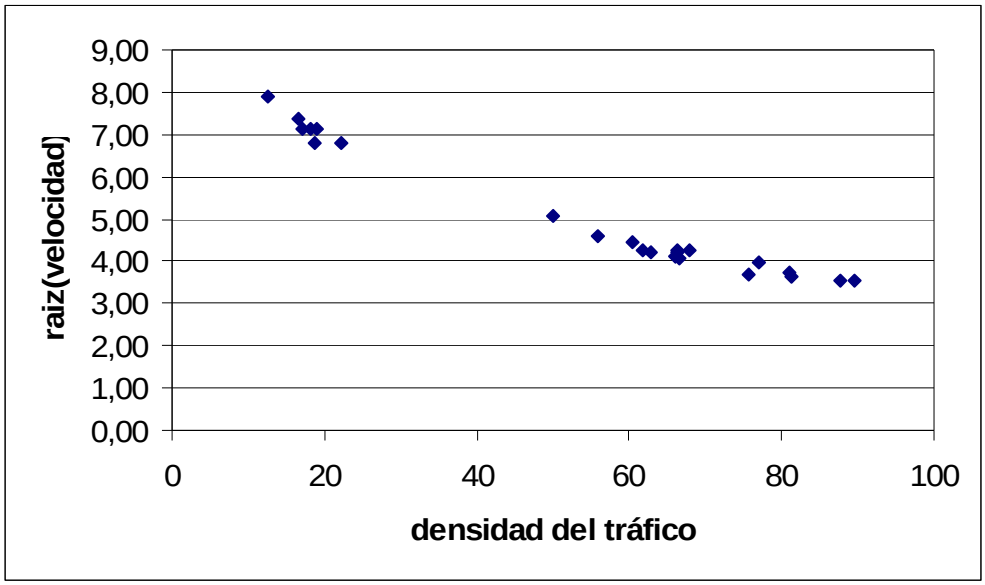
que estima el valor medio de y, para cada valor de x.

Demostración (obtención de las ecuaciones normales en la pizarra)

Ejercicio 4. Con los datos del ejemplo 1 (densidad del tráfico) encontrar la recta de regresión que mejor ajusta la velocidad en función de la densidad del tráfico. Por razones que se verán más adelante, tomar la raíz cuadrada de la velocidad.

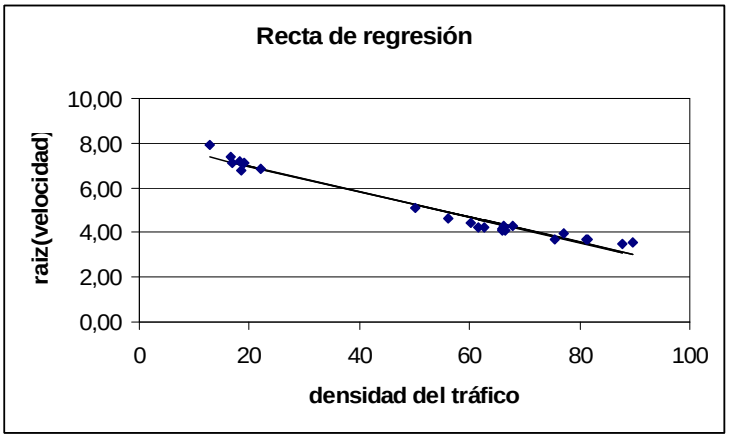
densidad x	raiz(velocidad) y	densidad x	raiz(velocidad) y
12,7	7,90	18,3	7,16
17	7,12	19,1	7,13
66	4,14	16,5	7,40
50	5,09	22,2	6,82
87,8	3,52	18,6	6,80
81,4	3,66	66	4,11
75,6	3,70	60,3	4,45
66,2	4,23	56	4,60
81,1	3,71	66,3	4,28
62,8	4,23	61,7	4,24
77	3,97	66,6	4,07
89,6	3,55	67,8	4,28

gráfico de dispersión



densidad x	raiz(velocidad) y	x^2	y^2	xy
12,7	7,90	161,29	62,40	100,32
17	7,12	289	50,70	121,05
66	4,14	4356	17,10	272,92
50	5,09	2500	25,90	254,46
87,8	3,52	7708,84	12,40	309,18
81,4	3,66	6625,96	13,40	297,97
75,6	3,70	5715,36	13,70	279,82
66,2	4,23	4382,44	17,90	280,08
81,1	3,71	6577,21	13,80	301,27
62,8	4,23	3943,84	17,90	265,70
77	3,97	5929	15,80	306,07
89,6	3,55	8028,16	12,60	318,05
18,3	7,16	334,89	51,20	130,94
19,1	7,13	364,81	50,80	136,13
16,5	7,40	272,25	54,70	122,03
22,2	6,82	492,84	46,50	151,38
18,6	6,80	345,96	46,30	126,56
66	4,11	4356	16,90	271,32
60,3	4,45	3636,09	19,80	268,32
56	4,60	3136	21,20	257,84
66,3	4,28	4395,69	18,30	283,62
61,7	4,24	3806,89	18,00	261,77
66,6	4,07	4435,56	16,60	271,35
67,8	4,28	4596,84	18,30	290,04
1306,6	120,17	86390,92	652,2	5678,21

	x	y	x^2	y^2	xy
medias	54,44	5,01	3599,62	27,18	236,59
varianzas	635,73	2,10			
covarianza	-36,00				
correlación	-0,98				
coeficientes	β_0	8,09			
	β_1	-0,06			



densidad x	raiz(velocidad) y	modelo $\hat{y}=\beta_0+\beta_1 x$	residuo
12,7	7,90	7,37	0,53
17	7,12	7,13	-0,01
66	4,14	4,35	-0,22
50	5,09	5,26	-0,17
87,8	3,52	3,12	0,40
81,4	3,66	3,48	0,18
75,6	3,70	3,81	-0,11
66,2	4,23	4,34	-0,11
81,1	3,71	3,50	0,22
62,8	4,23	4,53	-0,30
77	3,97	3,73	0,25
89,6	3,55	3,02	0,53
18,3	7,16	7,05	0,10
19,1	7,13	7,01	0,12
16,5	7,40	7,16	0,24
22,2	6,82	6,83	-0,01
18,6	6,80	7,04	-0,23
66	4,11	4,35	-0,24
60,3	4,45	4,68	-0,23
56	4,60	4,92	-0,31
66,3	4,28	4,34	-0,06
61,7	4,24	4,60	-0,35
66,6	4,07	4,32	-0,24
67,8	4,28	4,25	0,03
1306,6	120,17	120,17	0,00

Observaciones:
(1) La suma de los residuos es cero.
(2) La suma de los valores de y coincide con la suma de los valores de ŷ.

3.2 Estimación de la varianza σ^2

La desviación típica de la perturbación, σ, mide la **precisión del ajuste** de la recta de regresión. Para medir la variabilidad de los puntos alrededor de la recta utilizaremos la desviación típica residual (estimador de σ).

Se define la **varianza residual** como:

$$s_R^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2$$

donde $e_i = y_i - \hat{y}_i$ son los residuos del modelo.

Se divide por n-2 (y no por n), porque **los residuos no son independientes**, pues las ecuaciones normales inducen dos restricciones sobre ellos:

$$\sum_{i=1}^{n} e_i = 0, \quad \sum_{i=1}^{n} e_i x_i = 0.$$

Por tanto, solamente hay n-2 valores independientes o n-2 ***grados de libertad***.

Observación: En la práctica, existe otra forma para calcular la varianza residual, sin utilizar explícitamente los residuos:

$$s_R^2 = \frac{n(s_y^2 - \hat{\beta}_1^2 s_x^2)}{n-2}$$

Demostración (pizarra).

Atención: s_R^2 es una varianza y, por tanto, su valor debe ser **siempre positivo**. Si al calcularlo aparecen valores negativos, será debido a errores de redondeo. En caso de que esto ocurriera, se debe tomar una **mayor precisión decimal** en las estimaciones $\hat{\beta}_1^2, s_y^2, s_x^2$.

Consideraciones importantes:

- La recta de regresión y la desviación típica residual juegan el mismo papel que la media y la desviación típica de una distribución de datos:

  la recta de regresión indica el valor medio de y para cada valor de x, mientras que la desviación típica residual mide la desviación promedio de las observaciones alrededor de la recta.

Ejemplo: Hallar la varianza y desviación típica residual del **Ejemplo 1** (densidad del tráfico).

Recordemos que para este ejemplo *n=24,* y además habíamos calculado

$$s_x^2 = 635.73,\ s_y^2 = 2.10,\ \hat{\beta}_1^2 = -0.057.$$

Entonces, sustituyendo en la fórmula anterior, obtenemos que la varianza residual es igual a:

$$s_R^2 = \frac{n(s_y^2 - \hat{\beta}_1^2 s_x^2)}{n-2} = \frac{24(2.10 - (-0.057)^2 635.73)}{24-2} = 0.0377$$

y la desviación típica residual será su raíz cuadrada.

4. Propiedades de los estimadores de los parámetros β_1 y β_0

Si se toman muestras de y manteniendo constantes los valores de x, y para cada muestra se calculan los estimadores $\hat{\beta}_0, \hat{\beta}_1$ y s_R^2, éstos tomarán **valores distintos de una muestra a otra**. Se trata pues de **variables aletaorias** y, por tanto, tienen una distribución de probabilidad en el muestreo.

Las propiedades de los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ consisten en determinar su esperanza, varianza y distribución de probabilidad en el muestreo.

4.1 Propiedades del coeficiente de regresión $\hat{\beta}_1$ (demostraciones en la pizarra)

a) $\hat{\beta}_1$ tiene ley normal al ser combinación lineal de v.a. normales.

b) $\hat{\beta}_1$ es un estimador **insesgado** del parámetro β_1.

c) $\hat{\beta}_1$ es el estimador **más eficiente** del parámetro β_1. La varianza de $\hat{\beta}_1$ es:

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{n s_x^2},$$

cuyo estimador es

$$s_{\hat{\beta}_1}^2 = \frac{s_R^2}{n s_x^2} = \ldots = \frac{s_y^2 - \hat{\beta}_1^2 s_x^2}{(n-2)s_x^2}$$

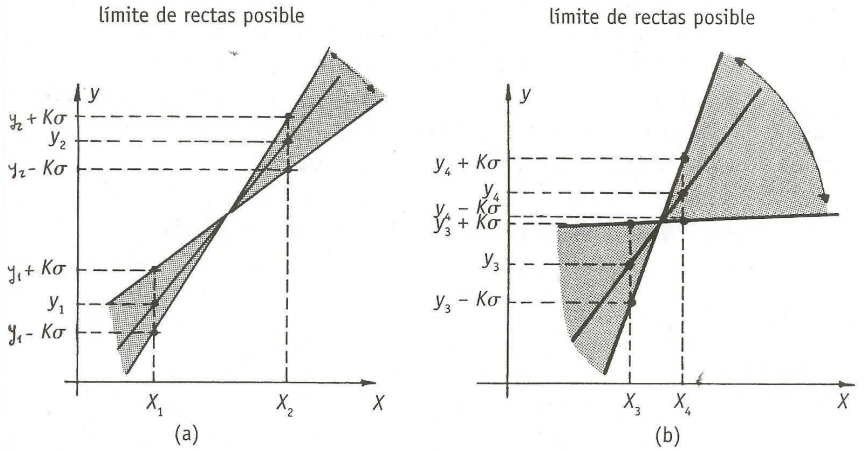
La varianza de $\hat{\beta}_1$ indica **el error de estimación** del parámetro β_1. Así, de la fórmula

$$s^2_{\hat{\beta}_1} = \frac{s^2_R}{n s^2_x}$$

se deduce que:

1) El error aumenta al aumentar la varianza residual, para x fijo,
2) El error disminuye al aumentar la dispersión de x,
3) El error disminuye al aumentar el tamaño muestral.

5.12 El error en la estimación de la pendiente es mayor cuando los puntos están muy próximos (caso b)



4.2 Propiedades del estimador $\hat{\beta}_0$

a) $\hat{\beta}_0$ tiene ley normal al ser combinación lineal de v.a. normales.

b) $\hat{\beta}_0$ es un estimador **insesgado** del parámetro β_0.

c) La varianza de $\hat{\beta}_0$ puede expresarse como suma de dos términos:

$$\text{var}(\hat{\beta}_0) = \frac{\sigma^2}{n}\left(1 + \frac{\bar{x}^2}{s^2_x}\right)$$

El primer término es σ^2/n, que es el error de estimación de $\bar{y}$.

El segundo término, $\frac{\sigma^2}{n s^2_x}\bar{x}^2 = \text{var}(\hat{\beta}_1)\bar{x}^2$, tiene en cuenta que el error de estimación de la pendiente de la recta se transmite a la ordenada en el origen en función de lo alejado que se encuentre $\bar{x}$ del origen (aumenta a medida que aumenta $\bar{x}$).

Conclusiones:

Para una muestra concreta,

- el valor calculado para $\hat{\beta}_1$ puede interpretarse como un valor extraído al azar de una distribución normal de media β_1 y varianza $\sigma^2/(ns_x^2)$. Esto equivale a decir que el estimador

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{ns_x^2}\right)$$

- el valor calculado para $\hat{\beta}_0$ puede interpretarse como un valor extraído al azar de una distribución normal de media β_0 y varianza $\frac{\sigma^2}{n}\left(1+\frac{\bar{x}^2}{s_x^2}\right)$. Esto equivale a decir que el estimador

$$\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2}{n}\left(1+\frac{\bar{x}^2}{s_x^2}\right)\right)$$

Relación entre los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$

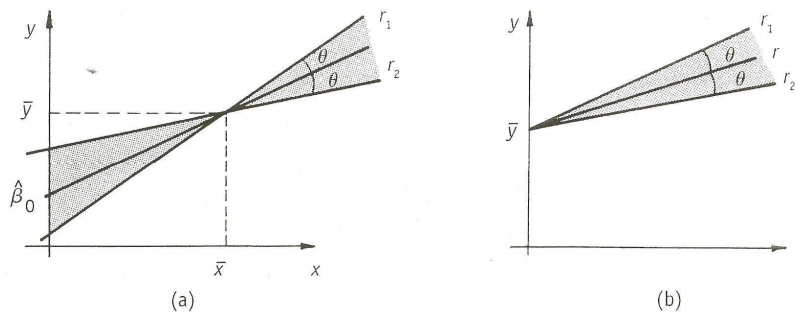
Tal como se ha obtenido $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$, está claro que $\hat{\beta}_0$ y $\hat{\beta}_1$ **no son independientes**. Puede demostrarse que:

$$\text{cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{x}\sigma^2}{ns_x^2} = -\bar{x}\,\text{var}(\hat{\beta}_1)$$

De donde se deduce que:

1) si $\bar{x} > 0$, la covarianza es negativa, indicando que errores por exceso en la pendiente producirán errores por defecto en $\hat{\beta}_0$, y viceversa.
2) La dependencia disminuye con $\bar{x}$ y con los factores que contribuyen a estimar la pendiente con mayor precisión.

Figura 5.14 Relación entre $\hat{\beta}_0$ y $\hat{\beta}_1$



5. Propiedades de la varianza residual

Recordemos que solamente hay *n-2* residuos independientes.

Puede demostrarse (lo veremos en general, cuando estudiemos el modelo de regresión lineal múltiple) que la suma cuadrática de los residuos de variables normales divida por σ^2 tiene ley chi-cuadrado con los grados de libertad que tengan los residuos.

$$\frac{1}{\sigma^2}\sum_{i=1}^{n} e_i^2 \sim \chi^2_{n-2} \underset{s_R^2=\frac{1}{n-2}\sum e_i^2}{\rightleftarrows} \frac{(n-2)s_R^2}{\sigma^2} \sim \chi^2_{n-2}$$

La **esperanza** y **varianza** del estimador s_R^2 son:

$$E(s_R^2)=\sigma^2,\ \ \mathrm{var}(s_R^2)=\frac{2\sigma^4}{n-2}.$$

(Se deducen a partir de la esperanza y varianza de la ley Gamma $\chi_k^2 = Gamma\left(\frac{k}{2},2\right) \Rightarrow E(\chi_k^2)=k,\ \ \mathrm{var}(\chi_k^2)=2k.$)

Propiedades de los estimadores y su dependencia de las hipótesis básicas

Pará-metro	Estimador	Esperanza	Varianza	Distribución
β_0	$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$	$E(\hat{\beta}_0)=\beta_0$ Linealidad	$\mathrm{var}(\hat{\beta}_0)=\frac{\sigma^2}{n}\left[1-\left(\frac{\bar{x}}{s_x}\right)^2\right]$ Homocedasticidad Independencia	*Normal* Normalidad
β_1	$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2}$	$E(\hat{\beta}_1)=\beta_1$ Linealidad	$\mathrm{var}(\hat{\beta}_1)=\frac{\sigma^2}{ns_x^2}$ Homocedasticidad Independencia	*Normal* Normalidad
σ^2	$s_R^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$	$E(s_R^2)=\sigma^2$ Linealidad Homocedasticidad Independencia	$\mathrm{var}(s_R^2)=\frac{2\sigma^4}{n-2}$ Linealidad Homocedasticidad Independencia Normalidad	$\frac{(n-2)s_R^2}{\sigma^2} \sim \chi^2_{n-2}$ Linealidad Homocedasticidad Independencia Normalidad

El incumplimiento de las hipótesis básicas afecta la estimación de los parámetros.

Descomposición de la variabilidad: Relación fundamental de la regresión

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{VT}=\underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{VE}+\underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{VNE}$$

Esta relación nos dice que la variabilidad de y (VT) descompone en dos términos independientes:

VE: contiene la **variabilidad explicada** o recogida en el modelo de regresión.

VNE: contiene la **variabilidad no explicada** por el modelo de regresión, que es debido al carácter estocástico de la relación.

Demostración de la relación fundamental de la regresión:

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

$$=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+2\underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})}_{=0\,?}$$

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})=\sum(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)(\hat{\beta}_0+\hat{\beta}_1x_i-\hat{\beta}_0-\hat{\beta}_1\bar{x})$$

$$=\hat{\beta}_1\sum(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)(x_i-\bar{x})=\hat{\beta}_1\sum(y_i-\hat{\beta}_1x_i)(x_i-\bar{x})-\hat{\beta}_1\hat{\beta}_0\underbrace{\sum(x_i-\bar{x})}_{=0}$$

$$=\hat{\beta}_1\sum(y_i-\hat{\beta}_1x_i)(x_i-\bar{x})=\hat{\beta}_1\sum(y_ix_i-y_i\bar{x}-\hat{\beta}_1x_i^2+\hat{\beta}_1x_i\bar{x})$$

$$=\hat{\beta}_1n(\overline{xy}-\bar{x}\,\bar{y}-\hat{\beta}_1\overline{x^2}+\hat{\beta}_1\bar{x}^2)=\hat{\beta}_1n\underbrace{(s_{xy}-\hat{\beta}_1s_x^2)}_{=0}=0$$

Es conveniente **descomponer la varianza** en una tabla ADEVA (análisis de la varianza) o ANOVA (*analysis of variace*) de la forma siguiente:

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

$$VNE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^{\,2} = (n-2)s_R^2 = n(s_y^2 - \hat{\beta}_1^2 s_x^2)$$

$$VE = VT - VNE = ns_y^2 - n(s_y^2 - \hat{\beta}_1^2 s_x^2) = n\hat{\beta}_1^2 s_x^2$$

Tabla ANOVA:

Fuente de variación	Sumas de Cuadrados	Grados de libertad	Cociente o varianza
VE	$n\hat{\beta}_1^2 s_x^2$	*1*	$n\hat{\beta}_1^2 s_x^2 / 1$
VNE	$(n-2)s_R^2$	*n-2*	$(n-2)s_R^2/(n-2) = s_R^2$
VT	ns_y^2	*n-1*	

Ejemplo. Con los datos del **Ejemplo 1** (densidad del tráfico) encontrar la tabla ADEVA (ANOVA).

Recordemos que ***x***=*densidad del tráfico*, ***y***=*raíz cuadrada de la velocidad*.

$$n = 24,\ \sum_{i=1}^{n} y_i = 120.17,\ \sum_{i=1}^{n} x_i = 1306.6,$$

$$\sum_{i=1}^{n} y_i^{\,2} = 652.2,\ \sum_{i=1}^{n} x_i^{\,2} = 86390.92,\ \sum_{i=1}^{n} x_i y_i = 5678.21,$$

$$\hat{\beta}_1 = -0.057.$$

Para construir la tabla ANOVA vamos a calcular primero $s_x^{\,2}$, $s_y^{\,2}$ y $s_R^{\,2}$:

$$s_x^2 = \overline{x^2} - \bar{x}^2 = \frac{86390.92}{24} - \left(\frac{1306.6}{24}\right)^2 = 635.73$$

$$s_y^2 = \overline{y^2} - \bar{y}^2 = \frac{652.2}{24} - \left(\frac{120.17}{24}\right)^2 = 2.10$$

$$s_R^2 = \frac{n(s_y^2 - \hat{\beta}_1^2 s_x^2)}{n-2} = \frac{24}{22}\left(2.10 - (-0.57)^2 635.73\right) = 0.0377$$

La tabla ANOVA es:

Fuente de variación	Sumas de Cuadrados	Grados de libertad	Cociente o varianza
VE	$n\hat{\beta}_1^2 s_x^2 = 49.5717$	1	49.5717
VNE	$(n-2)s_R^2 = 0.8294$	(n-1)-1=23-1=22	0.0377
VT	$ns_y^2 = 50.4$	n-1=23	

Observación: Generalmente se calculan solamente VE y VT y se obtiene VNE=VT-VE.

6. El coeficiente de determinación

La varianza residual es un índice de precisión de la recta de regresión, pero **no es útil para comparar rectas de regresión** de variables distintas, porque depende de las unidades de medida de la variable respuesta.

Una medida más adecuada del ajuste es **la proporción de variabilidad explicada**. Se define el **coeficiente de determinación** del modelo como:

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

En la práctica, utilizaremos la siguiente expresión:

$$R^2 = \frac{n\hat{\beta}_1^2 s_x^2}{ns_y^2} = \hat{\beta}_1^2 \frac{s_x^2}{s_y^2}$$

Ejercicio: Expresar el coeficiente de determinación en función de la varianza residual.

Relación entre el coeficiente de determinación y el coeficiente de correlación lineal de Pearson

Puesto que $\hat{\beta}_1 = s_{xy} \big/ s_x^2$, entonces $s_{xy} = \hat{\beta}_1 s_x^2$ y, por tanto:

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\hat{\beta}_1 s_x^2}{s_x s_y} = \hat{\beta}_1 \frac{s_x}{s_y}$$

elevando al cuadrado, tenemos que:

$$r^2 = \hat{\beta}_1^2 \frac{s_x^2}{s_y^2} = R^2$$

Atención: $r^2 = R^2 \Rightarrow r = \pm\sqrt{R^2}$

Propiedades del coeficiente de determinación

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

1) $0 \le R^2 \le 1$

$R^2 \ge 0$, al ser un cociente de varianzas,

$R^2 \le 1$, puesto que $VE \le VT = VE + VNE$

2) Si $R^2 = 1 \Rightarrow VE = VT \Rightarrow VNE = 0 \Rightarrow s_R^2 = 0$

es decir, todos los residuos son cero

3) Si $R^2 = 0 \Rightarrow VE = 0 \Rightarrow VNE = VT$

la variación de y no es debida a x, sinó al error.

Ejemplo. Con los datos del **Ejemplo 1** (densidad del tráfico) encontrar el valor del coeficiente de determinación e interpretarlo.

$$R^2 = \frac{VE}{VT} = \frac{49.5717}{50.4} = 0.9835 \Rightarrow 98.35\%$$

El modelo $y=8.09-0.057x$ resume el 98.35% de la variabilidad de y. Queda sin explicar el 100-98.35=1.65% de la variabilidad de los datos.

La velocidad de los vehículos queda explicada por la densidad del tráfico en un 98.35%. La influencia de otros factores en la velocidad es del 1.65%.

Ejercicio 5 (longitud de la línea de la mano). En un estudio para relacionar la longitud de la línea de la vida en la mano izquierda y la vida de una persona, se han observado los siguientes datos de 50 personas con los siguientes resultados:
x= *longitud de línea (en cm)*
y= *edad al morir (en años)*

$$\sum_{i=1}^{50} y_i = 3333, \quad \sum_{i=1}^{50} y_i^{\,2} = 231933, \quad \sum_{i=1}^{50} x_i y_i = 30549,$$
$$\sum_{i=1}^{50} x_i = 459.9, \quad \sum_{i=1}^{50} x_i^{\,2} = 4308.57.$$

a) Se pide construir una recta de regresión de ***y*** sobre ***x*** y encontrar la varianza residual.
b) Descomponer la variabilidad y expresarla en una tabla ADEVA (ANOVA).

Ejercicio 6 (salario-escolarización). Un investigador considera que el salario que percibe un individuo es función lineal de sus años de escolarización, esto es, $y_i = \beta_0 + \beta_1 x_i + u_i$, donde y_i representa el salario mensual del individuo i-ésimo, x_i los años de estudio de dicho individuo y u_i es el término de error, que supondremos que verifica las hipótesis habituales del modelo de regresión. El investigador ha obtenido una muestra aleatoria de 100 individuos, de la que conocemos la siguiente información:

$$\sum_{i=1}^{100} y_i = 1180, \quad \sum_{i=1}^{100} y_i^2 = 25543, \quad \sum_{i=1}^{100} x_i y_i = 13469,$$

$$\sum_{i=1}^{100} x_i = 1000, \quad \sum_{i=1}^{100} x_i^2 = 12820.$$

Se pide:

a) *Obtener estimadores insesgados de los parámetros* β_0, β_1 y σ^2**,** así como el coeficiente de determinación. Interpretar los resultados.

b) Hallar la descomposición de la variabilidad (tabla ADEVA).