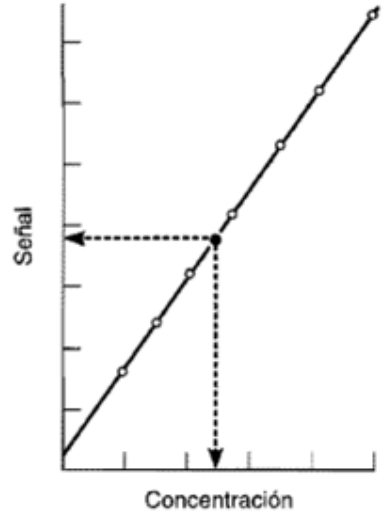
RELACIÓN ENTRE DOS VARIABLES NUMÉRICAS. REGRESIÓN LINEAL SIMPLE. CORRELACIÓN.

Los métodos de regresión se usan para estudiar la relación entre dos variables numéricas. Este tipo de problemas aparecen con frecuencia en el contexto de química analítica cuando se desea realizar el calibrado en análisis instrumental.

El procedimiento habitual es el siguiente: el analista toma una serie de materiales (pueden ser 3 ó 4 ó más aún) en los que conoce la concentración del analito. Estos patrones de calibración se miden con el instrumento analítico en las mismas condiciones en las que se trabajará en los ensayos con el material desconocido. Una vez establecido el gráfico de calibrado puede obtenerse la concentración del analito como se muestra en el siguiente gráfico:



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

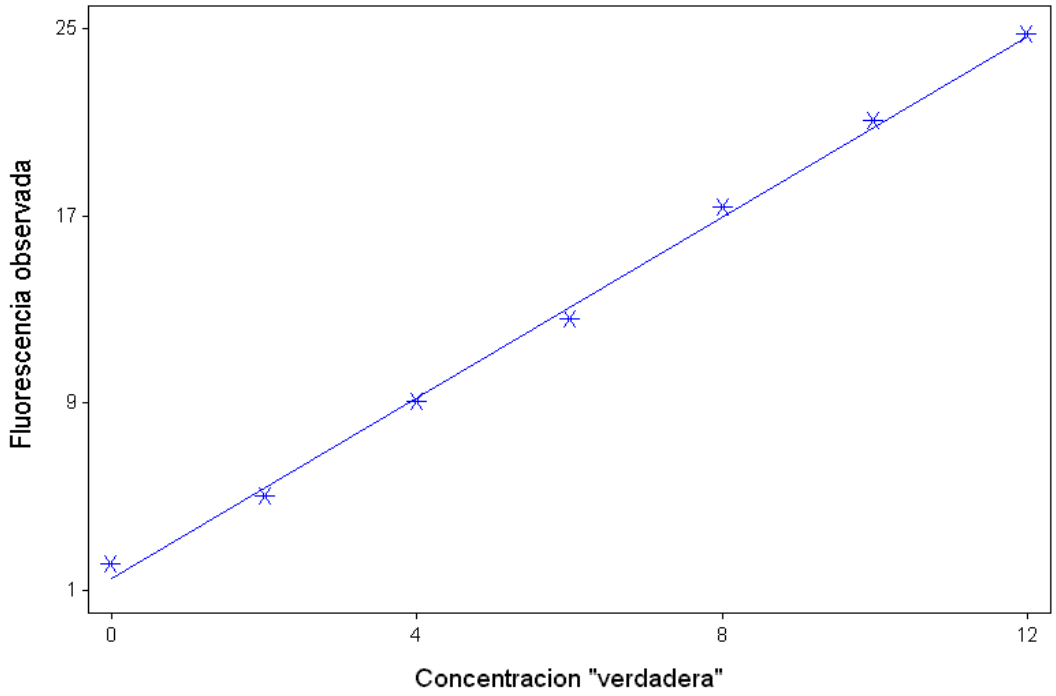
Veamos un ejemplo numérico:

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

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

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

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



Para ayudar a visualizar la relación, hemos agregado a los puntos del gráfico de dispersión una recta que se llama "recta de regresión" o "recta de cuadrados mínimos". Para ello, basta marcar (en el Statistix) donde dice "Display Regression Line".

Recordemos que la ecuación de una recta es de la forma

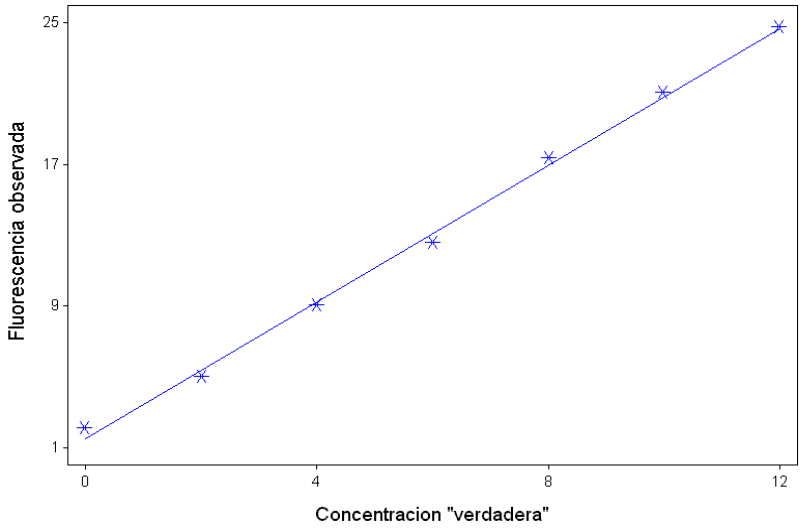
$$y = \alpha + \beta\ x$$

Ordenada al origen **pendiente**

Recta de cuadrados mínimos.

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

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



UNWEIGHTED LEAST SQUARES LINEAR REGRESSION OF FLUORESCE

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

R-SQUARED	0.9978	RESID. MEAN SQUARE (MSE)	0.18736
ADJUSTED R-SQUARED	0.9973	STANDARD DEVIATION	0.43285

SOURCE	DF	SS	MS	F	P
REGRESSION	1	417.343	417.343	2227.53	0.0000
RESIDUAL	5	0.93679	0.18736		
TOTAL	6	418.280			

CASES INCLUDED 7 MISSING CASES 0

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

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

o

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

Por ejemplo, si la concentración de fluoresceína de una muestra es 8, la ordenada de la recta es

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

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

Modelo de regresión lineal

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

La palabra "simple" es porque consideramos una sola variable independiente o predictora (X). Se generaliza en forma natural al caso en que hay más variables independientes y en ese caso se llama "modelo de regresión lineal múltiple".

Las suposiciones del modelo de regresión lineal simple son las siguientes.

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

$$y_i = \alpha + \beta_i x + e_i \qquad (2)$$

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

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

Para obtener algunos resultados alcanzan las suposiciones 1) a 3), pero para otros es necesario agregar:

4) $e_i \sim$ Normal

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

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

Observación:

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

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

En otras situaciones como podría ser en un problema en el que se desea estudiar la relación entre estatura y perímetro cefálico de recién nacidos, las covariables x_i's son aleatorias. Justificaremos los resultados sobre estimadores, IC e tests sólo para el modelo con x_i's fijas, que es más simple, pero casi todos estos resultados son los mismos para ambos modelos.

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

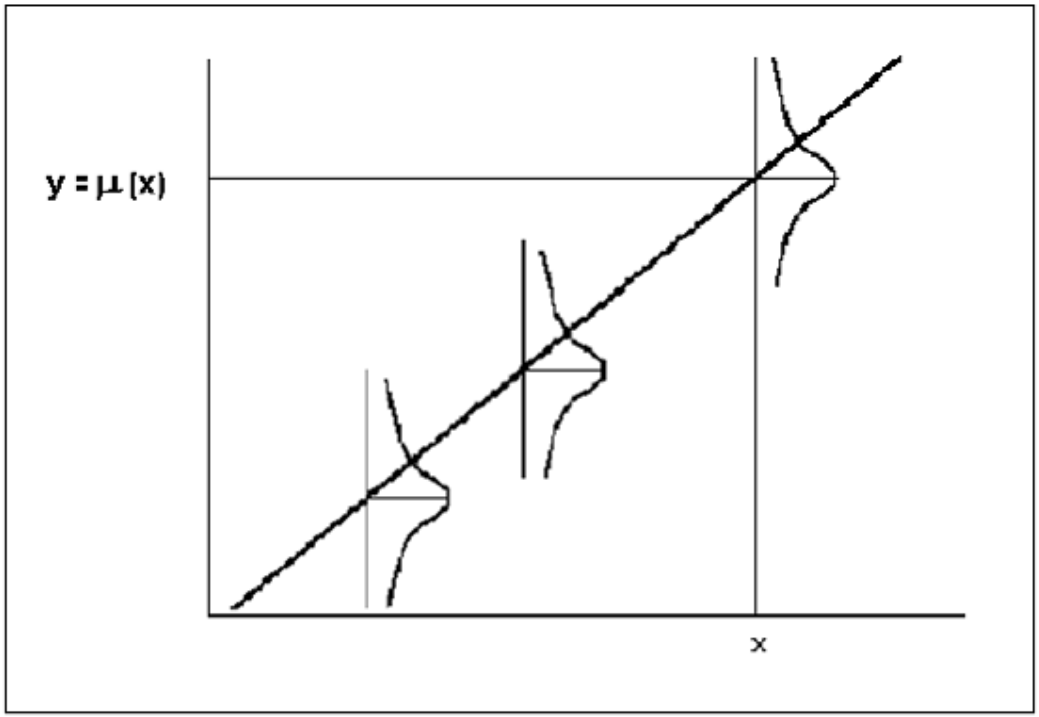
1*) $E(y_i) = \alpha + \beta x_i$ (para i=1,...,n)
2*) $Var(y_i) = \sigma^2$ (para i=1,...,n)
3*) $y_1, y_2, ..., y_n$ son v. a. independientes
4*) $y_i \sim$ Normal

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

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

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

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



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

Estimadores de α y β por el método de cuadrados mínimos

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

¿Cómo hallamos a y b ?

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

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

Mostraremos resolviendo estas ecuaciones que

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

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

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

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

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

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

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

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

Valores predichos: Llamamos valores predichos a

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

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

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

es un estimador insesgado de σ^2.

Varianza de $\hat{\alpha}$ y $\hat{\beta}$ **:** Se puede demostrar que:

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

y además

$$\text{COV}(\bar{y}, \hat{\beta}) = 0 \qquad (7)$$

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

Intervalo de confianza para β

Llamemos

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

El intervalo

$$\hat{\beta} \pm t_{n-2;\alpha/2}\, ES(\hat{\beta}) \qquad (8)$$

es un IC para β con nivel 1-α. Si la suposición 4) (de normalidad) no se cumple, este intervalo, bajo condiciones muy generales, tiene nivel asintótico 1-α.

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

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

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

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

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

Puede demostrarse que

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

Se cumple que

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

Significado del valor de r^2

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

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

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

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

Siempre es

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

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

En el ejemplo de la fluorescencia, se ve en la salida del Statistix que

R-SQUARED 0.9978

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

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

Estimación del valor esperado de y para un valor fijado de x y su intervalo de confianza.

Si fijamos un valor de la variable independiente, digamos en x_0: **¿cuál es el valor esperado de y para ese valor de la variable independiente?**

Por el modelo supuesto, por la suposición 1) o 1*) el valor esperado de y es

$$E(y)= \alpha + \beta\, x_0$$

Su estimador es

$$\hat{\alpha} + \hat{\beta}\; x_0$$

Usando (6) y (7) se puede demostrar que la varianza de este estimador es:

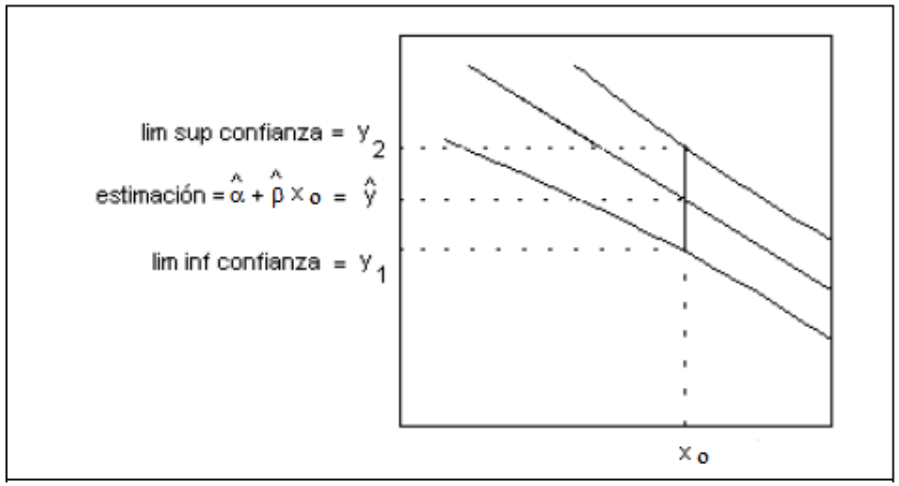
$$\mathrm{Var}(\hat{\alpha} + \hat{\beta}\, x_0) \quad = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] \qquad (11)$$

y que el intervalo de extremos

$$\left[\hat{\alpha} + \hat{\beta}\, x_0 - t_{n-2;\alpha/2} \sqrt{s^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]} \quad ; \quad \hat{\alpha} + \hat{\beta}\, x_0 + t_{n-2;\alpha/2} \sqrt{s^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]} \right] \qquad (12)$$

es un IC con nivel 1-α para el valor esperado de y, para $x = x_0$.

Gráficamente quedaría así:



Predicción de un nuevo valor de Y conocido el valor de x e intervalo de predicción.

Los estimadores de los parámetros del modelo se basaron en una muestra de n observaciones (x_i, y_i) (i=1,...,n).

Supongamos ahora que hacemos una nueva observación, pero sólo conocemos su valor de x (llamémoslo x_{n+1}), no conocemos el valor correspondiente de y, que llamaremos y_{n+1}.

Queremos dar un valor aproximado para y_{n+1} (se dice que queremos "predecir" y_{n+1}) y un intervalo que contiene a y_{n+1} con una probabilidad 0.95 (o 1-α) (que se llama intervalo de predicción para y_{n+1}).

Supondremos que el nuevo individuo observado cumple el mismo modelo que los n anteriores. Entonces:

$$y_{n+1} = \alpha + \beta \; x_{n+1} + e_{n+1}$$

donde e_{n+1} es una v.a. con esperanza cero y es independiente de e_1, e_2, ..., e_n.

Es intuitivamente razonable que el mejor predictor de y_0 sea:

$$\hat{y}_{n+1} = \hat{\alpha} + \hat{\beta} \; x_{n+1} \qquad (13)$$

El error de predicción es:

$$y_{n+1} - \hat{y}_{n+1} = (\alpha + \beta\, x_{n+1}) + e_{n+1} - (\,\hat{\alpha} + \hat{\beta}\;\; x_{n+1})$$

Se puede demostrar que este error de predicción tiene esperanza cero y varianza

$$Var(y_{n+1} - \hat{y}_{n+1}) = Var(e_{n+1}) + Var(\hat{\alpha} + \hat{\beta}\, x_{n+1}) \quad = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]$$

y que el intervalo de extremos

$$\left[\hat{y}_{n+1} - t_{n-2;\alpha/2} \sqrt{s^2 \left[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]};\quad \hat{y}_{n+1} + t_{n-2;\alpha/2} \sqrt{s^2 \left[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]} \right] \qquad (14)$$

es un "**intervalo de predicción**" con nivel 1-α para una nueva observación y_{n+1}.

Aplicación a un ejemplo: Volvamos al ejemplo de la fluorescencia. De la salida del programa mostrada anteriormente obtenemos:

$$\hat{\alpha} = 1.51786 \quad ; \qquad \hat{\beta} = 1.93036 \quad ; \qquad s^2 = 0.18736$$

$$ES(\hat{\beta}) = \sqrt{\hat{Var}(\hat{\beta})} = 0.04090$$

No aparece directamente en la salida el IC para β, pero es fácil obtenerlo usando (8).

Si queremos un IC al 95%, necesitamos el valor de t con 7-2=5 gl, con p=0.05 en las dos colas. Buscando en SX o en tablas obtenemos: $t_{5;\ 0.025} = 2.57$ y, reemplazando en (8):

$$1.93036 \pm 2.57 * 0.04090$$

$$1.93036 \pm 0.10511$$

o, redondeando

IC para β con nivel 95%: [1.83, 2.04]

El IC al 95% para α se obtiene en forma análoga:

$$1.51786 \pm 2.57*0.29494$$

redondeando:

$$1.52 \pm 0.76$$

IC para α con nivel 95%: [0.76, 2.59]

Predicción: Vamos a calcular ahora el predictor de la medición de fluorescencia y un intervalo de predicción para una nueva observación cuya concentración de fluoresceína es 8 pci/ml.

El predictor es fácil de calcular:

$$\hat{y}_{n+1} = \hat{\alpha} + \hat{\beta}\, x_{n+1} \quad = 1.51786 + 1.93036*8 = 16.96$$

Para obtener el intervalo de predicción para y_{n+1} hay que usar la expresión (14).

Statistix calcula automáticamente dicho intervalo. Para ello, inmediatamente después de obtener la salida de la regresión, marcamos "Results", "Prediction", ponemos en la ventana "Predictor Values" el número 8 y obtenemos:

PREDICTED/FITTED VALUES OF FLUORESCE

LOWER **PREDICTED** BOUND	15.753	LOWER FITTED BOUND	16.491
PREDICTED VALUE	16.961	FITTED VALUE	16.961
UPPER **PREDICTED** BOUND	18.169	UPPER FITTED BOUND	17.431
SE (PREDICTED VALUE)	0.4699	SE (FITTED VALUE)	0.1829

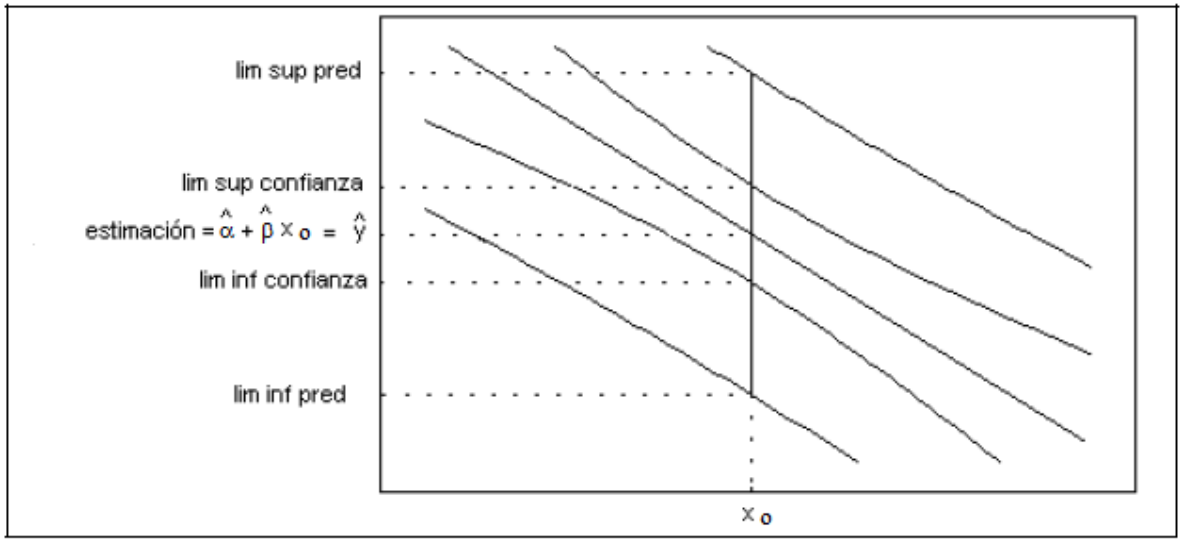
UNUSUALNESS (LEVERAGE)	0.1786
PERCENT COVERAGE	95.0
CORRESPONDING T	2.57

PREDICTOR VALUES: CONCENTRA = 8.0000

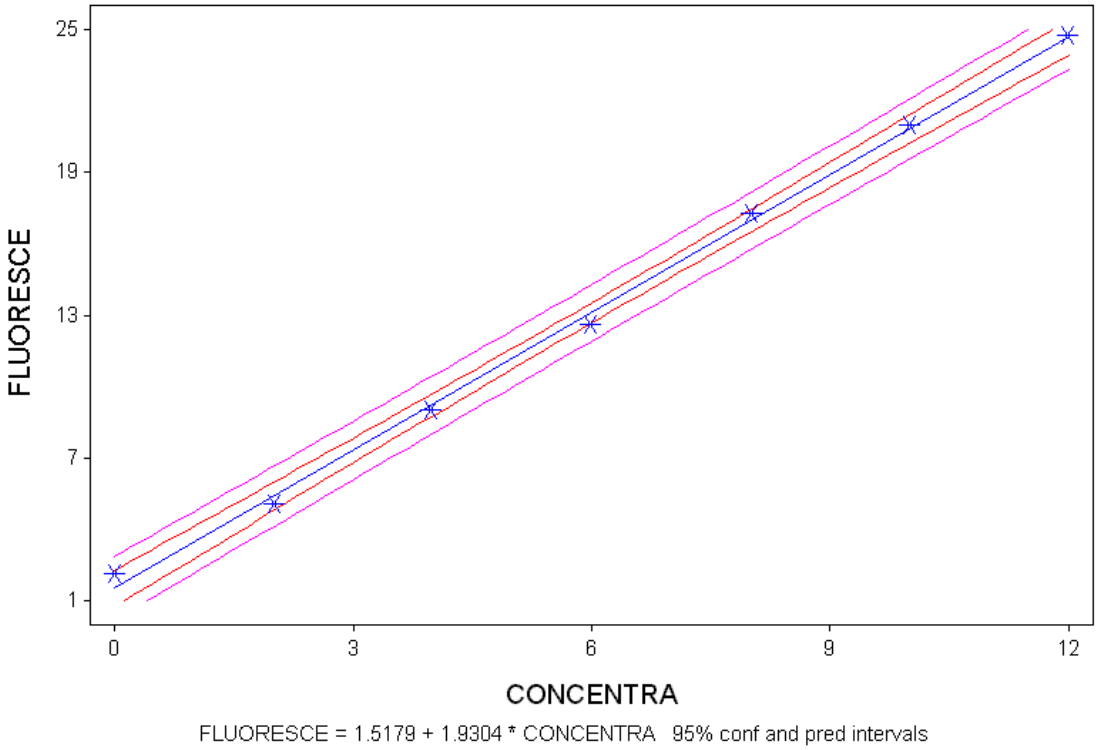
Vemos que el predictor o valor predicho es 16.961 y el intervalo de predicción al 95% es [15.753 ; 18.169].
También se muestra en esta salida el IC al 95% para el valor esperado de la medición de fluorescencia para muestras con concentración de fluoresceína=8. Observemos que este último intervalo tiene menor longitud que el de predicción.

Pregunta: *¿Es intuitivamente razonable que el IC para el valor esperado tenga menor longitud?*

Gráficamente los dos intervalos quedarían así:



Con SX también podemos representar gráficamente los intervalos de prediccion y los IC para el valor esperado de y, para diferentes valores de x. Para ello, siempre a partir de la salida de la regresión lineal, vamos a "Results", "Plots", "Simple Regression Plot" y obtenemos:

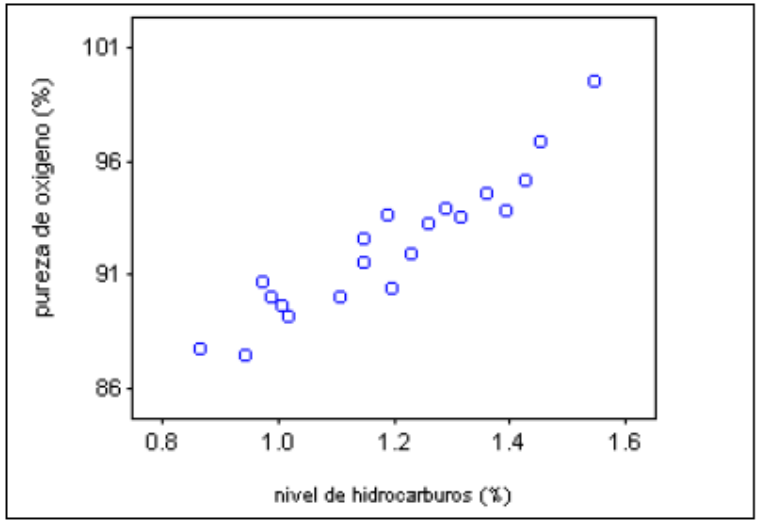


Observación**:** aunque no se nota mucho en este gráfico ni el IC para el valor esperado de y ni el intervalo de predicción tienen longitud constante, ¿para que valores de x_0 o x_{n+1} es menor la longitud?

Aquí mostramos los resultados en otro ejemplo:

Interesa estudiar la relación entre la pureza de oxígeno (y) producido en un proceso de destilación y el porcentaje de hidrocarburos (x) presentes en el condensador principal de un destilador. Los datos se muestran en la tabla y scatter plot siguientes:

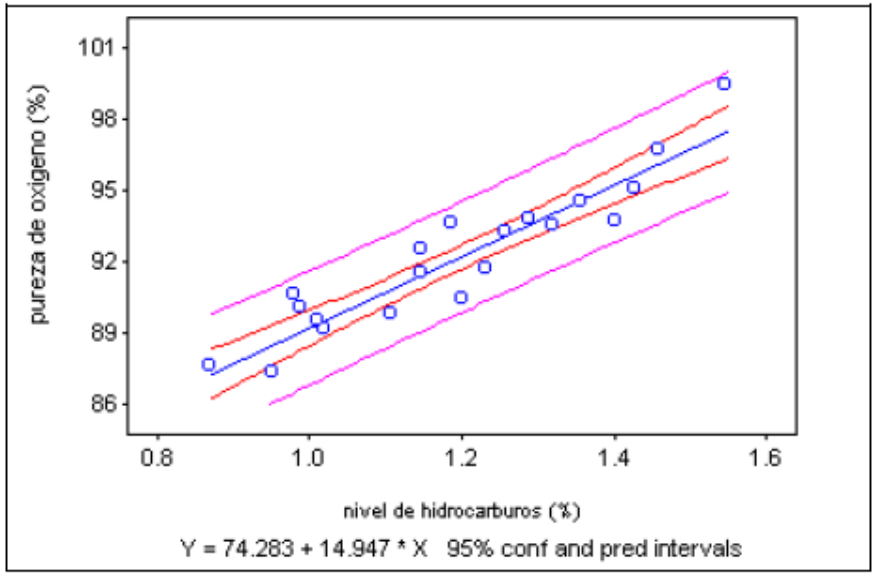
x(%)	y(%)	x(%)	y(%)	x(%)	y(%)	x(%)	y(%)
0.99	90.01	1.36	94.45	1.19	93.54	1.2	90.39
1.02	89.05	0.87	87.59	1.15	92.52	1.26	93.25
1.15	91.43	1.23	91.77	0.98	90.56	1.32	93.41
1.29	93.74	1.55	99.42	1.01	89.54	1.43	94.98
1.46	96.73	1.4	93.65	1.11	89.85	0.95	87.33



PREDICTOR VARIABLES	COEFFICIENT	STD ERROR	STUDENT'S T	P
CONSTANT	74.2833	1.59347	46.62	0.0000
X	14.9475	1.31676	11.35	0.0000

R-SQUARED	0.8774	RESID. MEAN SQUARE (MSE)	1.18055
ADJUSTED R-SQUARED	0.8706	STANDARD DEVIATION	1.08653

Recta ajustada junto con las
bandas de confianza y de predicción del 95%



Predicción inversa: predicción de de un nuevo valor de x conocido el valor de y cálculo de un intervalo de confianza.

Los estimadores de los parámetros del modelo se basaron en una muestra de n observaciones (x_i, y_i) (i=1,...,n).

Supongamos ahora que hacemos una nueva observación, pero sólo conocemos su valor de y, no conocemos su valor x. Queremos calcular un “estimador” de x y un intervalo que contiene a x con una probabilidad $1-\alpha$.

Hemos dicho que hay dos modelos de regresión lineal simple: uno con x's fijas y otro con x's aleatorias. Pero en ambos modelos y es aleatoria.

- En el caso en el que la variable x también es aleatoria, si queremos predecir X conocido Y una solución es cambiar el modelo: intercambiar en (2) el papel de las variables “Y” y “X” y luego aplicar "predicción" (o sea (13) y (14)).

- Pero si la variable x es fija (fijada por el experimentador), como suele ocurrir en los experimentos de calibración, no se la puede considerar como variable de respuesta "y" en (2), ya que no se cumplirían las suposiciones del modelo de regresión.

Consideremos entonces el caso x fija.

Supondremos que el nuevo individuo observado cumple el mismo modelo que los n anteriores, luego

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

donde e es una v.a. con esperanza cero y es independiente de e_1, e_2, ..., e_n.

Despejando x

$$x = \frac{y - \alpha - e}{\beta}$$

Como no tenemos información ninguna sobre e y, además, de α y β sólo conocemos los estimadores, es intuitivamente razonable estimar x con:

$$\hat{x} = \frac{y - \hat{\alpha}}{\hat{\beta}} \qquad (15)$$

Como $\hat{x}$ es un cociente de variables aleatorias, no es fácil calcular su varianza, pero se puede encontrar una expresión **aproximada**.

El estimador de esta aproximación de la varianza es

$$\hat{\mathrm{V}}\mathrm{ar}(\hat{x}) = \frac{s^2}{\hat{\beta}^2}\left[1 + \frac{1}{n} + \frac{(Y - \overline{Y})^2}{\hat{\beta}^2 \sum_{i=1}^{n}(x_i - \overline{x})^2}\right] \quad (16)$$

Llamando

$$ES(\hat{x}) = \sqrt{\hat{\mathrm{V}}\mathrm{ar}(\hat{x})} \quad (17)$$

el intervalo

$$\hat{x} \pm t_{n-2;\alpha/2}\, ES(\hat{x}) \quad (18)$$

es un intervalo de confianza con nivel aproximado 1-α para x.

Supongamos ahora que, para obtener mayor precisión, un químico hace "m" mediciones para la misma muestra. La muestra tiene un valor x desconocido y llamamos $\overline{Y}_m$ al promedio de las m observaciones Y's hechas en esa muestra. Entonces (46) y (47) se modifican así:

$$\hat{x} = \frac{\overline{y}_m - \hat{\alpha}}{\hat{\beta}} \qquad (15^*)$$

$$\hat{Var}(\hat{x}) = \frac{s^2}{\hat{\beta}^2} \left[\frac{1}{m} + \frac{1}{n} + \frac{(\overline{y}_m - \overline{y})^2}{\hat{\beta}^2 \sum_{i=1}^{n} (x_i - \overline{x})^2} \right] \qquad (16^*)$$

Quedando (17) y (18) sin cambios.

Ejemplo: SX no calcula la predicción inversa, que es el objetivo principal de un experimento de calibración. Podemos hacer las cuentas "a mano". Continuamos con el ejemplo de la fluorescencia.

Ahora medimos una muestra de la que no conocemos la concentración de fluoresceína. La medición de fluorescencia es 13.5. ¿Cuál es la verdadera concentración de fluoresceína de la muestra?

Llamemos x a esta verdadera concentración desconocida. Su estimador se calcula con (15):

$$\hat{x} = \frac{y - \hat{\alpha}}{\hat{\beta}} = \frac{13.5 - 1.518}{1.930} = 6.21$$

El estimador de la concentración es 6.21 pg/ml.

Una medida de la precisión de esta estimación la dan su Error Standar y también el IC al 95%. Necesitamos primero calcular (16). Vemos que todo lo que se necesita para calcular (16) puede encontrarse en la salida de la regresión lineal, salvo $\bar{y}$ y $\sum (x_i - \bar{x})^2$. En este experimento en que hay n=7 pares de datos, se podrían hacer las cuentas con una calculadora. Otra forma puede ser calcular en Summary Statistics, Descriptive Statistics:

DESCRIPTIVE STATISTICS

VARIABLE	N	MEAN	SD	VARIANCE
CONCENTRA	7	6.0000	4.3205	**18.667**
FLUORESCE	7	**13.100**	8.3495	69.713

Luego $\bar{y}$ = 13.10

$\sum (x_i - \bar{x})^2$ no lo tenemos directamente, pero tenemos la varianza que es igual a $\sum (x_i - \bar{x})^2 / (n-1)$. Por lo tanto multiplicando la varianza por (n-1) obtenemos

$$\sum (x_i - \bar{x})^2 = \mathbf{18.667*6 = 112.0}$$

Reemplazamos ahora en (16):

$$\hat{Var}(\hat{x}) = \frac{0.18736}{1.93036^2}\left[1 + \frac{1}{7} + \frac{(13.5-13.10)^2}{1.93036^2 * 112.0}\right] = 0.05748$$

Luego

$$ES(\hat{x}) = \sqrt{0.05748} = 0.240$$

Aplicando (18) obtenemos que

$$6.21 \pm 2.57*0.240$$
$$6.21 \pm 0.62$$

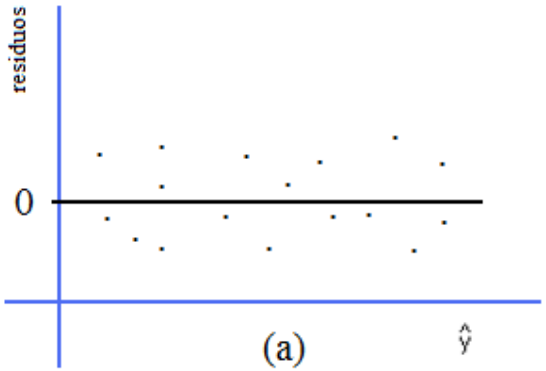
son los límites de confianza al 95% para la concentración de fluoresceína en la nueva muestra observada.

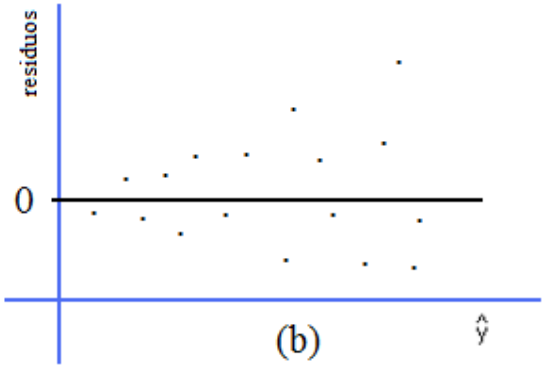
¿Como se debería tomar la muestra en el experimento de calibración para disminuir la longitud de los intervalos de confianza para x?

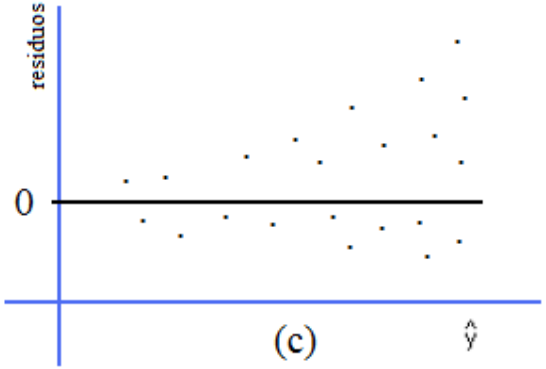
Diagnóstico del modelo de regresión.

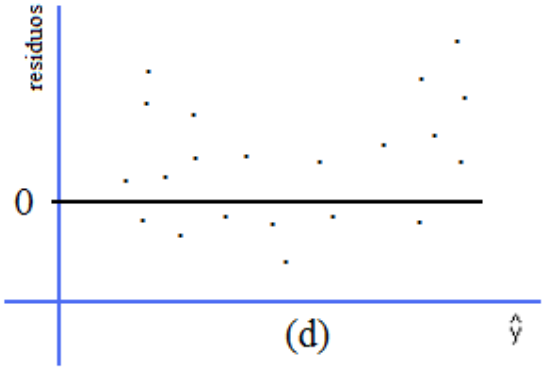
En regresión simple la validación de los supuestos del modelo se realiza en base a los datos y a los residuos del modelo ajustado. El diagrama de dispersión permite tener una idea del supuesto de linealidad y de la condición de homocedasticidad. Se realizan diversos gráficos: de los valores predichos vs. los residuos, que no debería mostrar ninguna estructura particular, y de la covariable vs. los residuos para evaluar el ajuste y también boxplots y qq-plots de los residuos para evaluar la normalidad de los errores.

Los siguientes gráficos muestran algunas situaciones que podemos encontrar.







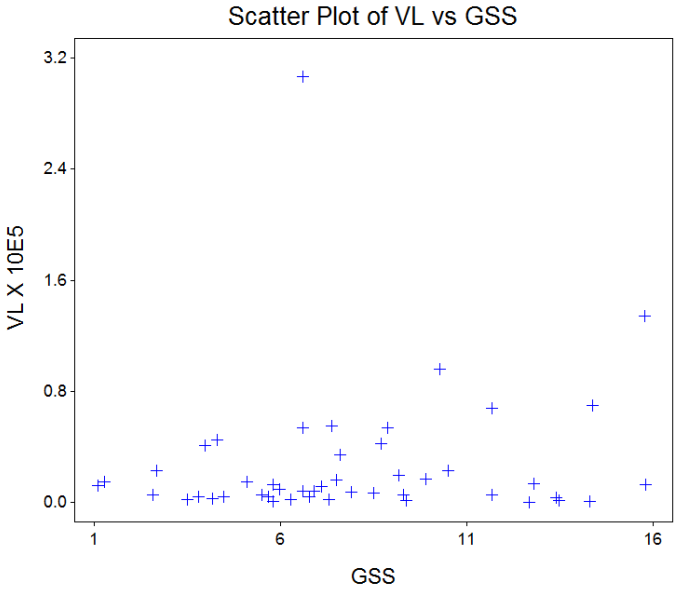


(a)
Representa la situación esperable si el modelo se cumple: una nube de residuos alrededor del 0 sin estructura.

(b) y (c)
Muestran gráficos en los que el supuesto de igualad de varianzas no se cumple.

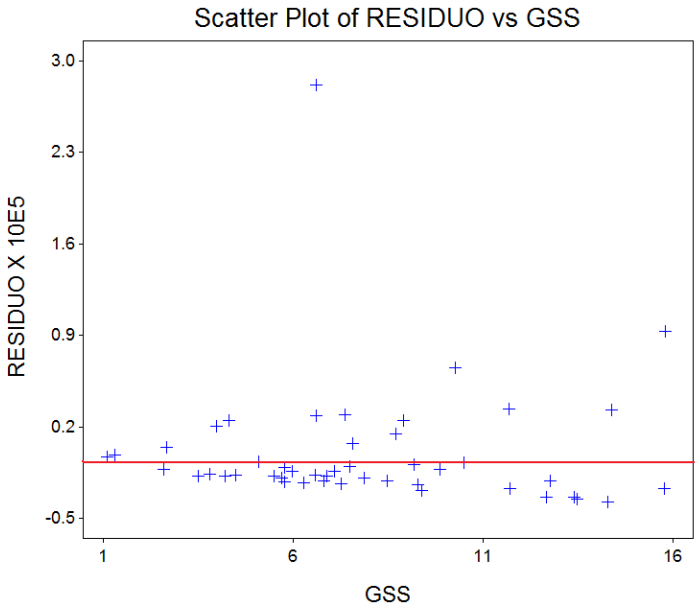
(d) El supuesto de linealidad no se satisface.

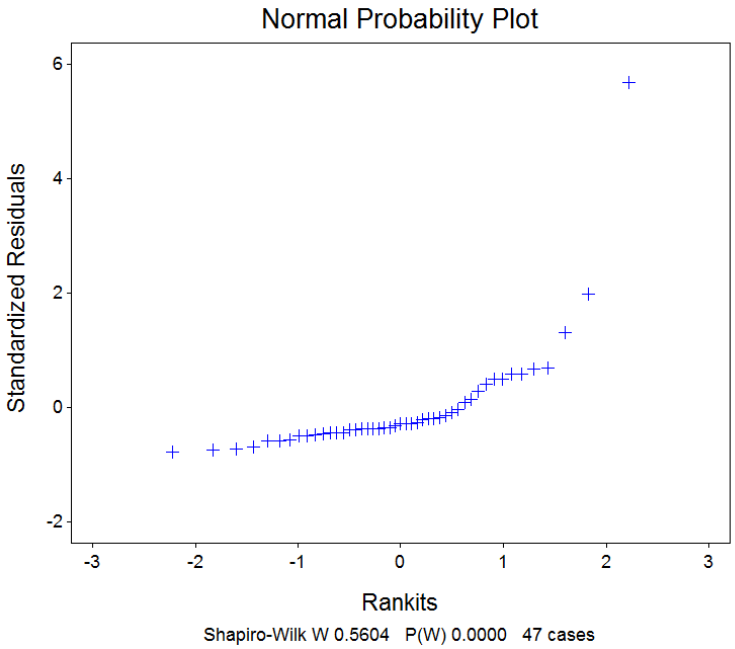
Ejemplo 1: Consideramos los datos correspondientes a 47 pacientes de HIV para los que se registraron las variables VL: carga viral y GSS que corresponde a un score genético del paciente.



En primer término realizamos un scatter plot de los datos originales y detectamos la presencia de un outlier así como cierto efecto de “abanico”.

En segundo término graficamos los residuos obtenidos después del ajuste de un modelo lineal usando como covariable GSS y respuesta VL. Aquí el efecto de “abanico” se encuentra reforzado.





El QQ-plot muestra un importante apartamiento de la normalidad y el test de Shapiro-Wilk tiene un p-valor inferior a 0.0001

Como vemos el SX realiza este gráfico usando “residuos standarizados” en lugar de los residuos $\hat{e}_i$ que hemos definido. ¿Cómo se definen?

En realidad los residuos no son igualmente distribuidos, se puede probar que

$$V(\hat{e}_i) = \sigma^2(1-h_{ii})$$

donde

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

mide la distancia de la i-ésima observación al promedio muestral.

Los h_{ii} reciben el nombre de **palanca o leverage** de la observación i-ésima.

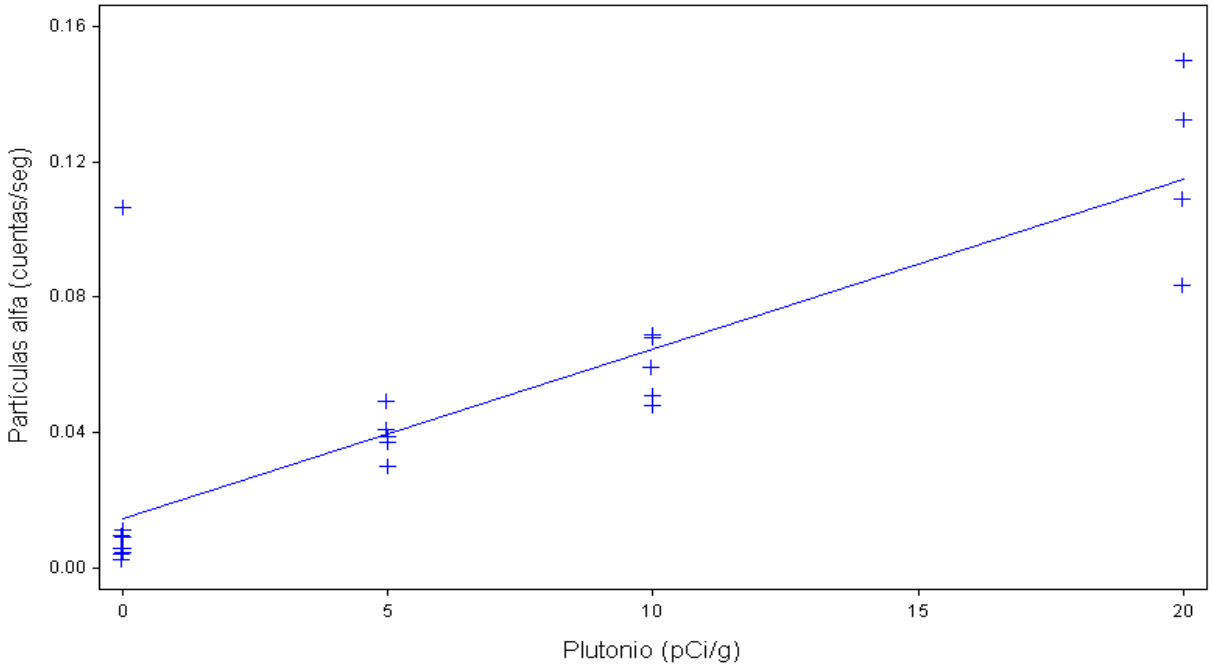
Teniendo en cuenta la varianza de los residuos definimos los residuos standarizados como:

$$r_i = \frac{y_i - \hat{y}_i}{s(1-h_{ii})^{1/2}}$$

Ejemplo 2: Cuando el plutonio se encuentra en pequeñas cantidades una forma de detectarlo es mediante las partículas alfa que emite. En un experimento de calibración se midieron varias veces 4 materiales standards para los que se conoce la actividad de plutonio (0, 5, 10 y 20 picocuries por gramo (pCi/g). Los resultados de estas mediciones se muestran a continuación y en el siguiente gráfico se puede apreciar la relación entre las dos variables.

0	5	10	20
0,004	0,030	0,069	0,150
0,011	0,041	0,068	0,109
0,004	0,037	0,048	0,083
0,009	0,039	0,059	0,132
0,009	0,049	0,051	
0,006			
0,004			
0,006			
0,002			
0,106			

Observemos el diagrama de puntos ("Statistics","Summary Statistics", "Scatter Plot"):



En este diagrama se observa que los datos no siguen el modelo de regresión lineal habitual: hay un claro dato atípico y no parece cumplirse la suposición de varianza constante.

Una posible forma de detectar fallas en el modelo, es estimar los parámetros del modelo y luego hacer gráficos para el “diagnóstico”. Para esto vamos a "Statistics", "Linear Models", "Linear Regression" y obtenemos:

UNWEIGHTED LEAST SQUARES LINEAR REGRESSION OF PARTALFA

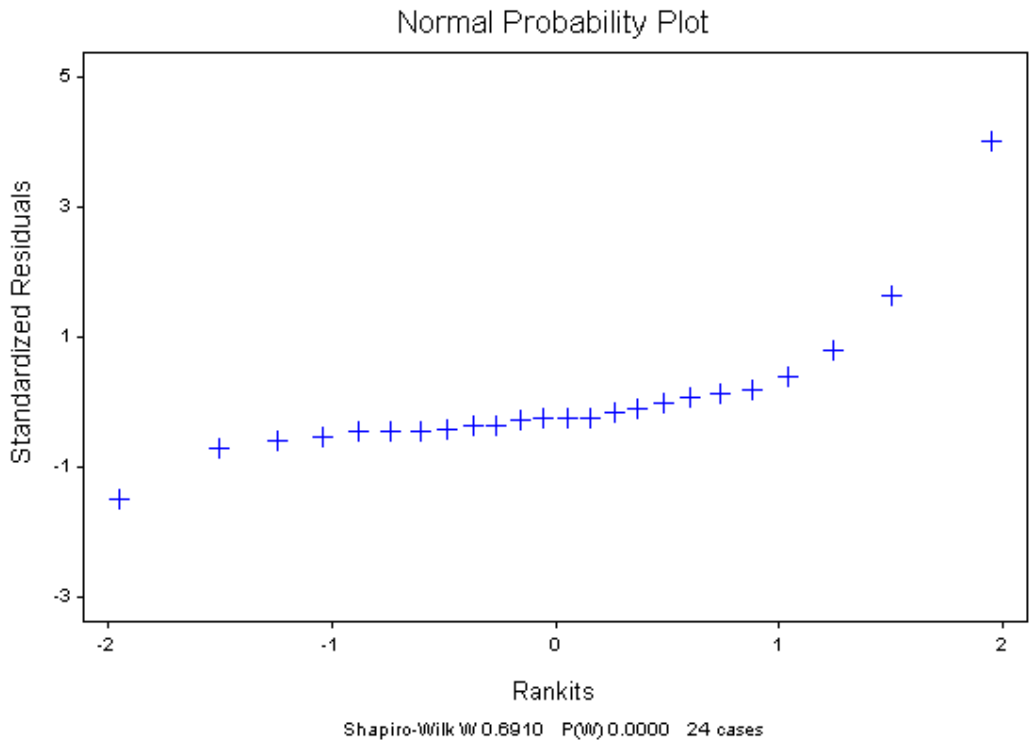
PREDICTOR VARIABLES	COEFFICIENT	STD ERROR	STUDENT'S T	P
CONSTANT	0.01453	0.00653	2.23	0.0366
PLUTONIO	0.00501	6.778E-04	7.40	0.0000

R-SQUARED	0.7133	RESID. MEAN SQUARE (MSE)	5.623E-04
ADJUSTED R-SQUARED	0.7003	STANDARD DEVIATION	0.02371

SOURCE	DF	SS	MS	F	P
REGRESSION	1	0.03078	0.03078	54.74	0.0000
RESIDUAL	22	0.01237	5.623E-04		
TOTAL	23	0.04315			

CASES INCLUDED 24 MISSING CASES 0

Podemos elegir del menú las opciones: "Results", "Plots", "Normal Probability Plot" y el programa hace el siguiente gráfico que permite evaluar la normalidad de los residuos:



En el gráfico se observa la presencia de un valor atípico y el test de Shapiro Wilk rechaza la hipótesis de normalidad (P<0.0001).
Si excluímos el dato atípico y volvemos a estimar los parámetros de la regresión y hacer gráficos con los residuos, resulta:

UNWEIGHTED LEAST SQUARES LINEAR REGRESSION OF PARTALFA

PREDICTOR VARIABLES	COEFFICIENT	STD ERROR	STUDENT'S T	P
CONSTANT	**0.00703**	**0.00360**	**1.95**	**0.0641**
PLUTONIO	0.00554	3.659E-04	15.13	0.0000

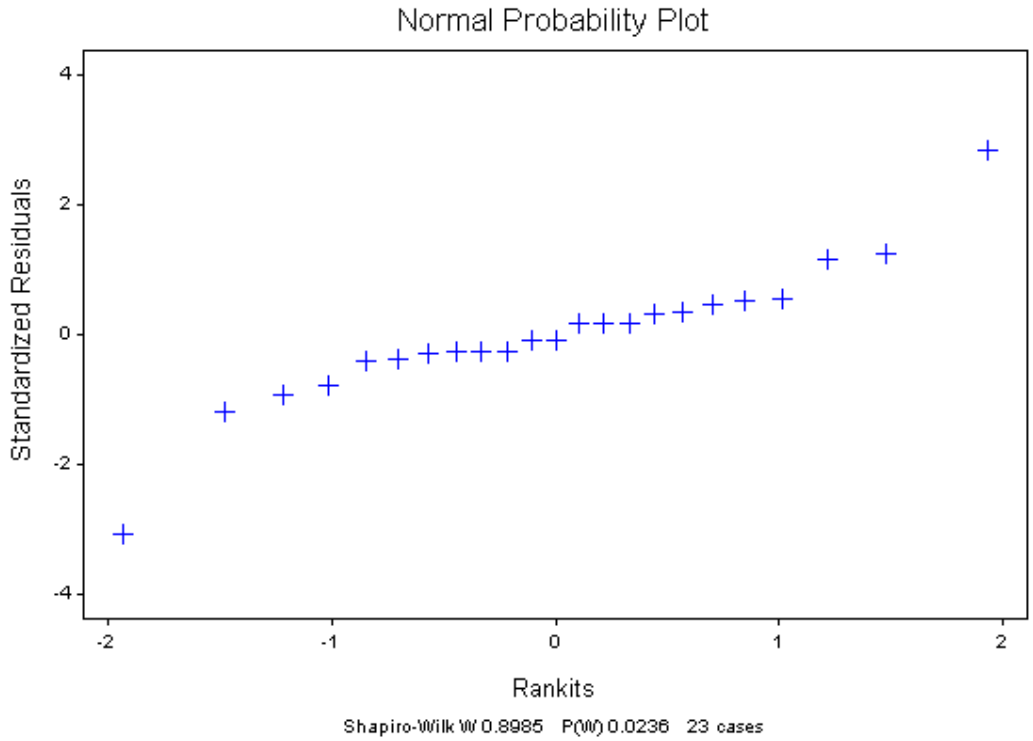
R-SQUARED	0.9160	RESID. MEAN SQUARE (MSE)	1.580E-04
ADJUSTED R-SQUARED	0.9120	STANDARD DEVIATION	0.01257

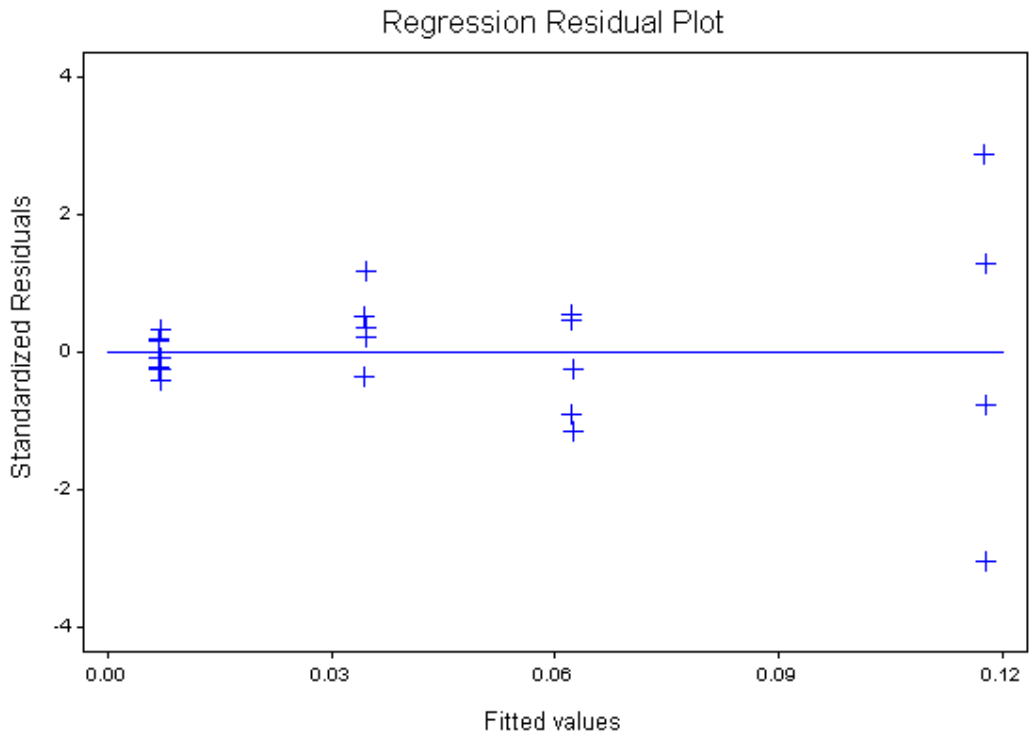
SOURCE	DF	SS	MS	F	P
REGRESSION	1	0.03619	0.03619	229.00	0.0000
RESIDUAL	21	0.00332	1.580E-04		
TOTAL	22	0.03951			

CASES INCLUDED 23 MISSING CASES 0

Notemos las diferencias con respecto al análisis anterior en la estimación de la intercept y de su significación, así como las diferencias entre los R-squared.

Marcamos ahora Results, Plots y representamos los 2 gráficos que nos ofrece el programa: grafico de probabilidad normal y grafico de residuos y valores ajustados.


Normal Probability Plot
Standardized Residuals
4
2
0
-2
-4
-2
-1
0
1
2
Rankits
Shapiro-Wilk W 0.8985 P(W) 0.0236 23 cases


Regression Residual Plot
Standardized Residuals
4
2
0
-2
-4
0.00
0.03
0.06
0.09
0.12
Fitted values

Esto sugiere que no es correcto ni conveniente usar para estos datos el método de cuadrados mínimos. No hay una solución automática para datos que no cumplen las suposiciones del modelo de regresión lineal. En este caso en que la dispersión aumenta con el valor esperado, se han propuesto dos tipos de soluciones.

Una es la de aplicar “cuadrados mínimos ponderados”, la otra es la de aplicar transformaciones a los datos.

Cuadrados mínimos ponderados.

Recordemos que en el modelo lineal clásico con el que hemos trabajado asumimos que:

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

$$\mathbf{y_i = \alpha + \beta\, x_i + e_i} \quad \textbf{(para i=1,...,n)} \qquad (2)$$

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

1) $E(e_i) = 0$ para todo i
2) $Var(e_i) = \sigma^2$ (o sea es siempre la misma para todas las observaciones)
3) $e_1, e_2, ..., e_n$ son v. a. independientes

En algunos problemas, se sabe de antemano o se observa en los datos que no se cumple la suposición de homocedasticidad, es decir que los errores tienen igual varianza, sino que la varianza cambia con x, digamos en general que es de la forma:

$$Var(e_i) = f(x_i)$$

donde en principio la función es desconocida.

En problemas en los que e_i es el error de medición puede ser conocida de antemano la relación entre la variancia del error y x_i. Si tenemos la suerte de conocer de antemano esta relación, o proponemos esta relación observando los datos, la solución es simple. Las relaciones más usadas son que la varianza o la desviacion standard son proporcionales a x, o sea

a) $Var(e_i) = cte.\ x_i$ b) $DS\ (e_i) = cte.\ x_i$

Tanto a) como b) pueden escribirse como:

$$Var(e_i) = \theta\ v_i \qquad (19)$$

donde θ es una constante conocida o más frecuentemente un parámetro a estimar y v_i son constantes conocidas.

Supongamos que se cumple el modelo

$$\mathbf{Y_i = \alpha + \beta\ x_i + e_i \quad (para\ i=1,...,n)}$$

con las suposiciones 1) y 3), pero cambiando 2) por $Var(e_i) = \theta\ v_i$. Entonces si dividimos por $\sqrt{v_i}$ ambos miembros de (2) y llamamos

$$y_i^* = \frac{Y_i}{\sqrt{v_i}}. \quad ; \quad x_i^* = \frac{x_i}{\sqrt{v_i}}; \quad e_i^* = \frac{e_i}{\sqrt{v_i}}$$

obtenemos

$$y_i^* = \alpha \frac{1}{\sqrt{v_i}} + \beta\, x_i^* + e_i^* \quad \text{(para i=1,...,n)} \qquad (20)$$

donde ahora e_i^* cumple las suposiciones 1) a 3) del modelo lineal "clásico". Luego, para estimar los parámetros α y β se aplica cuadrados mínimos en (20), que equivale a minimizar

$$\sum_{i=1}^{n} \frac{1}{v_i} (y_i - (a + bx_i))^2 \qquad (21)$$

por lo que el método de estimación se llama **cuadrados mínimos pesados o ponderados**.

El peso de cada observación es inversamente proporcional a su varianza, lo que es intuitivamente razonable.

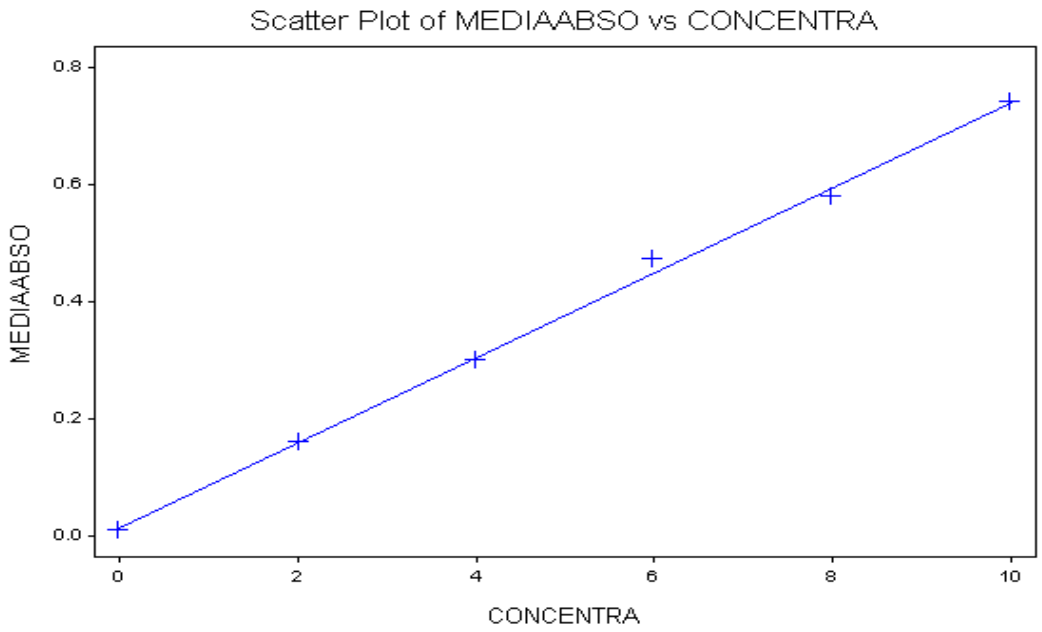
Statistix permite calcular los estimadores de cuadrados mínimos pesados.

Observación: Cuando para cada valor de x se hacen **varias observaciones** de y, se pueden estimar las varianzas de los errores por las varianzas muestrales en lugar de hacer suposiciones como en (19). Luego, se emplean estas varianzas estimadas en el método de cuadrados mínimos ponderados. Veremos a continuación un ejemplo en el que aplicaremos este método. Este método no es recomendable si hay pocas observaciones para cada x.

Ejemplo 3: En un experimento de calibración se analizaron soluciones standard con concentración conocida. Cada solución fue medida 10 veces. Se muestran las medias y los desvíos standard (DS) de las absorbancias observadas:

Concentración	0	2	4	6	8	10
Absorbancia						
Promedio	0.009	0.158	0.301	0.472	0.577	0.739
DS	0.001	0.004	0.010	0.013	0.017	0.022

Los datos de concentración y promedio de absorbancia se grafican a continuación:



Se observa en el gráfico que la relación es lineal. Pero en la tabla se ve que a medida que la verdadera concentración aumenta, crece la DS. Asi que es insostenible la suposición 2) del modelo de regresión en este ejemplo y es evidente que

$$Var(e_i) = f(x_i) = v_i$$

donde la función es creciente. Si no tenemos idea previa de la forma de esta función, se suele simplemente estimar cada v_i con el cuadrados de la DS correspondiente. Por ejemplo para $x_i=0$ estimamos v_i con el cuadrado de 0.001, etc. El estimador de mínimos cuadrados ponderados usa como pesos las inversas de estos v_i estimados. El Statistix permite calcular

cuadrados mínimos ponderados. Ingresamos los datos, calculamos los pesos y obtenemos la siguiente salida:

WEIGHTED LEAST SQUARES LINEAR REGRESSION OF MEDIAABSO

WEIGHTING VARIABLE: PESOS

PREDICTOR VARIABLES	COEFFICIENT	STD ERROR	STUDENT'S T	P
CONSTANT	0.00908	0.00105	8.67	0.0010
CONCENTRA	0.07376	0.00106	69.33	0.0000

R-SQUARED	0.9992	RESID. MEAN SQUARE (MSE)	1.12503
ADJUSTED R-SQUARED	0.9990	STANDARD DEVIATION	1.06067

Si, por error hubiésemos calculado la recta de cuadrados mínimos sin ponderaciones, hubiésemos obtenido:

UNWEIGHTED LEAST SQUARES LINEAR REGRESSION OF MEDIAABSO

PREDICTOR VARIABLES	COEFFICIENT	STD ERROR	STUDENT'S T	P
CONSTANT	0.01329	0.01056	1.26	0.2767
CONCENTRA	0.07254	0.00174	41.60	0.0000

R-SQUARED	0.9977	RESID. MEAN SQUARE (MSE)	2.128E-04
ADJUSTED R-SQUARED	0.9971	STANDARD DEVIATION	0.01459

Aunque en este ejemplo los estimadores de α y β son parecidos, debido a que los puntos están muy cerca de una recta, en otros ejemplo podría haber diferencias importantes, sin embargo cambia la significación de la ordenada al origen.