Tema 6: Regresión

Universidad Francisco de Vitoria UFV Madrid

Alberto Nogales

alberto.nogales@ufv.es

Curso 2022-2023

Índice

- Definición del problema
- Regresión lineal: coeficientes, bondad del ajuste, análisis de residuos, contraste de hipótesis.
- Regresión lineal multiple.
- Regresión logística.
- Regresión no linesal.

Introducción

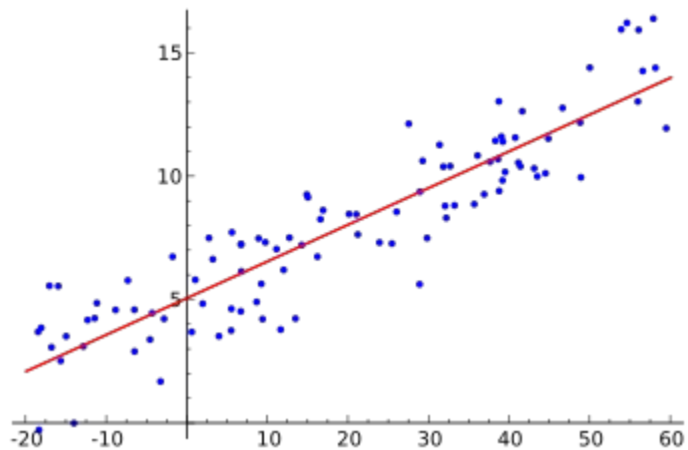
A partir de dos características numéricas de una población concreta. Establecer si existe una relación entre ellas o no, para predecir el valor de una en concreto.

Aplicaciones:

1. Cómo influye la altura de un padre en la de un hijo.
2. Relación entre los gastos de promoción de una empresa y sus ingresos.
3. Estimar el valor de una vivienda según su tamaño.

Definición del problema

- **Objetivo:** obtener estimaciones razonables de Y para distintos valores de X a partir de una muestra de n pares de valores (x1 , y1), . . . , (xn , yn).



Definición del problema

1) Si ambas variables están realmente relacionadas entre sí o si, por el contrario, pueden considerarse independientes.

2) Conocer el "grado de relación", así como el "tipo" de relación entre ambas.

3) Predecir la variable que es considerada como dependiente a partir de los valores de la otra, que es considerada independiente, y si es así, con qué precisión.

Definiciones

- Variable independiente (X): también llamada explicativa o exógena. La que se usa para predecir.

- Variable dependiente (Y): también llamada respuesta o endógena. La que se predice.

- Análisis de correlación: fuerza y dirección de la relación lineal.

- Análisis de regresión: predice o estima una variable en función del valor de otra variable.

Definiciones

- **Correlación:** existe una dependencia entre las variables. Determina cuál es el grado de dependencia y tipo de relación

- **Regresión:** Permite estimar los valores de Y a partir de X.

Definiciones

Una manera de saber si 2 variables están relacionadas es mediante la covarianza. El problema es que depende de las unidades de medida de las variables.

- **Coeficiente de correlación:** Cuantifica la intensidad de la relación lineal entre dos variables. El valor es independiente de las unidades utilizadas por las variables y es sensible a las anomalías.

Definiciones

Coeficiente de correlación de la muestra (r de Pearson):

Nos indica que tipo de correlación lineal podría tener.

$$r = \frac{cov\ muestra\ (X,Y)}{\sqrt{{s_x}^2 . sy^2}} = \frac{\sum_{i=1}^{n} \frac{(Xi - \bar{X})(Yi - \bar{Y})}{n-1}}{\sqrt{\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1} . \sum_{i=1}^{n} \frac{(Y_i - \bar{Y})^2}{n-1}}}$$

Coeficiente de correlación de la población (valor p):

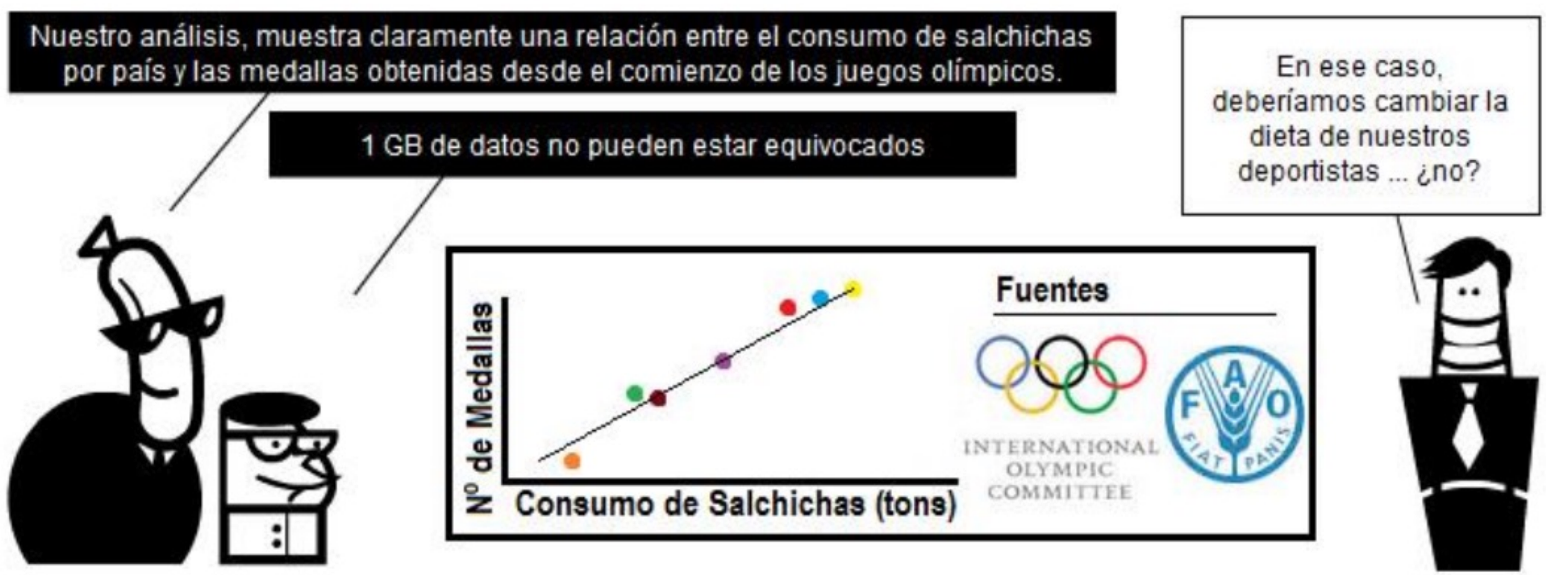
Nos indica si la relación es por azar o debido a la naturaleza de los datos.

$$\rho = \frac{cov(X,Y)}{\sqrt{var(X).var(Y)}} = \frac{E(X - \mu_X)(Y - \mu_Y)}{\sqrt{E[(X - \mu_X)^2] . E[(Y - \mu_Y)^2]}}$$

Definiciones

Si el coeficiente de correlación demuestra que hay relación entre variables **NO IMPLICA CAUSALIDAD!!!!**

PREGUNTA: ¿Cómo mejorar el rendimiento de nuestros deportistas olímpicos?



Viñeta elaborada por David Martín-Moncunill con http://stripgenerator.com

Definiciones

Interpretaciones de la r de Pearson:

Cuando vale 0, **NO** hay correlación **LINEAL** puede haber de otro tipo.

Cuando vale 1, implica qué al aumentar una variable, aumenta la otra.

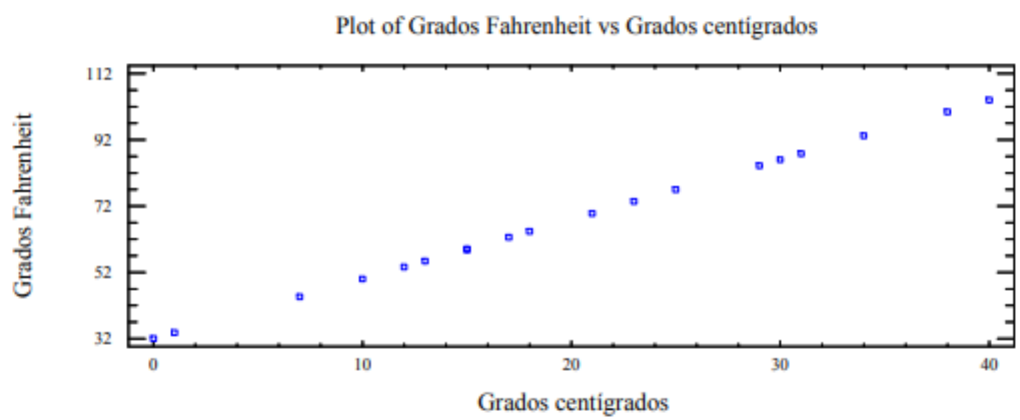
Cuando vale -1, implica que al aumentar una variable, la otra disminuye.

Definiciones: tipos de correlación

- **Determinista:** Conocido el valor de X, el valor de Y queda perfectamente establecido. Son del tipo: $y = f(x)$

  Ejemplo: La relación existente entre la temperatura en grados centígrados (X) y grados Fahrenheit (Y) es:

  $y = 1{,}8x + 32$

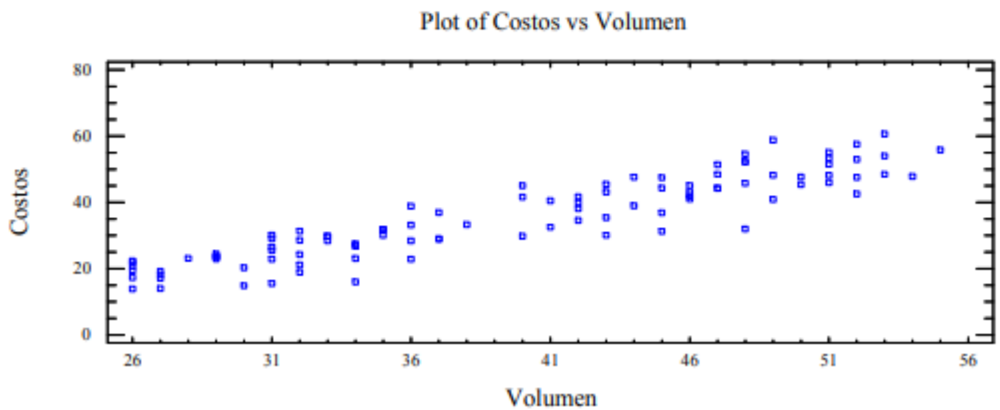


Definiciones: tipos de correlación

- **No determinista:** Conocido el valor de X, el valor de Y no queda perfectamente establecido. Son del tipo:

  $y = f(x) + u$ donde u es una perturbación desconocida (variable aleatoria).

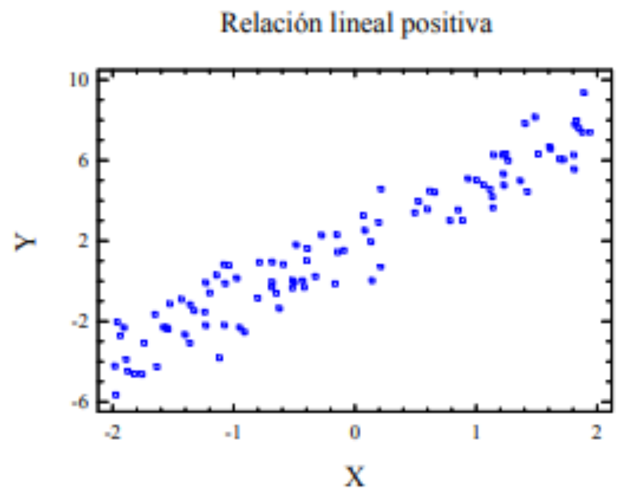
  Ejemplo: Volumen de producción (X) y el coste total (Y) asociado a un producto en un grupo de empresas.



Definiciones: tipos de correlación

- **Lineal positiva:** Cuando la función f (x) es lineal, $f(x) = \beta_0 + \beta_1 x$. Tiene aspecto de recta y la variable Y aumenta según aumenta X ($\beta_1 > 0$).

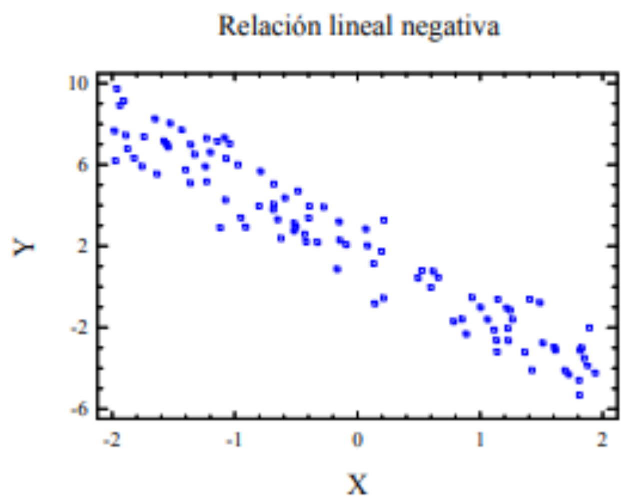
  Ejemplo: Precio del petróleo y el precio de los billetes de avión.



Definiciones: tipos de correlación

- **Lineal negativa:** Cuando la función f (x) es lineal, $f(x) = \beta_0 + \beta_1 x$. Tiene aspecto de recta y la variable Y disminuye según aumenta X ($\beta_1 < 0$).

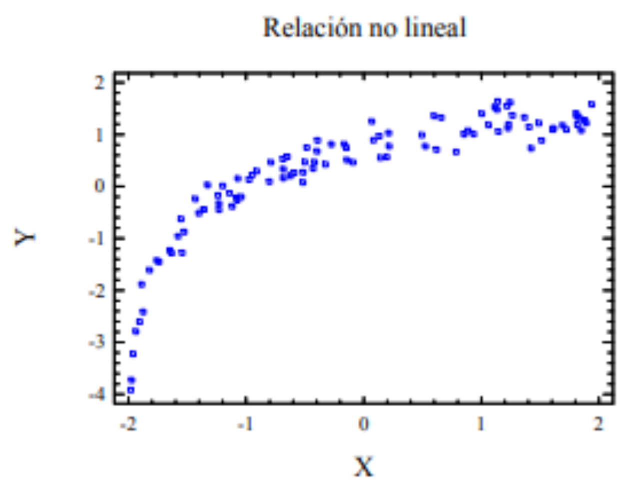
  Ejemplo: Velocidad de un vehículo y tiempo de reacción.



Definiciones: tipos de correlación

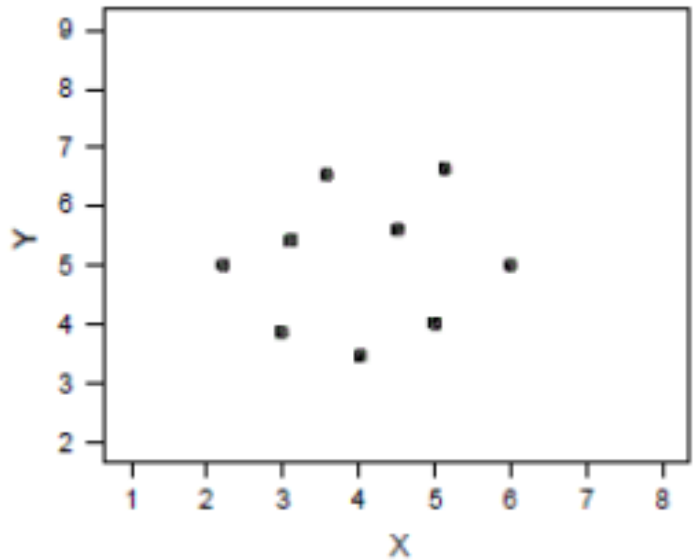
- **No lineal:** Cuando la función f(x) no es lineal. Por ejemplo, f (x) = log(x)

Ejemplo: un estudio sobre la relación entre capacidad de atención y duración de las conferencias.



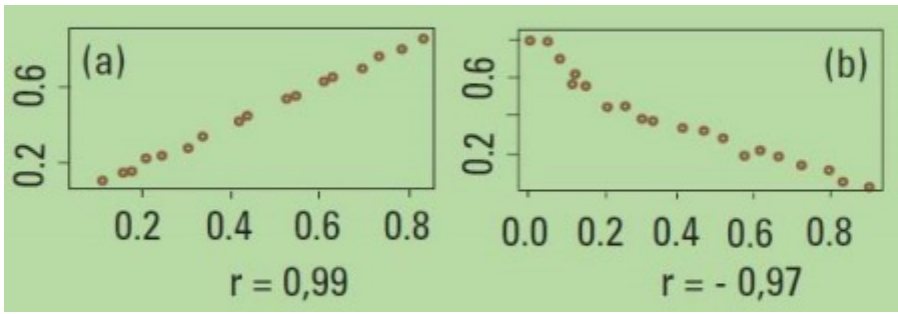
Definiciones: tipos de correlación

- **No correlados:** la dispersión entre individuos es tan grande que es imposible encontrar una relación.

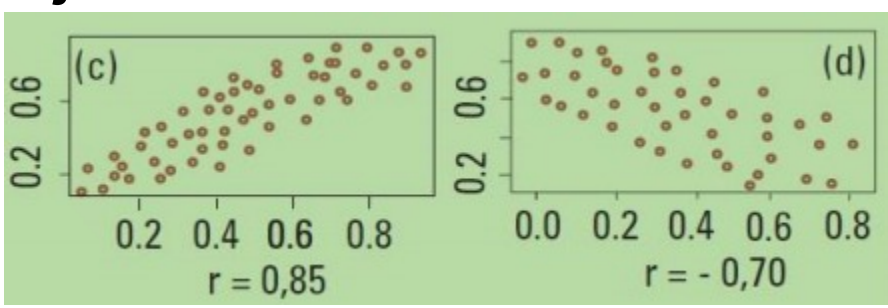


Definiciones: tipos de relación

Si r se acerca a 1 o -1 alto índice de relación lineal. El diagrama de dispersión se asemeja más a una línea.

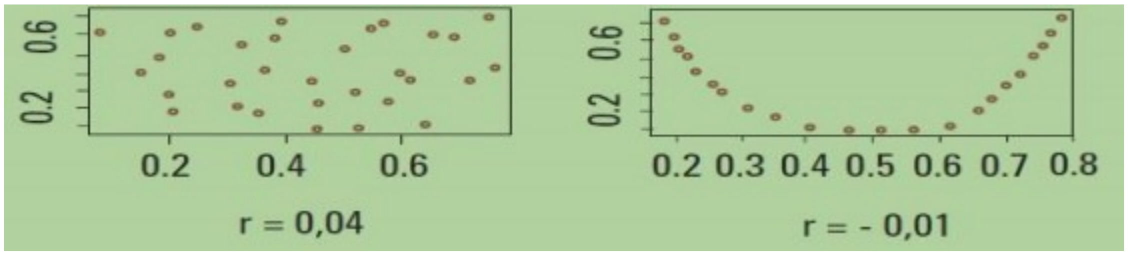


Si r se va alejando de 1 o -1 habrá relación lineal aunque menos.



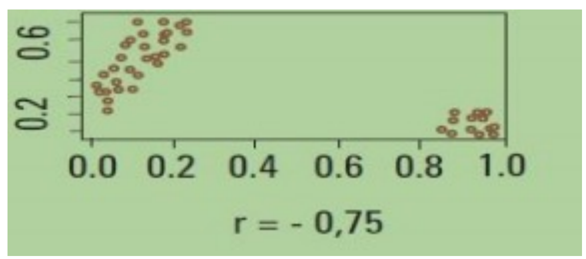
Definiciones: tipos de relación

Si r es cercano a 0 no hay relación **lineal**



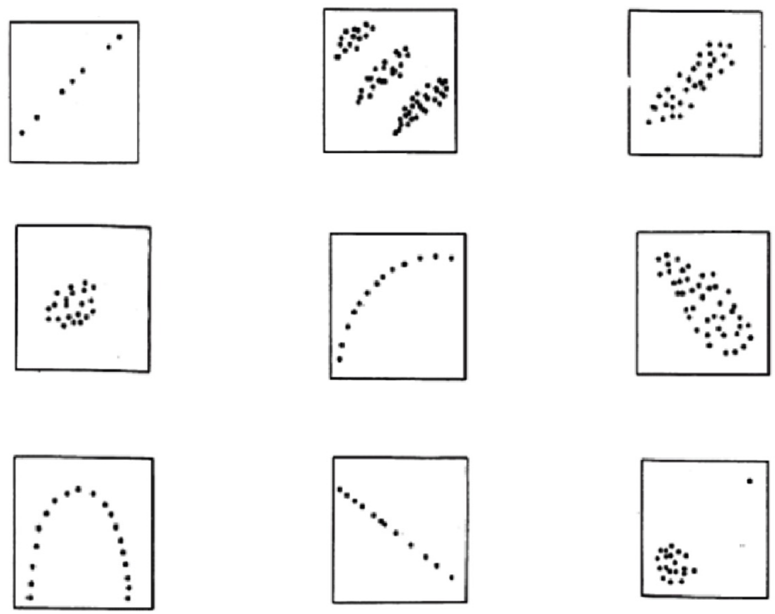
Pero **OJO!!!!**

No siempre se cumple (no lineal y no cercano a 0)



Definiciones: tipos de relación

Ejercicio: ¿Cuál es su r? ¿Tipo de relación?



Regresión lineal simple

Tomando la ecuación de una línea en geometría:

- $y = m*x + b$
- m es la pendiente
- b constante donde corta con el eje y

Usamos la ecuación de una línea en regresión:

- $\hat{y} = \beta_0 + \beta_1 x$
- y es la predicción
- x la variable explicativa
- El resto corresponden a la pendiente (β_1) y el corte con el eje y (β_0)

Regresión lineal simple

Dos objetivos:

1) Correlación: cómo de fuerte es la relación (r).

2) Regresión: estudiar como los valores de una variable pueden ser utilizados para predecir el valor de otra.

Qué se obtiene:

Relación lineal o no.

Modelo de regresión lineal.

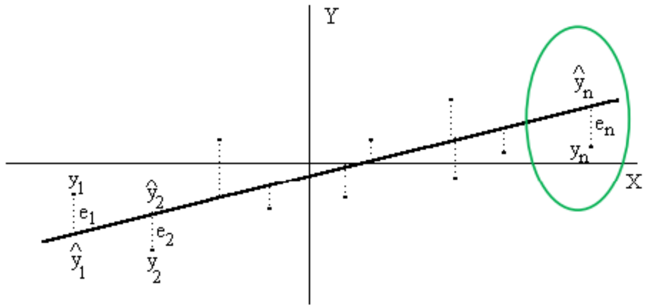
- Intercepto: punto medio del modelo, la media.
- Pendiente: efecto que tiene X sobre Y.
- Residuo: error que se comete al predecir.

Regresión lineal simple

Por lo tanto para regresión tendremos que encontrar la recta que mejor se adapte al conjunto de datos.

Para ello se usará el método de los mínimos cuadrados. Consiste en minimizar la suma de los cuadrados de los errores.

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



La suma de los cuadrados de las diferencias entre los valores reales observados y los valores de estimados.

Regresión lineal simple

Para la pendiente usaremos las medias muestrales de X e Y respectivamente. Para el intercepto se divide la covarianza muestral de X e Y entre la varianza muestral de X al cuadrado.

Pendiente (β_1); $\frac{S_{XY}}{S_X^2}$

Intercepto; $\bar{y} - \beta_1 \bar{x}$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

$$S_X^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

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

Regresión lineal simple

Ejemplo (Acidez total (Y) y libre (X) en la miel):

1) Se calculan las variables anteriores con:

$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ = 37,998 $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ = 33,8727 $S_X^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$ = 90,786 $S_{XY} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n}$ = 89,8811

2) Se calculan los valores de la recta:

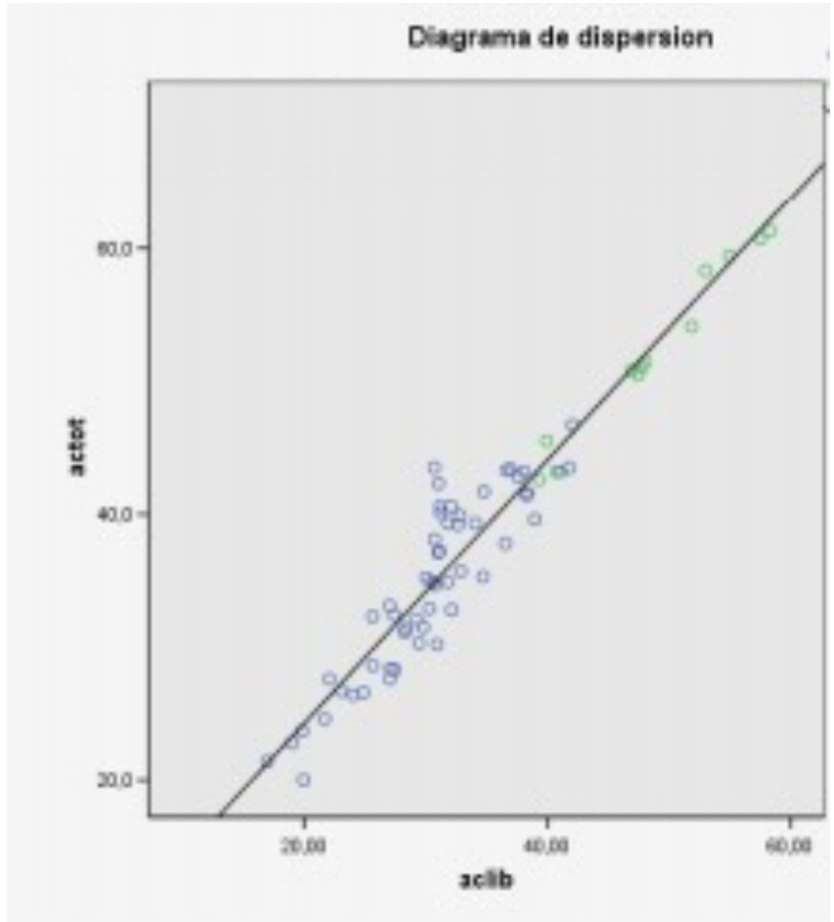
Pendiente = 89.8811/90,786 = 0,990 Intercepto = 33,8727 - 0,990*37,998 = 4,469

3) Se crea el modelo

$$\hat{Y} = 4.469 + 0.990X$$

Regresión lineal simple

4) Se dibuja e interpreta



Regresión lineal simple

Ej (edad y peso):

Id	Edad	Peso
1	2	14
2	3	20
3	5	32
4	7	42
5	8	44

Sumatorios

X_i	Y_i	X_i^2	Y_i^2	X_i*Y_i
2	14	4	196	28
3	20	9	400	60
5	32	25	1024	160
7	42	49	1764	294
8	44	64	1936	352
25	**152**	**151**	**5320**	**894**

1) Tipo de correlación

$$\bar{x} = \frac{25}{5} = 5$$

$$\bar{y} = \frac{152}{5} = 30,4$$

$Sx^2 = 151/5 - 5^2 = 5,2$

$Sy^2 = 5320/5 - 30,4^2 = 139,84$

$Sxy = 894/5 - 5*30,4 = 26,8$

$$r = \frac{cov\ muestra\ (X,Y)}{\sqrt{S_x{}^2 . sy^2}}$$

$$r = \frac{26,8}{\sqrt{5,2 * 139.84}} = 0,99$$

Correlación lineal muy alta positiva

2) Modelos de regresión

Sxy/Sy²

$x - 5 = 0,192*(y - 30,4) \rightarrow x = 0,192*y - 0,76$

$y - 30,4 = 5,15*(x - 5) \rightarrow y = 5,15*x + 4,65$

Sxy/Sx²

Regresión lineal simple

Ej (Clientes que entran a un comercio y su distancia al transporte público):

Nº Clientes	Distancia
15	8
19	7
25	6
23	4
34	2

Regresión lineal múltiple

Las técnicas de regresión lineal múltiple parten de (k+1) variables cuantitativas, siendo Y la variable de respuesta y (X_1 ,X_2 , … , X_k) las variables explicativas.

Se trata de extender a las 'k' variables las técnicas de la regresión lineal simple. En esta línea, la variable Y se puede expresar mediante una función lineal de las variables (X_1 ,X_2 , … , X_k)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_K$$

Regresión lineal múltiple

Ejemplo (Productividad de un empleado):
Un empresario quiere examinar los efectos de los cursos de formación sobre la productividad de los trabajadores. Se tiene en cuenta que el salario de los trabajadores se establece según su experiencia y sus estudios.

La ecuación sería la siguiente:

Salario=f(educación,experiencia,formación)

educación: tiempo dedicado a estudios.
experiencia: años trabajando en el sector.
formación: tiempo dedicado a cursos de formación.

Regresión lineal múltiple

salario = β_0 + $\beta_1$educación + β_2 experiencia + β_3 $formación$ + u

β_0 = intercepto. Donde la línea corta con el eje Y.

β_1 β_2 β_3 = variables que influyen en el salario.

u = Residuo.

u lo determinan todos aquellos factores que influyen en el salario que no han sido variables incluidas en la ecuación. Por ejemplo: habilidad innata, calidad de la educación o el entorno familiar.

Regresión lineal múltiple

$$\text{salario} = 0{,}284 + 0{,}092\ \text{educación} + 0{,}0041\,\text{experiencia} + 0{,}022\ \text{formación}$$

El coeficiente 0,092 significa que si mantenemos fijos “experiencia” y “formación”, un año más de “educación” predice un aumento de 0,092 en salario, lo que se traduce en un incremento aproximadamente el 9,2%. Es decir que si escogemos a dos personas con los mismos niveles de experiencia laboral y formación en la empresa, el coeficiente de educación mide la diferencia proporcional en el valor predicho de su salario si sus niveles de formación académica difieren en un año.

Si aumentasen dos variables. Por ejemplo experiencia y formación:
0,0041 + 0,022 = 0,0261 = 2,6%

Regresión lineal múltiple

Ej (Gastos, ingresos y miembros):

Gasto de alimentación	Ingresos	Tamaño
0,43	2,10	3
0,31	1,10	4
0,32	0,90	5
0,46	1,60	4
1,25	6,20	4
0,44	2,30	3
0,52	1,80	6
0,29	1,00	5
1,29	8,90	3
0,35	2,40	2
0,35	1,20	4
0,78	4,70	3
0,43	3,50	2
0,47	2,90	3
0,38	1,40	4

Regresión lineal múltiple

$$Y=\begin{bmatrix}0,43\\0,31\\0,32\\0,46\\1,25\\0,44\\0,52\\0,29\\1,29\\0,35\\0,35\\0,78\\0,43\\0,47\\0,38\end{bmatrix}= X\beta + U = \begin{bmatrix}1 & 2,1 & 3\\1 & 1,1 & 4\\1 & 0,9 & 5\\1 & 1,6 & 4\\1 & 6,2 & 4\\1 & 2,3 & 3\\1 & 1,8 & 6\\1 & 1 & 5\\1 & 8,9 & 3\\1 & 2,4 & 2\\1 & 1,2 & 4\\1 & 4,7 & 3\\1 & 3,5 & 2\\1 & 2,9 & 3\\1 & 1,4 & 4\end{bmatrix}\begin{bmatrix}\beta_0\\\beta_1\\\beta_2\end{bmatrix}+\begin{bmatrix}u_0\\u_1\\u_2\end{bmatrix}$$

Gasto de alimentación	Ingresos	Tamaño
0,43	2,10	3
0,31	1,10	4
0,32	0,90	5
0,46	1,60	4
1,25	6,20	4
0,44	2,30	3
0,52	1,80	6
0,29	1,00	5
1,29	8,90	3
0,35	2,40	2
0,35	1,20	4
0,78	4,70	3
0,43	3,50	2
0,47	2,90	3
0,38	1,40	4

Regresión lineal múltiple

	Y_i	X_{1i}	X_{2i}	X_{1i}^2	X_{2i}^2	$X_{1i}X_{2i}$	$X_{1i}Y_i$	$X_{2i}Y_i$
	0,43	2,10	3	4,41	9	6,3	0,903	1,29
	0,31	1,10	4	1,2	16	4,4	0,341	1,24
	0,32	0,90	5	0,81	25	4,5	0,288	1,6
	0,46	1,60	4	2,56	16	6,4	0,736	1,84
	1,25	6,20	4	38,44	16	24,8	7,750	5
	0,44	2,30	3	5,29	9	6,9	1,012	1,32
	0,52	1,80	6	3,24	36	10,8	0,936	3,12
	0,29	1,00	5	1	25	5	0,29	1,45
	1,29	8,90	3	79,21	9	26,7	11,481	3,87
	0,35	2,40	2	5,76	4	4,8	0,84	0,7
	0,35	1,20	4	1,44	16	4,8	0,42	1,4
	0,78	4,70	3	22,09	9	14,1	3,666	2,34
	0,43	3,50	2	12,25	4	7	1,505	0,86
	0,47	2,90	3	8,41	9	8,7	1,363	1,41
	0,38	1,40	4	1,96	16	5,6	0,532	1,52
$\sum =$	**8,07**	**42**	**55**	**188,08**	**219**	**140,8**	**32,063**	**28,96**

Regresión lineal múltiple

$$\left.\begin{aligned}
\sum_{i=1}^{15} Y_i &= N\beta_0 + \beta_1 \sum_{i=1}^{15} X_{1i} + \beta_2 \sum_{i=1}^{15} X_{2i} \\
\sum_{i=1}^{15} X_{1i} Y_i &= \beta_0 \sum_{i=1}^{15} X_{1i} + \beta_1 \sum_{i=1}^{15} X_{1i}^2 + \beta_2 \sum_{i=1}^{15} X_{1i} X_{2i} \\
\sum_{i=1}^{15} X_{2i} Y_i &= \beta_0 \sum_{i=1}^{15} X_{2i} + \beta_1 \sum_{i=1}^{15} X_{1i} X_{2i} + \beta_2 \sum_{i=1}^{15} X_{2i}^2
\end{aligned}\right\} \Rightarrow \left.\begin{aligned}
15\beta_0 + 42\beta_1 + 55\beta_2 &= 8{,}07 \\
42\beta_0 + 188{,}08\beta_1 + 140{,}08\beta_2 &= 32{,}063 \\
55\beta_0 + 140{,}08\beta_1 + 219\beta_2 &= 28{,}96
\end{aligned}\right\}$$

$$\begin{bmatrix} 15 & 42 & 55 \\ 42 & 188{,}08 & 140{,}08 \\ 55 & 140{,}08 & 219 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 8{,}07 \\ 32{,}063 \\ 28{,}96 \end{bmatrix}$$

Regresión lineal múltiple

$$\begin{bmatrix}\beta_0\\ \beta_1\\ \beta_2\end{bmatrix} = \begin{bmatrix}15 & 42 & 55\\ 42 & 188{,}08 & 140{,}08\\ 55 & 140{,}08 & 219\end{bmatrix}^{-1} \begin{bmatrix}8{,}07\\ 32{,}063\\ 28{,}96\end{bmatrix}$$

$$\begin{bmatrix}\beta_0\\ \beta_1\\ \beta_2\end{bmatrix} = \begin{bmatrix}1{,}398 & -0{,}096 & -0{,}289\\ -0{,}096 & 0{,}016 & 0{,}013\\ -0{,}289 & 0{,}013 & 0{,}068\end{bmatrix} \begin{bmatrix}8{,}07\\ 32{,}063\\ 28{,}96\end{bmatrix}$$

$$\begin{bmatrix}\beta_0\\ \beta_1\\ \beta_2\end{bmatrix} = \begin{bmatrix}-0{,}165\\ 0{,}114\\ 0{,}538\end{bmatrix} \Rightarrow$$

$Y = -0{,}166 + 0{,}114X_1 + 0{,}538X_2 + u$

Regresión lineal múltiple

Ej (Ganancias (Y), experiencia(X_1) y horas dedicadas(X_2)):

$$\sum_{k=0}^{n} Yi = 1625{,}5 \qquad \sum_{k=0}^{n} Yi\,X1i = 4862{,}9 \qquad \sum_{k=0}^{n} X1i^2 = 46$$

$$\sum_{k=0}^{n} X1i = 14 \qquad \sum_{k=0}^{n} Yi\,X2i = 63196{,}9 \qquad \sum_{k=0}^{n} X2i^2 = 7477$$

$$\sum_{k=0}^{n} X2i = 193 \qquad \sum_{k=0}^{n} X1iX2i = 549 \qquad n = 5$$

$$M^{-1} = \begin{bmatrix} 76{,}650 & 2{,}3045 & -2{,}1477 \\ 2{,}3045 & 0{,}2450 & -0{,}0775 \\ -2{,}1477 & -0{,}0775 & 0{,}0613 \end{bmatrix}$$

Regresión lineal múltiple: análisis residuos

Calculo de u

Residuos ordinarios: Se define el residuo (ordinario) asociado a una observación muestral como la diferencia entre la observación y la predicción.

$$e_i = y_i - \hat{y}_i$$

Una familia gasta 0,43 cuando ingresa 2,10 y tiene 3 miembros.

$\hat{y}_i$ =-0,166+0,144X_1+0,077X_2=-0,166+0,144*(2,10)+0,077*(3)= 0,3674

e_i = 0,43 – 0,3674 = 0,0262

Regresión lineal múltiple: análisis residuos

Ej (Gastos, ingresos y miembros):

$$\sum_{i=1}^{15}(Y_1 - \hat{Y}_1)^2 = 0{,}0721$$

Predicciones	Residuos	u_i^2
0,3839	0,046	0,0021
0,3119	-0,002	0,0000
0,3591	-0,039	0.0015
0,3864	0,074	0,0054
1,0718	0,178	0,0318
0,4137	0,26	0,0007
0,5702	-0,050	0,0025
0,374	-0,084	0,0071
1,3971	-0,107	0,0115
0,3516	-0,002	0,0000
0,3268	0,023	0,0005
0,7713	0,009	0,0001
0,5155	-0,086	0,0073
0,5031	-0,033	0,0011
0,3566	0,023	0,0005

El modelo final es:

$$Y = -0{,}16 + 0{,}149X_1 + 0{,}077X_2 + 0{,}0721$$

Regresión lineal múltiple: análisis residuos

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{\substack{\text{SCT} \\ \text{suma cuadrados total} \\ (n-1)\text{ grados libertad}}} = \underbrace{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}_{\substack{\text{SCE} \\ \text{suma cuadrados explicada} \\ k\text{ grados libertad}}} + \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{\substack{\text{SCR} \\ \text{suma cuadrados residual} \\ (n-k-1)\text{ grados libertad}}}$$

SCT: la desigualdad total (entre lo que gastan las familias).

SCE: la cantidad que depende de las variables explicativas (miembros e ingresos).

SCR: la cantidad que depende de otras variables.

Regresión lineal múltiple: análisis residuos

SCT=1,4316 SCE=0,0721 SCR=1,3595

Los tenemos que interpretar como medida de desigualdad entre los gastos de las familias igual a 1,4316, de los cuales 1,3595 son imputables a la los ingresos familiares y al número de miembros, mientras que 0,0721 representa la cantidad de la variación de gastos no explicada por las variables exógenas y que, por tanto, se puede asignar a otras causas.

Regresión lineal múltiple: bondad ajuste

Una vez tenemos el plano de regresión, es necesario saber si el ajuste que ofrece sobre la nube de puntos es suficientemente bueno.

Es decir, se trata de saber si el modelo que se ha ajustado para relacionar las variables X e Y es un modelo consistente. Mediante el Coeficiente de Determinación.

Regresión lineal múltiple: bondad ajuste

El coeficiente de determinación (R^2) representa la proporción de varianza de Y explicada por las variables implicadas en el modelo de regresión ajustado a los datos (X en el modelo de regresión lineal simple).

Este coeficiente oscilará siempre entre 0 y 1, de modo que cuanto más próximo sea a 1, indicará mejor bondad de ajuste del modelo de regresión a la distribución conjunta de las variables. Si R^2 es igual a 1, el ajuste será perfecto.

Regresión lineal múltiple: bondad ajuste

Para ello usamos el Coeficiente de Determinación:

$$R^2 = \frac{SCE}{SCT} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} = \frac{1,3595}{1,4316} = 0,9496$$

Indica que el ajuste es grande en concreto de un 94,95%.

Regresión lineal múltiple: intervalos confianza

Es el intervalo en el que variará la variable dependiente Y con respecto a alguna variable independiente X.

Ejemplo de cómo varia la distancia del morro al pectoral (Y) de un pez con respecto a su distancia del morro a la aleta (X).

Dado un intervalo de confianza para X de 0,606 a 0,811 cm y la confianza es 0,95. Esto quiere decir que para un 95 % de la muestra por cada cm de X se asegura que Y se incrementa por término medio entre 0.606 y 0.811 cm.

Regresión lineal múltiple: intervalos confianza

Para calcularlo, se usará la distribución de t-student con un % confianza y con G grados de libertad.

La confianza viene dada por el problema (a veces en %).

G se calcula por la diferencia entre el número de individuos de la población y el número de parámetros a estimar. En el caso del gastos familiar.

$$G = 15 - 3 = 12$$

15 familias

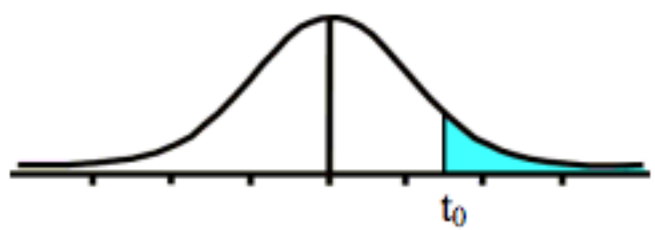
Ingresos, miembros y gasto

Regresión lineal múltiple: intervalos confianza

Los valores de la confianza vienen dados por la ecuación 1 - $\alpha = confianza$

Si buscásemos una confianza del 90%; 1 - α =0,9→ α =0,1

Cómo hay un t a cada lado de la distribución α se divide entre 2, siendo 0,05.



Regresión lineal múltiple: intervalos confianza

Teniendo los grados de libertad y α.

Podemos calcular la $t_{\alpha G}$ usando la tabla t-student. En este caso es 1,7823. 12 grados de libertad (G) y 0,05 (α)

Grados de libertad	α 0.25	0.1	0.05	0.025	0.01	0.005
1	1.0000	3.0777	6.3137	12.7062	31.8210	63.6559
2	0.8165	1.8856	2.9200	4.3027	6.9645	9.9250
3	0.7649	1.6377	2.3534	3.1824	4.5407	5.8408
4	0.7407	1.5332	2.1318	2.7765	3.7469	4.6041
5	0.7267	1.4759	2.0150	2.5706	3.3649	4.0321
6	0.7176	1.4398	1.9432	2.4469	3.1427	3.7074
7	0.7111	1.4149	1.8946	2.3646	2.9979	3.4995
8	0.7064	1.3968	1.8595	2.3060	2.8965	3.3554
9	0.7027	1.3830	1.8331	2.2622	2.8214	3.2498
10	0.6998	1.3722	1.8125	2.2281	2.7638	3.1693
11	0.6974	1.3634	1.7959	2.2010	2.7181	3.1058
12	0.6955	1.3562	1.7823	2.1788	2.6810	3.0545
13	0.6938	1.3502	1.7709	2.1604	2.6503	3.0123
14	0.6924	1.3450	1.7613	2.1448	2.6245	2.9768
15	0.6912	1.3406	1.7531	2.1315	2.6025	2.9467
16	0.6901	1.3368	1.7459	2.1199	2.5835	2.9208
17	0.6892	1.3334	1.7396	2.1098	2.5669	2.8982
18	0.6884	1.3304	1.7341	2.1009	2.5524	2.8784
19	0.6876	1.3277	1.7291	2.0930	2.5395	2.8609
20	0.6870	1.3253	1.7247	2.0860	2.5280	2.8453

Regresión lineal múltiple: intervalos confianza

Por último se calculan los intervalos de confianza para cada parámetro $IC(\beta_i)$=[$\widehat{\beta_i} \pm t * \sqrt{var(\beta_i)}$]

$\text{Var}(\beta_0) = 0{,}00816 \qquad \text{Var}(\beta_1) = 0{,}000096 \qquad \text{Var}(\beta_2) = 0{,}0004$

$\hat{\beta}_0 = -0{,}160 \qquad \hat{\beta}_1 = 0{,}149 \qquad \hat{\beta}_2 = 0{,}077 \qquad t_{0,05,12} = 1{,}782$

$$IC_{1-\alpha}(\beta_0) = \left[-0{,}160 \pm (1{,}782)\sqrt{0{,}00816}\right] = \left[-0{,}321 ; 0{,}001\right]$$

$$IC_{1-\alpha}(\beta_1) = \left[0{,}149 \pm (1{,}782)\sqrt{0{,}000096}\right] = \left[0{,}1315 ; 0{,}1665\right] \quad \text{(Ingreso)}$$

$$IC_{1-\alpha}(\beta_2) = \left[0{,}077 \pm (1{,}782)\sqrt{0{,}0004}\right] = \left[0{,}0414 ; 0{,}1126\right] \quad \text{(Tamaño)}$$

Regresión lineal múltiple: contraste hipótesis

El contraste de hipótesis es una metodología de inferencia diseñada para valorar si una propiedad de la población es compatible con la información muestral. Dicho de otra manera, como influyen las variables X en Y.

Para ello se trabajará con la hipótesis nula H_0 y la hipótesis alternativa H_1.

Regresión lineal múltiple: contraste hipótesis

Se denomina hipótesis nula, H_0, a la hipótesis que se desea contrastar frente a una hipótesis alternativa, H_1, que es la negación de la nula. El adjetivo «nula» indica que ésta es la hipótesis que se mantendrá como cierta a no ser que los datos indiquen su falsedad.

Para refutar la hipótesis nula:

Cero tiene que pertenecer al intervalo de confianza.

Regresión lineal múltiple: contraste hipótesis

Para el ejemplo actual, la hipótesis nula se cumple para la variable de los ingresos por familia y para la que índica el número de miembros que tiene.

$$IC_{1-\alpha}(\beta_1) = \left[0,1315 ; 0,1665\right] \quad \text{(Ingreso)}$$

$$IC_{1-\alpha}(\beta_2) = \left[0,0414 ; 0,1126\right] \quad \text{(Tamaño)}$$

Regresión lineal múltiple: intervalos confianza

Ejercicio (Gastos publicidad y visualizaciones):

Y = 0,5895 + 0,936*X1 + 0,1866*X2

Para un 95% de la población y con varianzas 0,3454, 1,5438 y 2,3167

Regresión logística

En general, la regresión logística es adecuada cuando la variable de respuesta Y es categórica múltiple (admite varias categorías de respuesta, tales como mejora mucho, empeora, se mantiene, mejora, mejora mucho), pero es especialmente útil en particular cuando solo hay dos posibles respuestas que es el caso más común.

Aplicaciones:

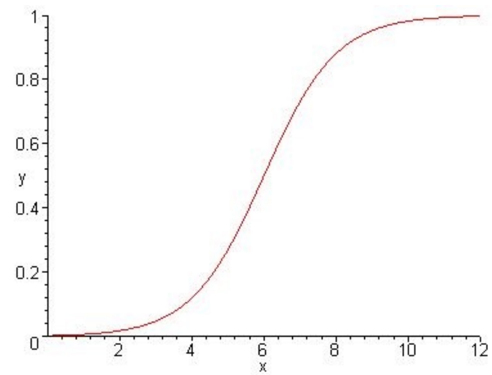
- Un paciente deja de fumar o no después del tratamiento.
- Un paciente muere o no antes del alta.
- Un paciente positivo al VIH está o no en el estado IV.

Regresión logística

Se basa en la función logística y se expresa de la siguiente manera:

$$P(C=1|\mathbf{x}) = P(C=1|X_1=x_1,\ldots,X_n=x_n) = \frac{1}{1+e^{-\left(\beta_0+\sum_{i=1}^{n}\beta_i x_i\right)}}$$

El modelo se formaliza a partir de de la función logística, donde se plantea una relación funcional, aquí la función es la logística, una función con forma de curva sigmoidal:



Regresión logística

Probabilidad de sufrir una enfermedad coronaria (C) teniendo en cuenta el nivel de colesterol (X_1), la edad (X_2) y el electrocardiograma (X_3)

$$\beta_0=-3{,}911 \quad \beta_1=0{,}652 \quad \beta_2=0{,}029 \quad \beta_3=0.342$$

¿Quién tiene mas probabilidad de sufrirla? Entre dos individuos de 40 años con un electrocardiograma normal. Uno con nivel de colesterol alto y otro no.

$$P(C=1|\mathbf{x}) = P(C=1|X_1=1, X_2=40, X_3=0) = \frac{1}{1+e^{-(-3{,}911+0{,}652(1)+0{,}029(40)+0{,}342(0))}} = 0{,}109$$

$$P(C=1|\mathbf{x}') = P(C=1|X_1=0, X_2=40, X_3=0) = \frac{1}{1+e^{-(-3{,}911+0{,}652(0)+0{,}029(40)+0{,}342(0))}} = 0{,}060$$

Regresión logística

El análisis de regresión logística tiene dos modalidades: la regresión logística binaria cuando se pretende explicar una característica dicotómico (estar desempleado o no), y la regresión logística multinomial en el caso de querer explicar una variable cualitativa politómica.

Este segundo caso se diferencia la situación en que la variable categórica es politómica nominal (la elección de una marca de un producto o la filiación política) o politómica ordinal (el nivel salarial o el grado de acuerdo sobre una cuestión).

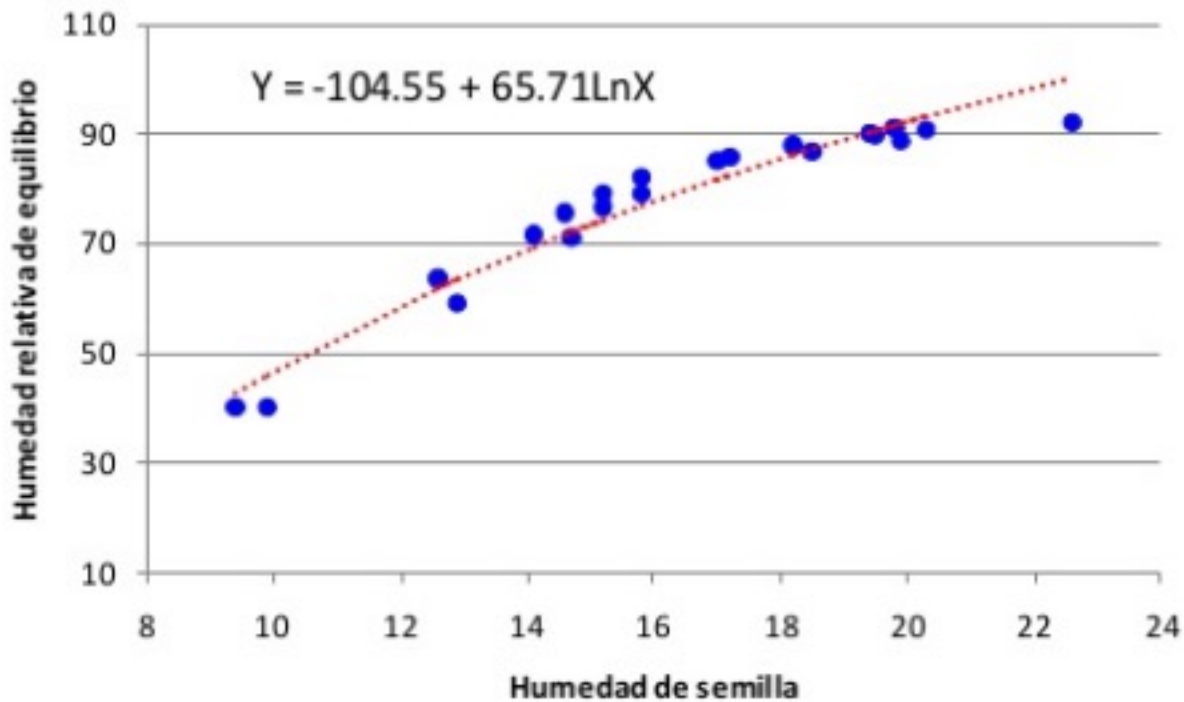
Regresión no lineal

Las variables pueden no tener una relación lineal. Por lo que hay que linealizarlos mediante la aplicación de logaritmos:

Modelo	Ecuación	Ecuación Linealizada
Logarítmico	$e^{Y} = b_0 X^{b_1}$	$\hat{Y} = Lnb_0 + b_1 LnX$
Exponencial	$\hat{Y} = b_0 b_1^X$	$Ln\hat{Y} = Lnb_0 + Lnb_1 X$
Exponencial	$\hat{Y} = b_0 e^{b_1 X}$	$Ln\hat{Y} = Lnb_0 + b_1 X$
Doble Logarítmico o Potencia	$\hat{Y} = b_0 X^{b_1}$	$Ln\hat{Y} = Lnb_0 + b_1 LnX$

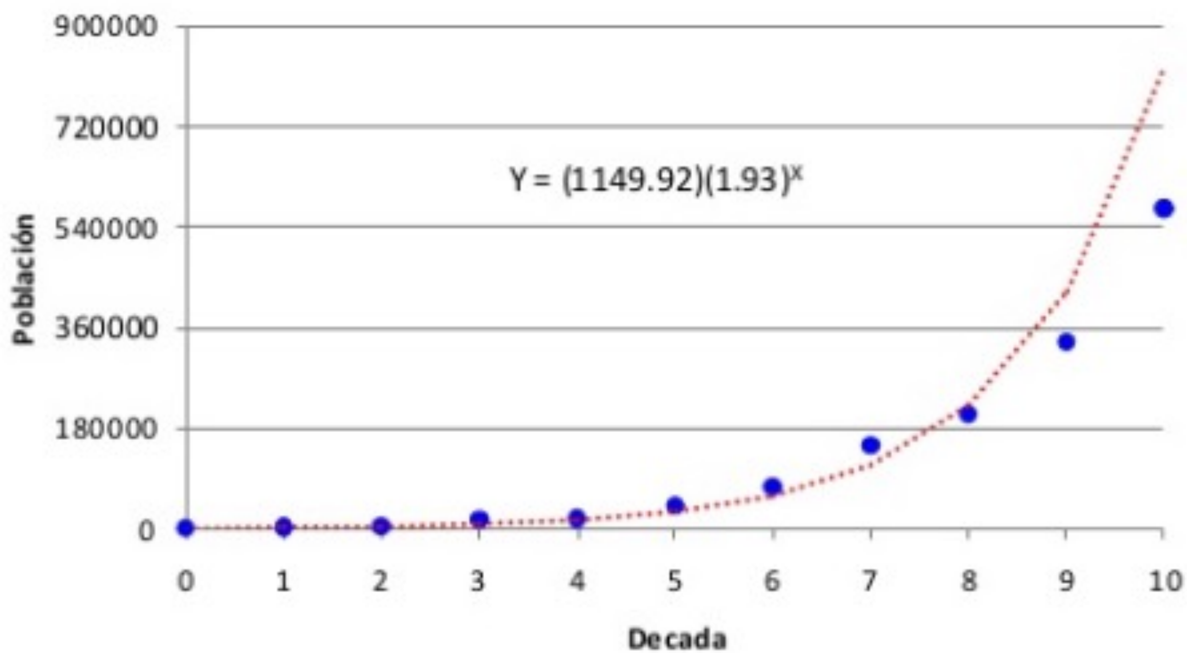
Regresión no lineal

Logarítmica (humedad semilla vs humedad relativa equilibrio):



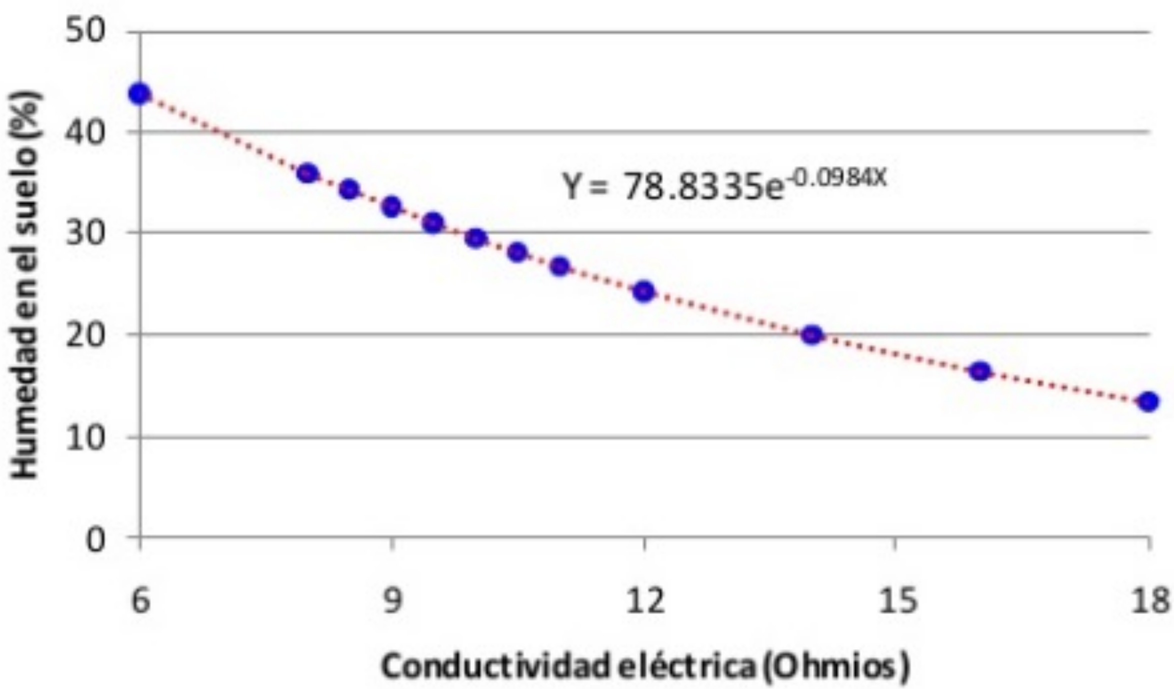
Regresión no lineal

Exponencial (tiempo vs población):



Regresión no lineal

Exponencial (humedad en suelo de yeso vs conductividad):



Regresión no lineal

Potencia (diámetro cebolla vs peso cebolla):