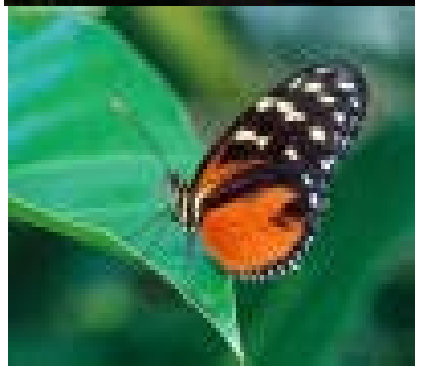
REGRESIÓN LINEAL SIMPLE

1. **El problema de la regresión lineal simple**
2. **Método de mínimos cuadrados**
3. **Coeficiente de regresión**
4. **Coeficiente de correlación lineal**
5. **El contraste de regresión**
6. **Inferencias acerca de los parámetros**
7. **Inferencias acerca de la predicción**
8. **Los supuestos del modelo de regresión lineal**
9. **Un ejemplo en donde no se cumplen los supuestos**

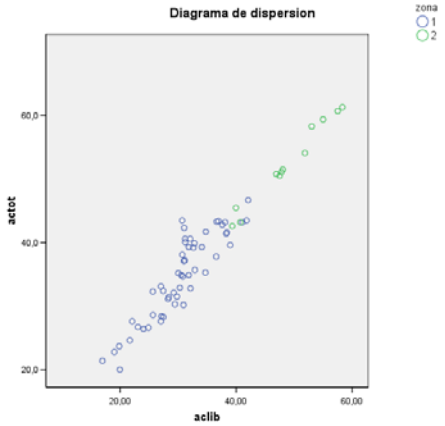


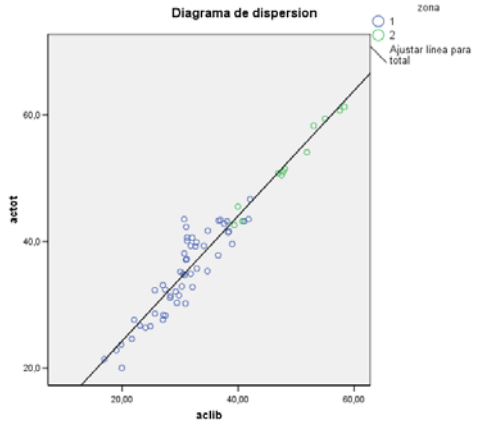
1. El problema de la regresión lineal simple

El objetivo de un modelo de regresión es tratar de explicar la relación que existe entre una variable dependiente (variable respuesta) *Y* un conjunto de variables independientes (variables explicativas) $X_1,\ldots, X_n$.

En un modelo de **regresión lineal simple** tratamos de explicar la relación que existe entre la variable respuesta *Y* y una única variable explicativa *X*.

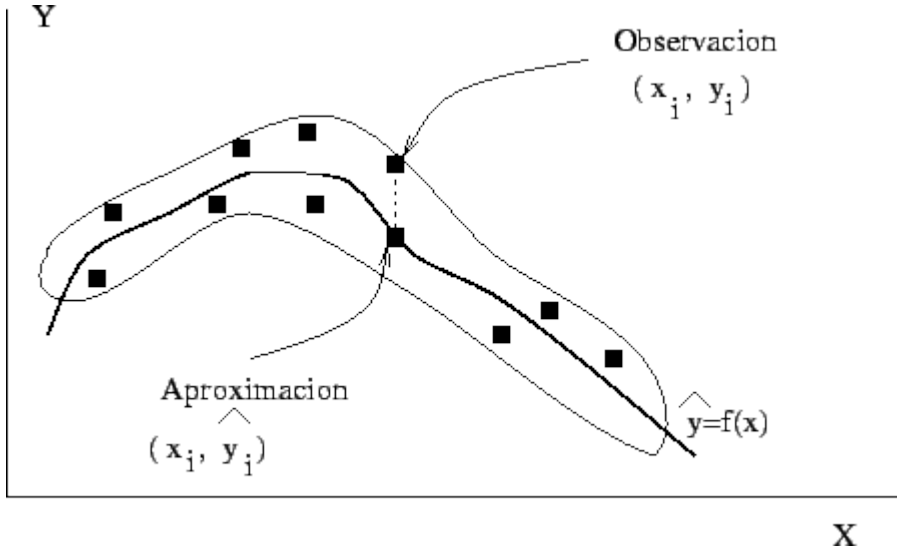
Ejemplo: En la muestra de la miel vamos a ver si existe relación lineal entre la acidez libre (AcLib) y la acidez total (AcTot). Para ver si un modelo de regresión lineal tiene sentido, comenzamos dibujando un diagrama de dispersión.





Mediante las técnicas de regresión de una variable *Y* sobre una variable *X*, buscamos una función que sea una buena aproximación de una nube de puntos (x_i,y_i), mediante una curva del tipo:

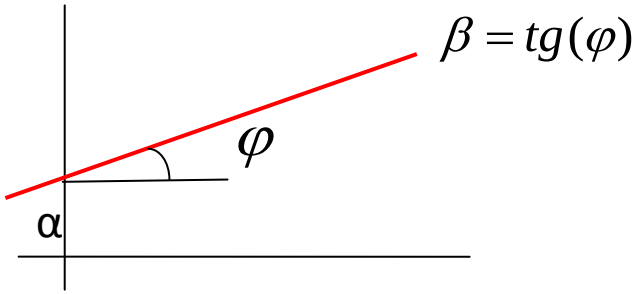
$$\hat{Y} = f(X)$$



El modelo de regresión lineal simple tiene la siguiente expresión:

$$Y = \alpha + \beta X + \varepsilon,$$

En donde α es la ordenada en el origen (el valor que toma *Y* cuando *X* vale 0), β es la pendiente de la recta (e indica cómo cambia *Y* al incrementar *X* en una unidad) y ε una variable que incluye un conjunto grande de factores, cada uno de los cuales influye en la respuesta sólo en pequeña magnitud, a la que llamaremos error. *X* e *Y* son variables aleatorias, por lo que no se puede establecer una relación lineal exacta entre ellas.



2. Método de mínimos cuadrados

Para hacer una estimación del modelo de regresión lineal simple, trataremos de buscar una recta de la forma:

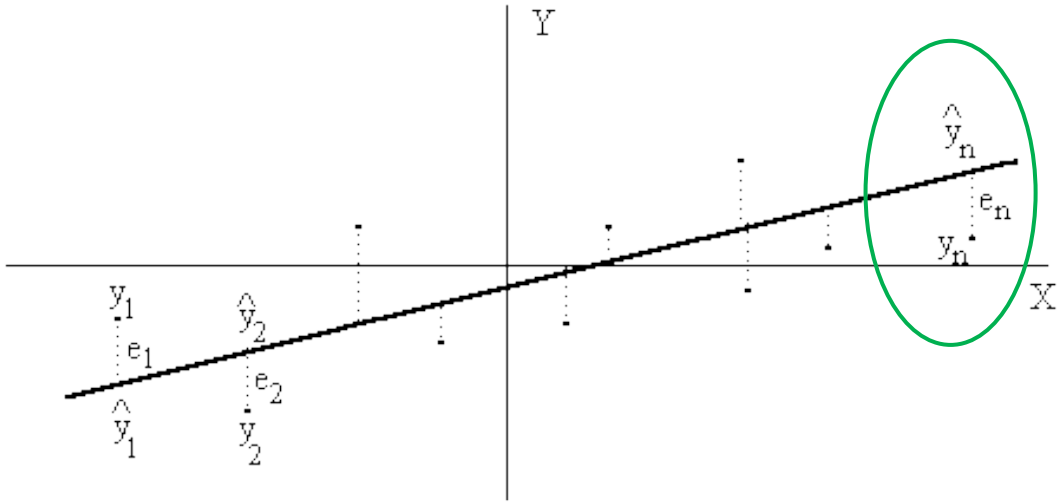
$$\boxed{\hat{Y} = \hat{\alpha} + \hat{\beta} X = a + bX}$$

de modo que se ajuste a la nube de puntos.

Para esto utilizaremos el método de mínimos cuadrados. Este método consiste en minimizarla suma de los cuadrados de los errores:

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

Es decir, la suma de los cuadrados de las diferencias entre los valores reales observados (y_i) y los valores estimados ($\hat{y}_i$).



Con este método, las expresiones que se obtiene para *a* y *b* son las siguientes:

$$a = \overline{y} - b\overline{x}, \quad b = \frac{S_{XY}}{S_X^2},$$

En donde $\overline{x}$ e $\overline{y}$ denotan las medias muestrales de *X* e *Y* (respectivamente), S_X^2 es la varianza muestral de *X* y S_{XY} es la covarianza muestral entre *X* e *Y*.

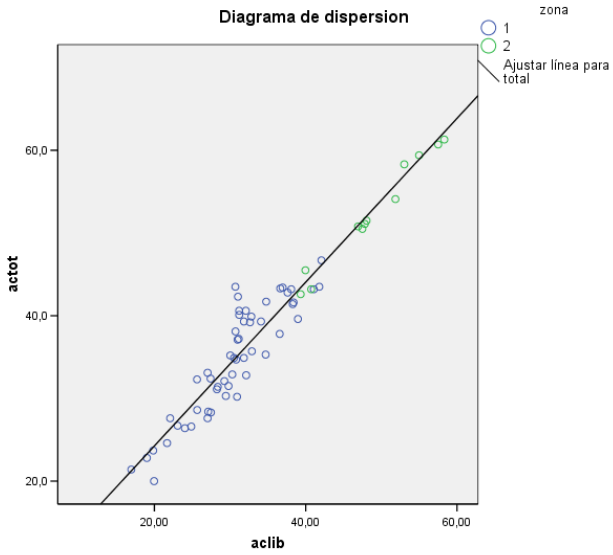
Estos parámetros se calculan como:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \quad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}, \quad S_X^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}, \quad S_Y^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n}, \quad S_{XY} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n}.$$

La cantidad b se denomina coeficiente de regresión de *Y* sobre X, lo denotamos por $b_{Y/X}$.

Ejemplo: Los estadísticos descriptivos anteriores para las variables AcTot y AcLib (acidez total y acidez libre) son los siguientes:

$$\bar{x} = 37.998, \quad \bar{y} = 33.8727,$$
$$S_X^2 = 90.786, \quad S_Y^2 = 85.459,$$
$$S_X = 9.5282 \quad S_Y = 9.24439.$$



La recta de regresión ajustada es la siguiente:

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

donde *Y* es la acidez total y *X* es la acidez libre.

Para calcular la recta de regresión de *X* sobre *Y* se hace aproximando *X* por $\hat{X}$, del modo

$$\hat{X} = a + bY$$

donde

$$a = \overline{x} - b\overline{y}, \quad b = \frac{S_{XY}}{S_Y^2},$$

Es totalmente incorrecto despejar X de la ecuación $\hat{Y} = a + bX$ para calcular la recta de regresión de *X* sobre *Y.*

Observación: La recta de regresión pasa siempre por el centro de gravedad de la nube de puntos, es decir por el punto $\left(\overline{X}, \overline{Y}\right)$ **.**

3. El coeficiente de regresión

El coeficiente de regresión nos da información sobre el comportamiento de la variable Y frente a la variable X, de manera que:

a) Si $b_{Y/X} = 0$, para cualquier valor de *X* la variable *Y* es constante (es decir, no cambia).

b) Si $b_{Y/X} > 0$, esto nos indica que al aumentar el valor de *X*, también aumenta el valor de *Y*.

c) Si $b_{Y/X} < 0$, esto nos indica que al aumentar el valor de *X*, el valor de *Y* disminuye.

En el ajuste de regresión lineal de la acidez total sobre la acidez libre se obtenía el modelo:

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

en donde *Y* es la acidez total y *X* es la acidez libre.

El coeficiente de regresión es $b_{Y/X} = 0.990 > 0$ y esto indica que al aumentar X aumenta y.

4. El coeficiente de correlación lineal

El coeficiente de correlación lineal entre *X* e *Y* viene dado por:

$$r = \frac{S_{XY}}{S_X S_Y},$$

y trata de medir la dependencia lineal que existe entre las dos variables. Su cuadrado se denomina coeficiente de determinación, r^2.

Propiedades del coeficiente de correlación:

a) No tiene dimensión, y siempre toma valores en [-1,1].

b) Si las variables son independientes, entonces r=0, pero el inverso no tiene por qué ser cierto.

c) Si existe una relación lineal exacta entre *X* e *Y*, entonces *r* valdría 1 (relación directa) ó -1 (relación inversa).

d) Si r>0, esto indica una relación directa entre las variables (es decir, que si aumentamos *X*, también aumenta *Y*).

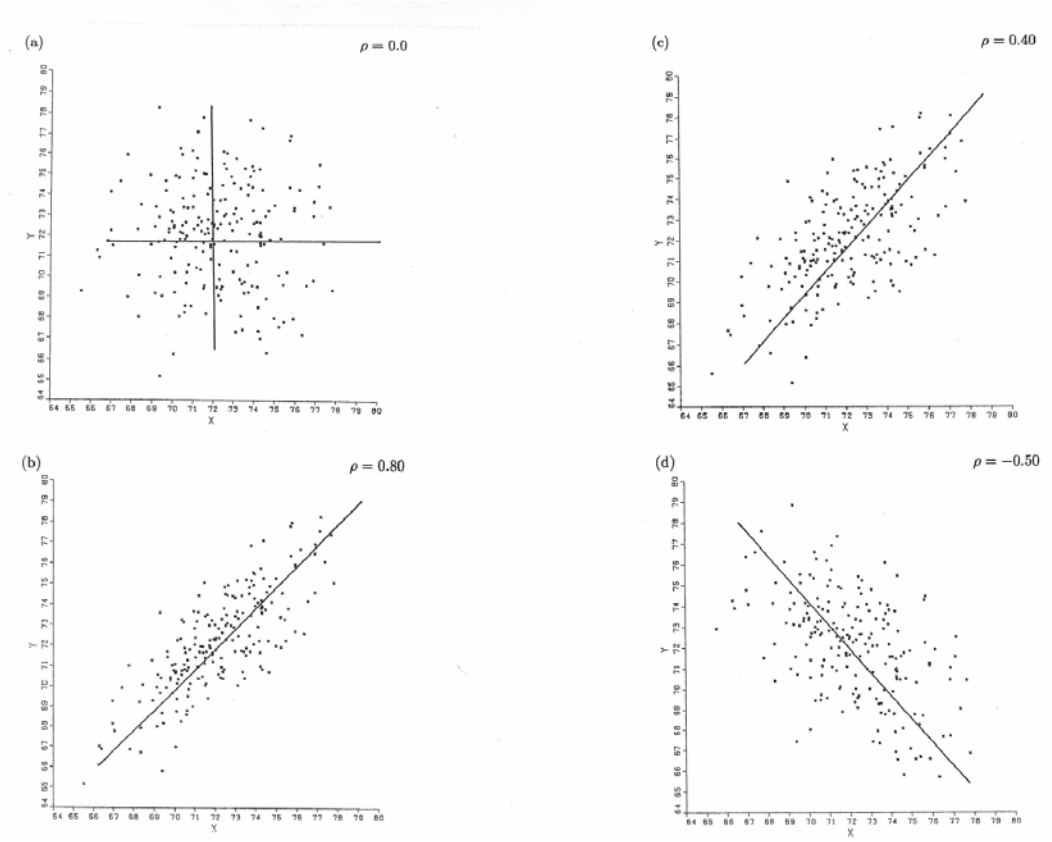
e) Si r<0, la correlación entre las variables es inversa (si aumentamos una, la otra disminuye).

Para nuestro **ejemplo** el valor de ***r* es 0.960**.
Como es positivo, esto indica que **existe una relación directa** entre las variables acidez total y acidez libre. Además su valor es próximo a 1 indicando una dependencia lineal muy fuerte.

Relación entre los coeficientes de regresión y de correlación:

$$b_{Y/X} = r\frac{S_Y}{S_X}, \quad b_{X/Y} = r\frac{S_X}{S_Y}.$$

Los dos coeficientes de regresión y el coeficiente de correlación tienen pues el mismo signo.



Descomposición de la variabilidad:

$$\sum \left(y_i - \overline{y}\right)^2 = \sum \left(\hat{y}_i - \overline{y}\right)^2 + \sum \left(y_i - \hat{y}_i\right)^2 \quad \Leftrightarrow$$
$$SCtot \quad = \quad SCR \quad + \quad SCres$$

Coeficiente de determinación (r^2):

$$r^2 = \frac{\sum \left(\hat{y}_i - \overline{y}\right)^2}{\sum \left(y_i - \overline{y}\right)^2} = \frac{SCR}{SCtot}$$

El coeficiente de determinación puede interpretarse como la **proporción de variabilidad de *Y* que es explicada por *X***. Mide la proximidad de la recta ajustada a los valores observados de Y.

5. El contraste de regresión

En el contraste de regresión contrastamos la hipótesis nula de que la pendiente de la recta es cero, es decir, que no existe relación o dependencia lineal entre las dos variables.

$$\begin{array}{|ll|}\hline H_0: & \beta = 0 \\ H_1: & \beta \neq 0 \\ \hline\end{array} \quad \Leftrightarrow \quad \begin{array}{|ll|}\hline H_0: & \rho = 0 \\ H_1: & \rho \neq 0 \\ \hline\end{array}$$

En la tabla ANOVA del análisis de regresión el estadístico F nos permite realizar dicho contraste.

Ejemplo: En el modelo de regresión para explicar la Acidez Total en función de la Acidez Libre, el análisis proporciona la siguiente tabla ANOVA:

ANOVA

S_r^2

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Regres	5442.731	1	5442.73	**759.92**	**.000**
Resid	458.379	64	7.162		
Total	5901.110	65			

Dado que la significación (P-valor) asociada al valor del estadístico F es del 0%, rechazamos la hipótesis nula de que β, o equivalentemente el coeficiente de correlación, sea nulo. Concluimos pues que **existe una relación de tipo lineal entre X e Y**.

6. Inferencias para los parámetros del modelo

a) Contrastes de hipótesis para los parámetros

- Contraste $H_0: \begin{cases} \beta = \beta^* \\ \beta \neq \beta^* \end{cases}$

Estadístico del contraste:

$$t = \frac{\hat{\beta} - \beta^*}{s(\hat{\beta})} \text{ que sigue bajo } H_0 \text{ una } t_{n-2} \qquad s(\hat{\beta}) = \frac{s_r}{\sqrt{n}} \frac{1}{s_x}$$

- Contraste $H_0: \begin{cases} \alpha = \alpha^* \\ \alpha \neq \alpha^* \end{cases}$

s_r = *error típico en la estimación* de la regresión(*lo da el SPSS*) = *media cuadrática de los residuos*

Estadístico del contraste:

$$t = \frac{\hat{\alpha} - \alpha^*}{s(\hat{\alpha})} \text{ que sigue bajo } H_0 \text{ una } t_{n-2} \qquad s(\hat{\alpha}) = \frac{s_r}{\sqrt{n}} \sqrt{1 + \frac{\overline{x}^2}{s_x^2}}$$

- **El contraste clave para este modelo es**

$$\text{Contraste} \quad H_0 : \begin{cases} \beta = 0 \\ \beta \neq 0 \end{cases}$$

Es decir el contraste de regresión.

- ✓ **Si aceptamos la hipótesis nula concluimos que no hay evidencias de que haya una relación lineal entre las variables y el modelo, en principio, no es apropiado. Puede haber una relación lineal en la población pero la muestra elegida no la detecta.**

- ✓ **Si rechazamos la hipótesis nula concluimos que el modelo lineal es apropiado. Puede que exista una relación NO-LINEAL pero los datos son también consistentes con un modelo lineal.**

b) Estimaciones por intervalo para los parámetros.

- $$IC \quad \text{para} \quad \beta: \quad (\hat{\beta} \mp \mathrm{t}_{n-2}^{\alpha/2} \frac{s_r}{\sqrt{n}} \frac{1}{s_x})$$

- $$IC \quad \text{para} \quad \alpha: \quad (\hat{\alpha} \mp \mathrm{t}_{n-2}^{\alpha/2} \frac{s_r}{\sqrt{n}} \sqrt{1+\frac{\overline{x}^2}{s_x^2}})$$

7. Inferencias acerca de la predicción

Nos puede interesar predecir el valor medio de la variable respuesta o bien el valor de la variable respuesta para un valor *x* que no ha sido considerado en la muestra. El estimador puntual es el mismo para las dos situaciones.

a) Estimación puntual del *valor medio de Y* para un valor *X=x* :

Estimador: $\hat{Y} = a + bx.$

$$s(\hat{y}) = \mathrm{s}_r \left(\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_i (x_i - \overline{x})^2} \right)^{1/2}$$ error típico de la estimación de la media de Y

➢ **Intervalo de confianza para el valor medio de la respuesta** *cuando X=x:*

$$(\hat{y}) \mp \mathrm{t}_{n-2}^{\alpha/2} \mathrm{s}_r \left(\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_i (x_i - \overline{x})^2} \right)^{1/2}$$

b) Estimación puntual del *valor de Y* para un valor *X=x* no observado

Estimador: $\hat{Y} = a + bx.$

$$s(\hat{y}) = \mathrm{s}_r \left(1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{\sum_i (x_i - \overline{x})^2} \right)^{1/2}$$ error típico de la estimación de Y

➢ Intervalo de confianza para el valor de la respuesta para una nueva observación *X=x:*

$$\left(\hat{y}\right) \mp \mathrm{t}_{n-2}^{\alpha/2} \mathrm{s}_r \left(1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{\sum_i (x_i - \overline{x})^2} \right)^{1/2}$$

c) Bandas de confianza

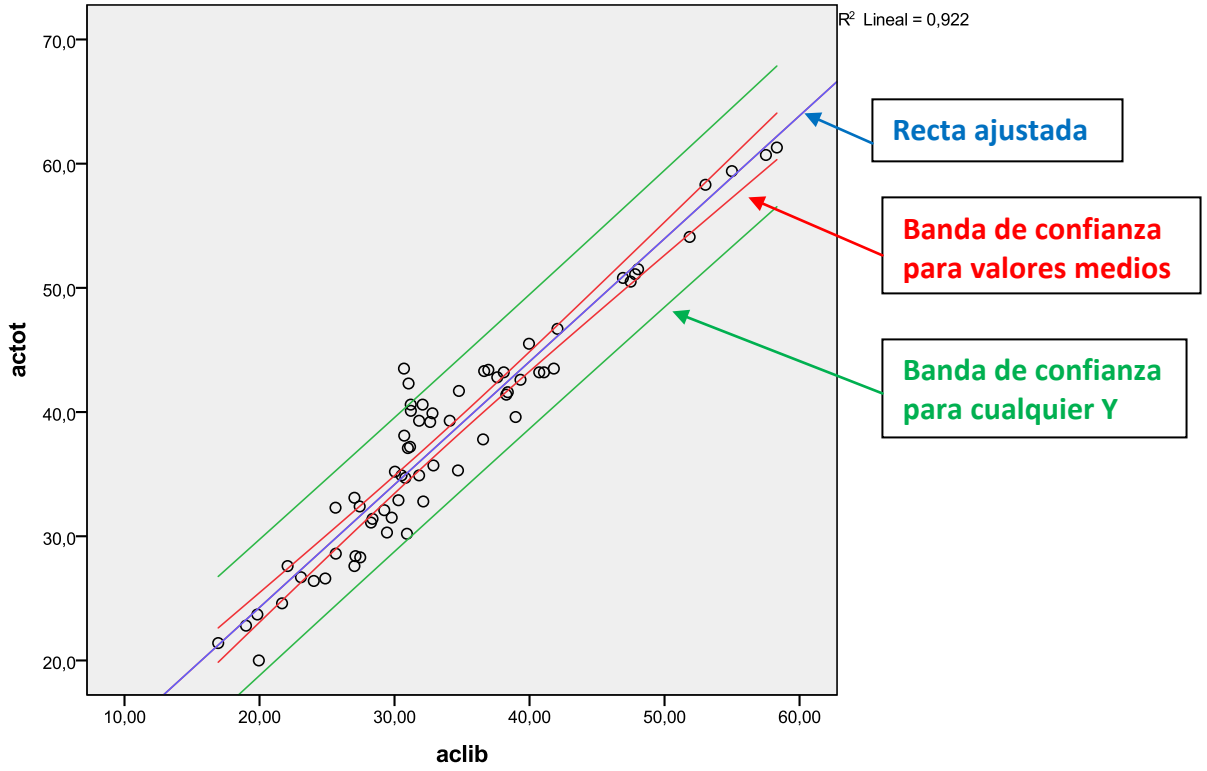
➢ **Banda de confianza *para predecir el valor medio de Y*** en cualquier valor de *X*:

$$\left(\hat{y}\right) \mp (2F_{2,n-2}^{\alpha})^{1/2} \mathrm{s}_r \left(\frac{1}{n} + \frac{\left(x-\overline{x}\right)^2}{\sum_i \left(x_i-\overline{x}\right)^2} \right)^{1/2}$$

➢ **Banda de confianza para *predecir Y*** en cualquier valor de *X*:

$$\left(\hat{y}\right) \mp (2F_{2,n-2}^{\alpha})^{1/2} \mathrm{s}_r \left(1 + \frac{1}{n} + \frac{\left(x-\overline{x}\right)^2}{\sum_i \left(x_i-\overline{x}\right)^2} \right)^{1/2}$$

El SPSS nos dibuja las correspondientes bandas de confianza.



Riesgos de la extrapolación:

Los límites de confianza calculados mediante las expresiones anteriores son válidos únicamente si el modelo es correcto. Un riesgo evidente de extrapolar el modelo fuera del rango de datos mediante el cual se ha construido, es que la relación entre las variables deje de ser lineal.

8. Los supuestos del modelo de regresión lineal

Hasta ahora explicamos cómo aproximar el modelo de regresión lineal

$$Y = \alpha + \beta X + \varepsilon,$$

por la recta

$$\hat{Y} = a + bX.$$

Para garantizar que esta aproximación es válida, se deben cumplir las siguientes condiciones:

1. **Independencia**: los residuos deben ser independientes entre sí.

2. **Homocedasticidad** (igualdad de varianzas): para cada valor de la variable *X*, la varianza de los residuos $e_i = \left(\hat{Y}_i - Y_i\right)$ debe ser la misma (es decir, que el ajuste es igual de preciso independientemente de los valores que tome *X*).

3. **Normalidad**: para cada valor de la variable *X*, los residuos e_i tienen distribución normal de media cero.

Por lo tanto, para ver si un modelo de regresión lineal ajustado es válido, debemos comprobar que se cumplen estas tres condiciones sobre los residuos.

Ejemplo. En el modelo de regresión ajustado para la *acidez total* sobre la *acidez libre*, debemos comprobar la validez del mismo. Para eso veremos que se cumplen las hipótesis de independencia, homocedasticidad y normalidad dos residuos.

- La independencia podemos comprobarla con el **estadístico de Durbin-Watson**. Si éste está entre 1.5 e 2.5, entonces podemos asumir que los residuos son independientes.

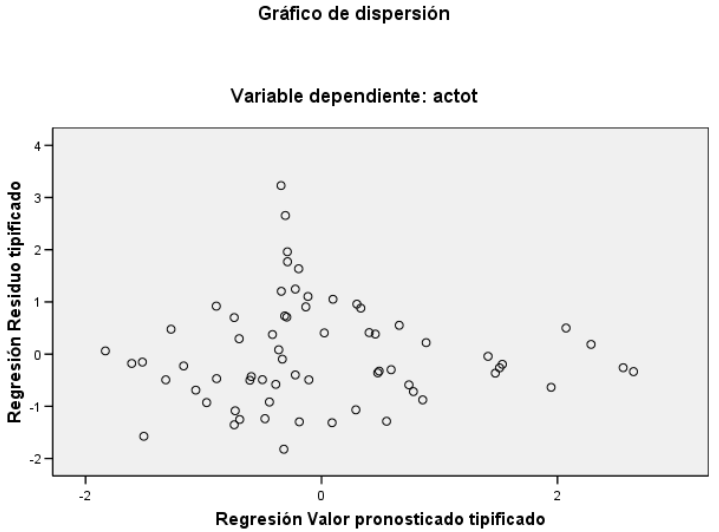
Modelo	R	R^2	R^2corregida	Error típ. de la estimac.	Durbin-Watson
1	.960	.922	.921	2.6762	1.624

1 Variables predictoras: (Constante), aclib

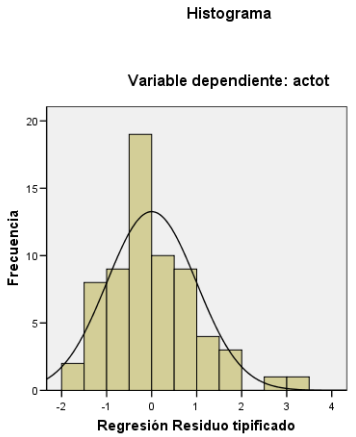
- Para comprobar la **homocedasticidad** de los residuos, haremos un procedimiento gráfico.

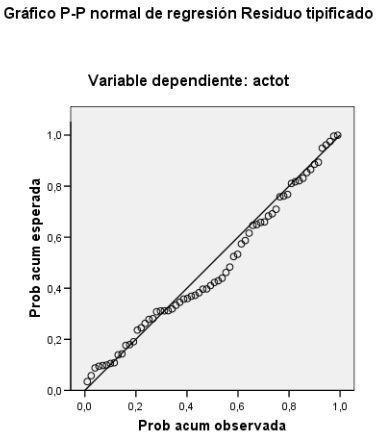
  Dibujaremos un diagrama de dispersión de las estimaciones (valores predichos por el modelo) tipificadas (ZPRED) frente a los residuos tipificados (ZRESID).

  Para garantizar que hay homocedasticidad, no se debe mostrar ninguna pauta de asociación (ningún patrón) en la nube de puntos.



- Para comprobar la **normalidad** hacemos los gráficos de normalidad y realizamos el contraste de normalidad (test de Kolmogorov-Smirnov ó test de Shapiro-Wilks)





Pruebas de normalidad

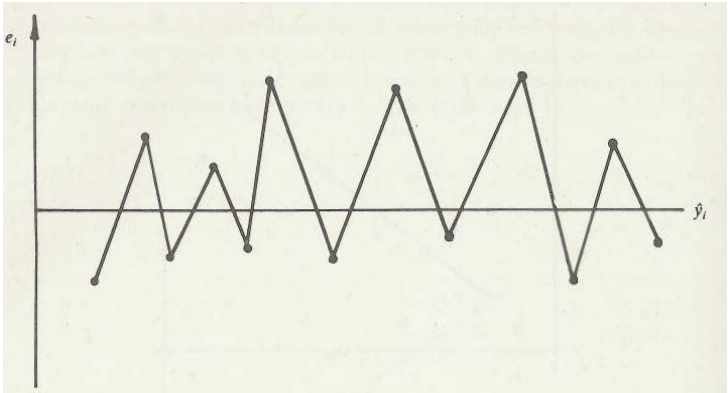
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
Standardized Residual	,107	66	,061	,959	66	,027

a. Corrección de la significación de Lilliefors

Algunos casos en los que no se cumplen los supuestos

i) Falta de independencia

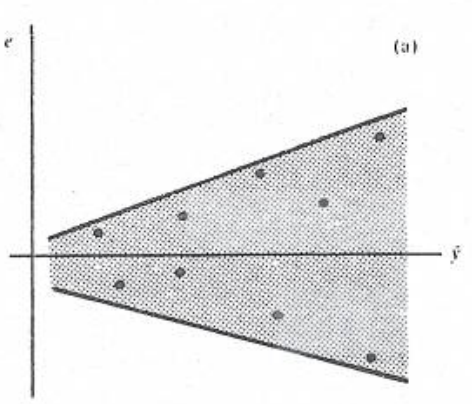
(*Autocorrelación negativa* en los resíduos,
valores por encima de la media tienden a ir seguidos de valores por debajo de ella)



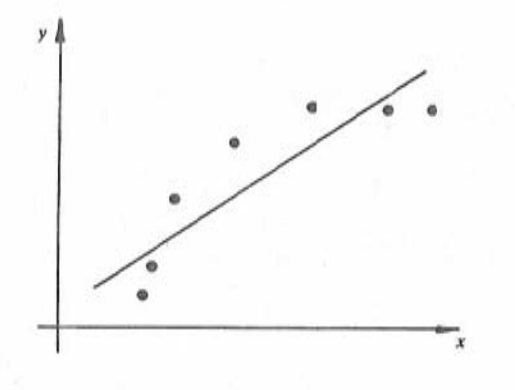
ii) Falta de homocedasticidad

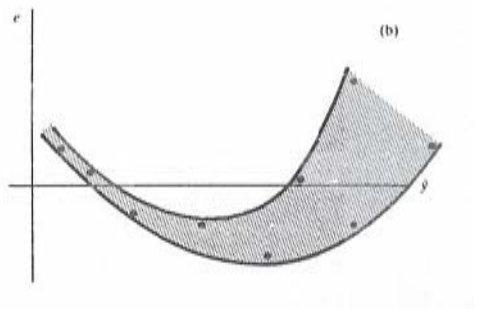
(La variablidad aumenta al aumentar las predicciones)

los valores de las



iii) Falta de linealidad





Transformaciones previas a la regresión:

Si después de efectuar la diagnosis del modelo vemos que no se cumplen algunas de sus hipótesis básicas, podemos actuar de dos maneras:

- Efectuar una transformación de los datos de manera que los datos ya cumplan todas las hipótesis del modelo.

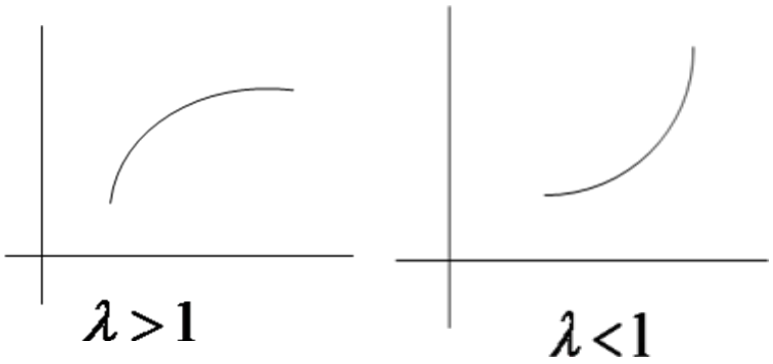
- Buscar otro tipo de modelo de regresión que no requiera las hipótesis que se han imcumplido, que ajuste adecuadamente a los datos y cuyas nuevas hipótesis sí sean verificadas.

Lo más frecuente es intentar primero una transformación de los datos. Las transformaciones más habituales son las de la familia **transformaciones de Box-Cox.**

Esta familia viene dada por la siguiente expresión:

$$h(y) = y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda} & \lambda \neq 0 \\ \ln y & \lambda = 0 \end{cases}$$

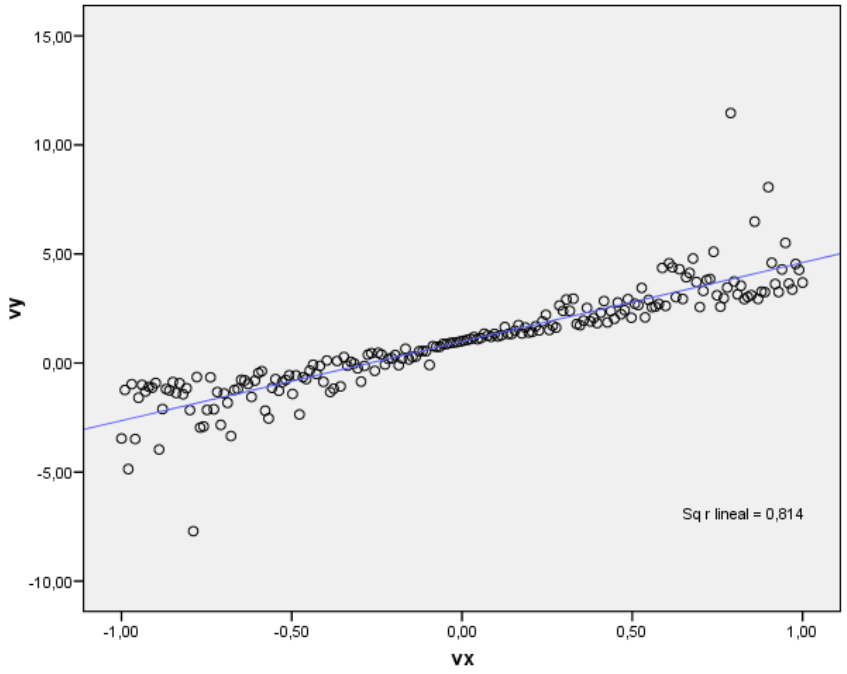
Cada λ produce una transformación diferente, de modo que escogiendo su valor estamos escogiendo la transformación que más nos convenga en cada caso.



9. Un ejemplo en el que no se cumplen los supuestos

Los supuestos que deben cumplirse para que el ajuste de un modelo de regresión lineal sea adecuado son: independencia, igualdad de varianzas y normalidad de los residuos. Pero, ¿qué ocurre cuando estos supuestos no se dan? ¿Cuál es el resultado que obtenemos?

Fijémonos en la gráfica siguiente (archivo *datos_NS.sav*). A simple vista, parecería que el ajuste lineal es adecuado: la línea recta ajusta bastante bien la nube de puntos, y el coeficiente de determinación es de 0.81.



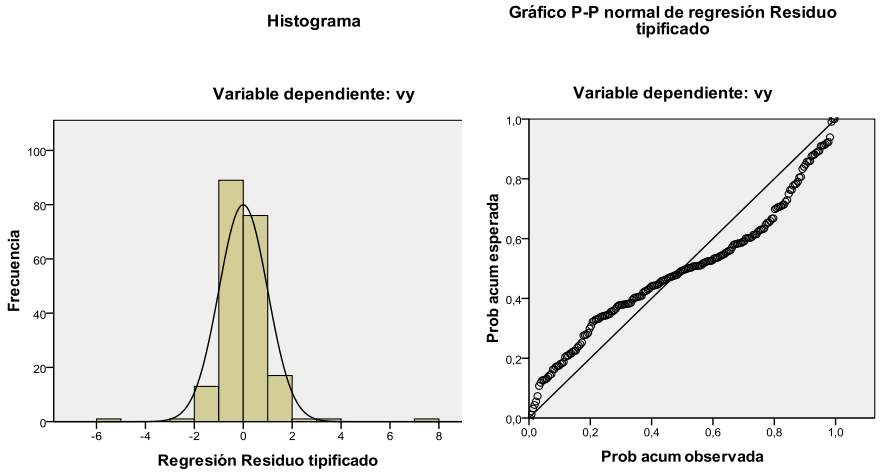
Para un modelo de regresión lineal sobre estos datos, los resultados son los siguientes:

Resumen del modelo(b)

Mod.	R	R^2	R^2 corregida	Error típ. de estimac.	Durbin-Watson
1	,902	,814	,813	1,00916	2,096

- **Los datos son incorrelados**

Veamos la normalidad:



Pruebas de normalidad

	Kolmogorov-Smirnov			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
Standardized Residual	,136	200	,000	,776	200	,000

a. Corrección de la significación de Lilliefors

- **Vemos que los datos no son normales.**

En cuanto a la igualdad de varianzas:

- Hay **heterocedasticidad** pues la varianza es mayor para valores pronosticados grandes o pequeños.