

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

Los métodos de regresión se usan para estudiar la relación entre dos variables numéricas.

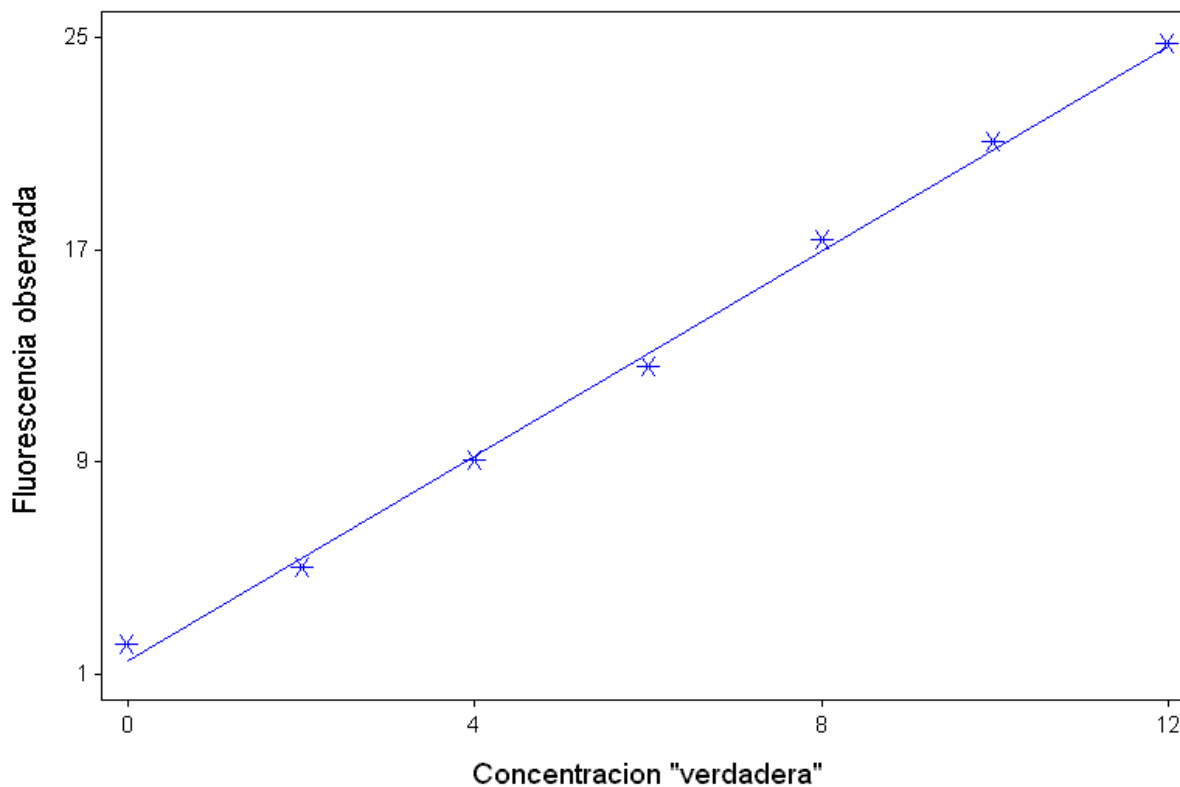
Puede interesar por ejemplo estudiar la relación entre estatura y perímetro cefálico de niños varones recién nacidos, o la relación entre la estatura del hijo y la estatura del padre (éste es un famoso ejemplo histórico de Galton 1880 que dió origen a la denominación "modelo de regresión").

**En química analítica se usa el modelo de regresión para calibrar un método de medición.**

**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" ("casi 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".

### **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:

$$\Sigma ( y_i - (a + b x_i) )^2 \quad (31)$$

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

$$y \cong 1.51786 + 1.93036 X$$

0

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

Por ejemplo si la verdadera concentración de fluoresceína de una muestra es 8, la ordenada de la recta es  $1.51786 + 1.93036 * 8 = 16.96$ . Obviamente esto no quiere decir que para las muestras que tengan concentración=8 la intensidad de la fluorescencia es 16.96 (ver gráfico, los puntos están muy cerca de la recta, pero no están sobre la recta).

## Modelo de regresión lineal

Para hacer inferencias (aplicar test de hipótesis y calcular intervalos de confianza) se necesita, como siempre, suponer un modelo, que se llama "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 varias variables independientes y en ese caso se llama "modelo de regresión lineal múltiple". Las suposiciones del modelo de regresión lineal simple (que es el que estudiaremos en este curso) son las siguientes.

**MODELO:** Se observan pares de valores  $(x_i, Y_i)$  para  $i=1, \dots, n$ , que cumplen:

$$Y_i = \alpha + \beta x_i + e_i \quad (\text{para } i=1, \dots, n) \quad (32)$$

donde  $e_1, e_2, \dots, e_n$  son variables aleatorias tales que

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

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

- 4)  $e_i \sim \text{Normal}$

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

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

Comentario: 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. El problema de estudiar la relación entre estatura y perímetro cefálico de recién nacidos es un ejemplo con  $x_i$ 's aleatorias. Justificaremos los resultados (estimadores, IC, tests) sólo para el modelo con  $x_i$ 's fijas, que es más simple, pero casi todos los resultados (IC y tests) 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  $i=1, \dots, n$ )
- 2\*)  $\text{Var}(Y_i) = \sigma^2$  (para  $i=1, \dots, n$ )
- 3\*)  $Y_1, Y_2, \dots, Y_n$  son vs as independientes
- 4\*)  $Y_i \sim \text{Normal}$

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

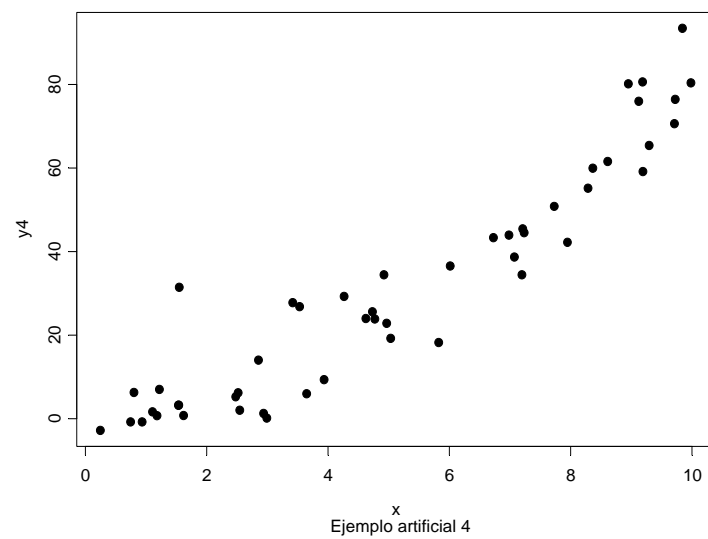
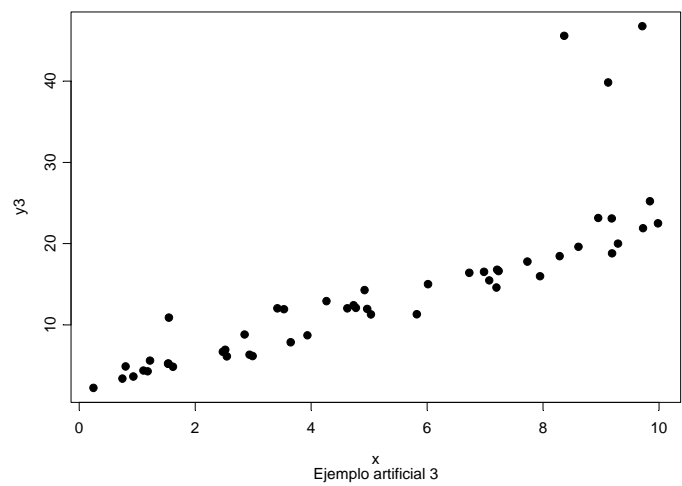
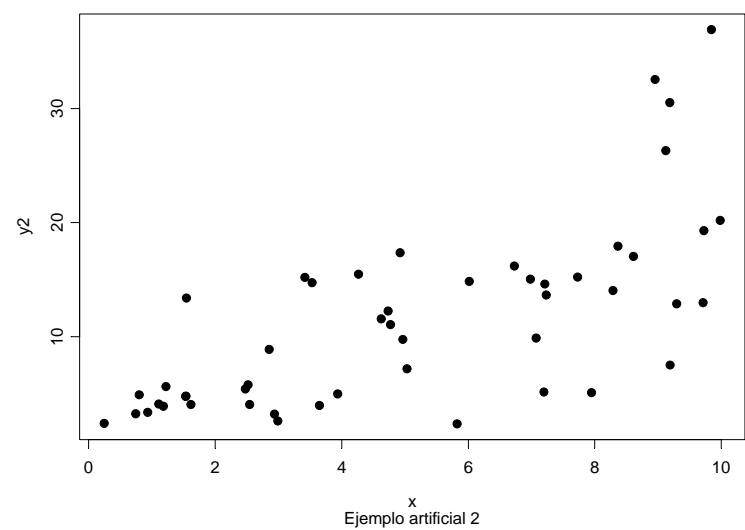
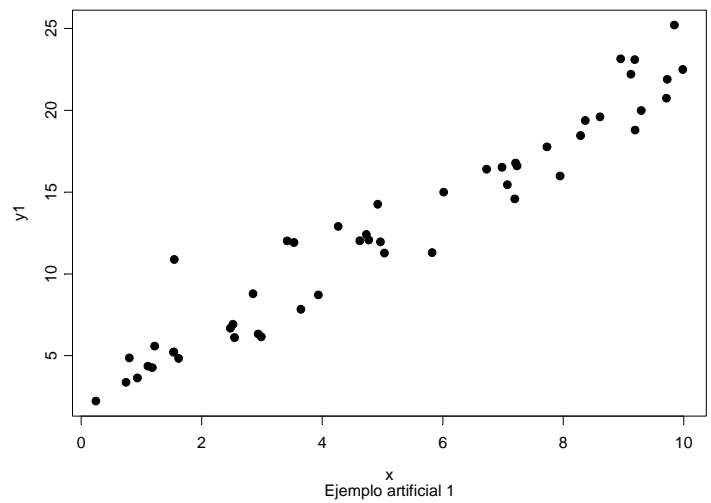
$$1*) \text{ a } 4*) \Leftrightarrow Y_i \text{ v. a. i.i.d. } N(\alpha + \beta x_i, \sigma^2)$$

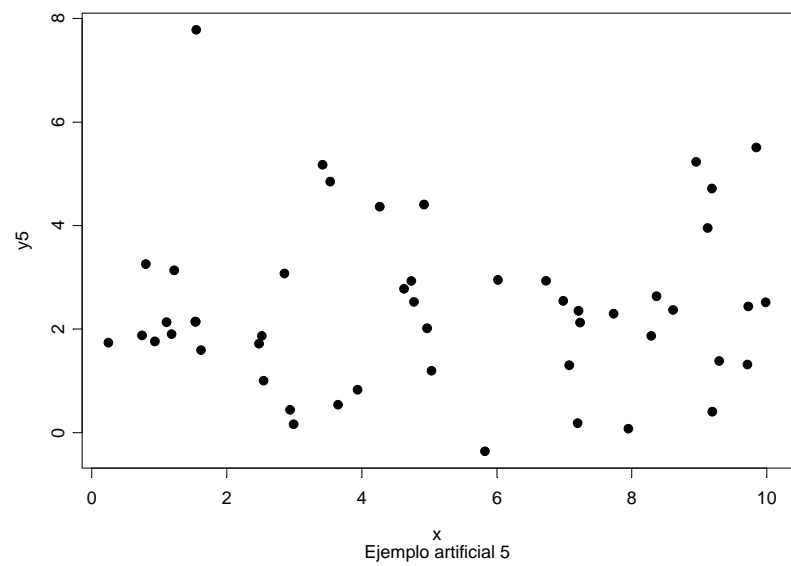
Comentario: en el modelo con  $x_i$ 's aleatorias, no hay que hacer ninguna suposición sobre la distribución de las  $x_i$ 's . Puede ser normal o no.

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.

Les presento a continuación gráficos de dispersión para cinco ejemplos artificiales, generados con un programa (generando números pseudoaleatorios). Algunos fueron generados de modo que cumplan todas las suposiciones del modelo de regresión lineal, otros no. ¿Detecta usted en cuáles de estos ejemplos no se cumple alguna de las suposiciones y cuál es la suposición que no se cumple?







## Estimadores de $\alpha$ y $\beta$ por el método de cuadrados mínimos

Llamemos  $\hat{\alpha}$  y  $\hat{\beta}$  a los valores de  $a$  y  $b$  que minimizan (31) que se llaman "estimadores de cuadrados mínimos" de  $\alpha$  y  $\beta$ . Se puede demostrar (derivando (31) e igualando a cero, hacerlo) que:

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{(\sum_{i=1}^n x_i Y_i) - n \bar{x} \bar{Y}}{(\sum_{i=1}^n x_i^2) - n \bar{x}^2} \quad (33)$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{x} \quad (34)$$

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:

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

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

$$s^2 = \frac{\sum_{i=1}^n \hat{e}_i^2}{n-2} = \frac{\sum_{i=1}^n (Y_i - (\hat{\alpha} + \hat{\beta} x_i))^2}{n-2} \quad (35)$$

es un estimador insesgado de  $\sigma^2$

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

$$E(\hat{\beta}) = \beta \quad (36)$$

$$\text{Var}(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (37)$$

y que

$$\text{cov}(\bar{Y}, \hat{\beta}) = 0 \quad (38)$$

Usando (36) a (38), se puede demostrar que

$$E(\hat{\alpha}) = \alpha \quad ; \quad \text{Var}(\hat{\alpha}) = \sigma^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

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

## Intervalo de confianza para $\beta$

Llamemos

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

El intervalo

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

(39)

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

## 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:

$$r^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2 - \sum_{i=1}^n (Y_i - (\hat{\alpha} + \hat{\beta}x_i))^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \quad (40)$$

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.

Siempre es

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

## 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 raíz 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:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (41)$$

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  $r = \text{raiz}(0.9978) = 0.9989$ . 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 (37) y (38) se puede demostrar que la varianza de este estimador es:

$$\text{Var}(\hat{\alpha} + \hat{\beta} x_0) = \sigma^2 \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \quad (42)$$

y que el intervalo de extremos

$$\left[ \hat{\alpha} + \hat{\beta} x_0 - t_{n-2; \alpha/2} \sqrt{s^2 \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]} ; \hat{\alpha} + \hat{\beta} x_0 + t_{n-2; \alpha/2} \sqrt{s^2 \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]} \right] \quad (43)$$

es un IC con nivel  $1-\alpha$  para el valor esperado de  $Y$ , para  $x = x_0$ .

## **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  $(x_i, Y_i)$  ( $i=1, \dots, n$ ).

Supongamos ahora que hacemos una nueva observación, pero sólo conocemos su valor de x (llamémoslo  $x_{n+1}$ ), no conocemos su valor de Y, que llamaremos  $Y_{n+1}$ . Queremos en esta sección 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  $1-\alpha$ ) (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  $e_1, e_2, \dots, e_n$ .

Es intuitivamente razonable que el mejor predictor de  $Y_0$  sea:

$$\hat{Y}_{n+1} = \hat{\alpha} + \hat{\beta} x_{n+1} \quad (44)$$

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

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

y que el intervalo de extremos

$$\left[ \hat{Y}_{n+1} - t_{n-2; \alpha/2} \sqrt{s^2 \left[ 1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]}; \hat{Y}_{n+1} + t_{n-2; \alpha/2} \sqrt{s^2 \left[ 1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]} \right] \quad (45)$$

es un "intervalo de predicción" con nivel  $1-\alpha$  para una nueva observación  $Y_0$ .

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

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

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

No aparece directamente en la salida el IC para  $\beta$ , pero es fácil obtenerlo usando (39). Si queremos un IC al 95%, necesitamos el valor de t con  $7-2=5$  gl, con  $p=0.05$  en las dos colas. En Statistix o en tablas obtenemos:  $t_{5; 0.025} = 2.57$  y, reemplazando en (39):

$$1.93036 \pm 2.57 * 0.04090$$

$$1.93036 \pm 0.10511$$

o, redondeando

**IC para  $\beta$  con nivel 95%: [1.83; 2.04]**

El IC al 95% para  $\alpha$  se obtiene en forma análoga:

$$1.51786 \pm 2.57 * 0.29494$$

redondeando:

$$1.52 \pm 0.76$$

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

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

Para obtener el intervalo de predicción para  $Y_{n+1}$  hay que usar la expresión (45). Pero 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 FLUORESCENCE**

LOWER <b>PREDICTED</b> BOUND	15.753	LOWER FITTED BOUND	16.491
<b>PREDICTED</b> VALUE	16.961	FITTED VALUE	16.961
UPPER <b>PREDICTED</b> BOUND	18.169	UPPER FITTED BOUND	17.431
<b>SE (PREDICTED VALUE)</b>	<b>0.4699</b>	<b>SE (FITTED VALUE)</b>	<b>0.1829</b>
UNUSUALNESS (LEVERAGE)	0.1786		
PERCENT COVERAGE	95.0		
CORRESPONDING T	2.57		

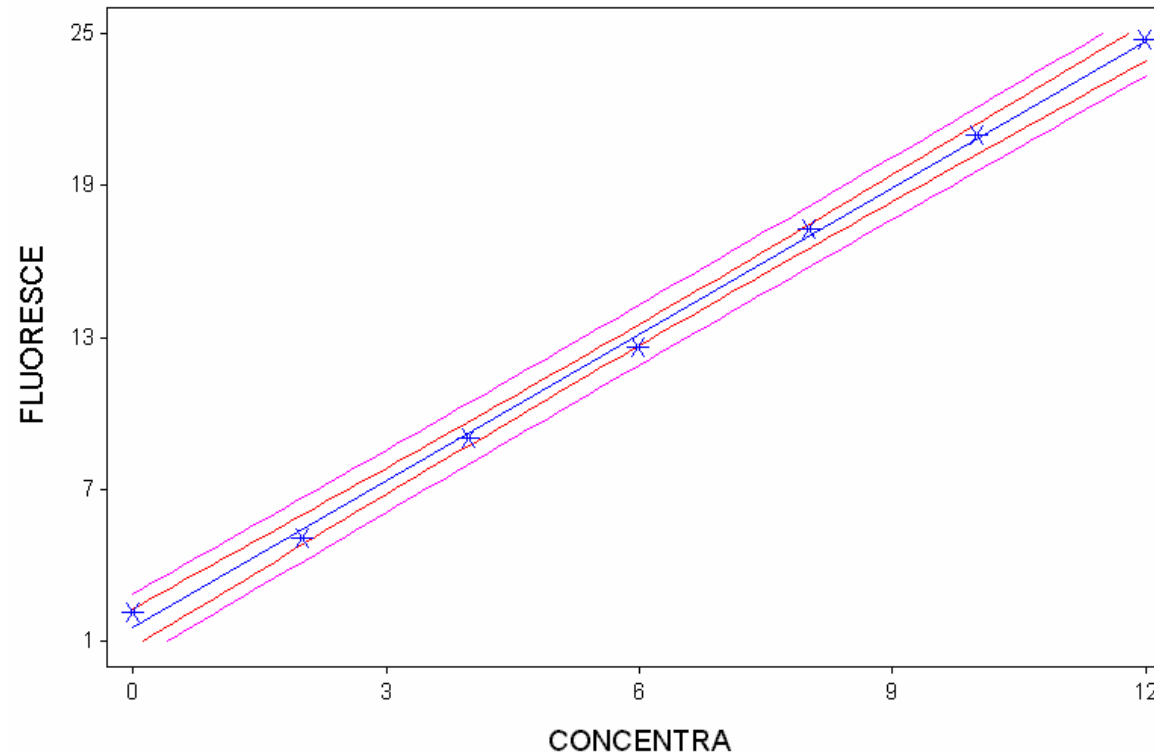
PREDICTOR VALUES: CONCENTRA = 8.0000

Vemos que el predictor 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. Observar este intervalo y ver que tiene menor longitud.

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



Con Statistix también podemos representar gráficamente los intervalos de predicción 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:



**Comentario:** 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?

## **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  $(x_i, Y_i)$  ( $i=1, \dots, n$ ).

Supongamos ahora que hacemos una nueva observación, pero sólo conocemos su valor de  $Y$ , no conocemos su valor  $x$ . Queremos en esta sección 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 (32) el papel de las variables “ $Y$ ” y “ $X$ ” y luego aplicar "predicción" (o sea (44) y (45)).
- 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 "Y" en (32) 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  $e_1, e_2, \dots, e_n$ .

Despejando  $x$

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

Como no tenemos información ninguna sobre  $e$  y, además, de  $\alpha$  y  $\beta$  sólo conocemos los estimadores, es intuitivamente razonable estimar  $x$  con:

$$\hat{x} = \frac{Y - \hat{\alpha}}{\hat{\beta}}$$

(46)

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

$$\hat{\text{Var}}(\hat{x}) = \frac{s^2}{\hat{\beta}^2} \left[ 1 + \frac{1}{n} + \frac{(Y - \bar{Y})^2}{\hat{\beta}^2 \sum_{i=1}^n (x_i - \bar{x})^2} \right] \quad (47)$$

Llamando

$$\text{ES}(\hat{x}) = \sqrt{\hat{\text{Var}}(\hat{x})} \quad (48)$$

el intervalo

$$\hat{x} \pm t_{n-2; \alpha/2} \text{ES}(\hat{x}) \quad (49)$$

es un intervalo de confianza con nivel aproximado  $1-\alpha$  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  $\bar{Y}_m$  al promedio de las m observaciones Y's hechas en esa muestra. Entonces (46) y (47) se modifican así:

$$\hat{x} = \frac{\bar{Y}_m - \hat{\alpha}}{\hat{\beta}} \quad (46^*)$$

$$\hat{\text{Var}}(\hat{x}) = \frac{s^2}{\hat{\beta}^2} \left[ \frac{1}{m} + \frac{1}{n} + \frac{(\bar{Y}_m - \bar{Y})^2}{\hat{\beta}^2 \sum_{i=1}^n (x_i - \bar{x})^2} \right] \quad (47^*)$$

Quedando (48) y (49) sin cambios.

**Ejemplo:** Lamentablemente Statistix 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 (46):

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

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 (47). Vemos que todo lo que se necesita para calcular (47) 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	<b>18.667</b>
FLUORESC	7	<b>13.100</b>	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

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

Reemplazamos ahora en (47):

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

Luego

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

Aplicando (49) 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 deberían tomar las muestras standard en el experimento de calibración para disminuir la longitud de los intervalos de confianza para x?

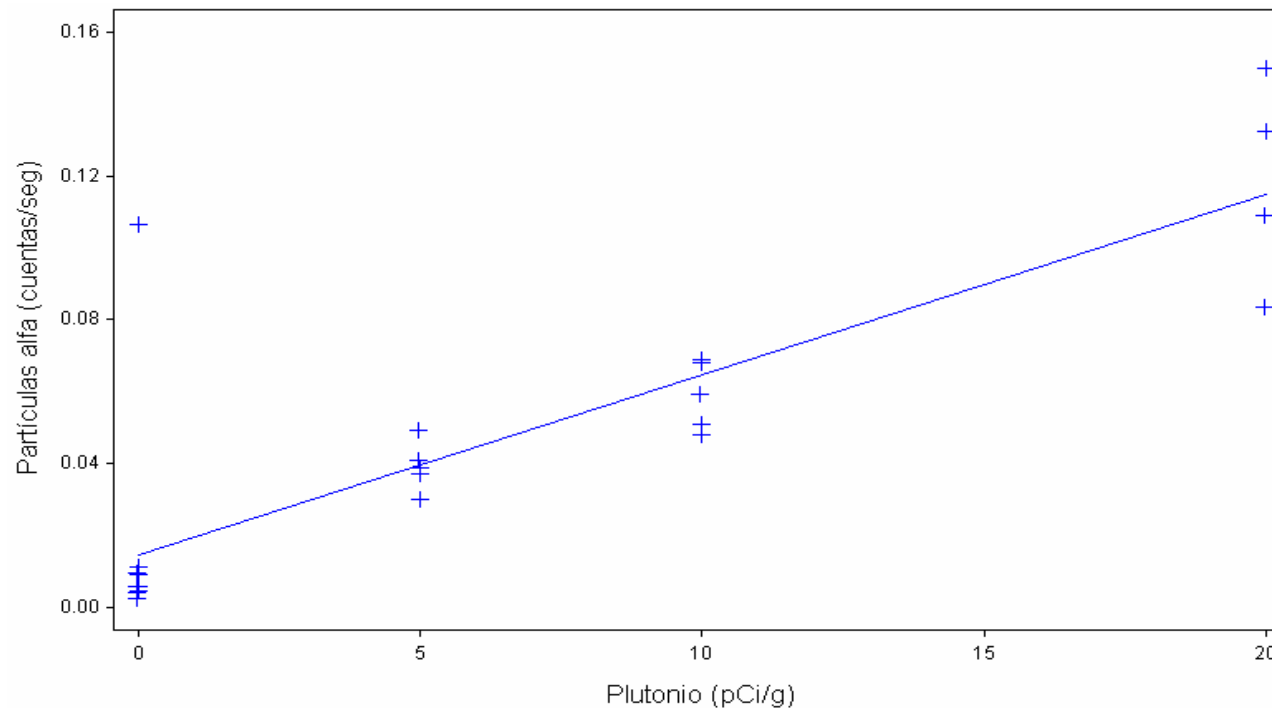


## **Diagnóstico del modelo de regresión.**

**Ejemplo 2:** En la clase de estadística descriptiva comenzamos a analizar los datos de un experimento de calibración. Dijimos en esa clase:

“Cuando el plutonio está presente en pequeñas cantidades mezclado con otros materiales es difícil detectarlo. Una forma de detectarlo es medir las partículas alfa que emite. En una investigación para estudiar la relación entre la cantidad de plutonio y la emisión de partículas alfa, se midieron varias veces cuatro materiales standards para los que se sabe que la actividad de plutonio (0, 5, 10 y 20 picocuries por gramo (pCi/g)). Los resultados de estas mediciones están en el archivo plutonio.xls.”

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



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

Otra forma que a veces ayuda a detectar fallas en el modelo, es estimar los parámetros del modelo y luego hacer gráficos para el "diagnóstico" del modelo. Para ello vamos a "Statistics", "Linear Models", "Linear Regression" y obtenemos:

UNWEIGHTED LEAST SQUARES LINEAR REGRESSION OF PARTALFA

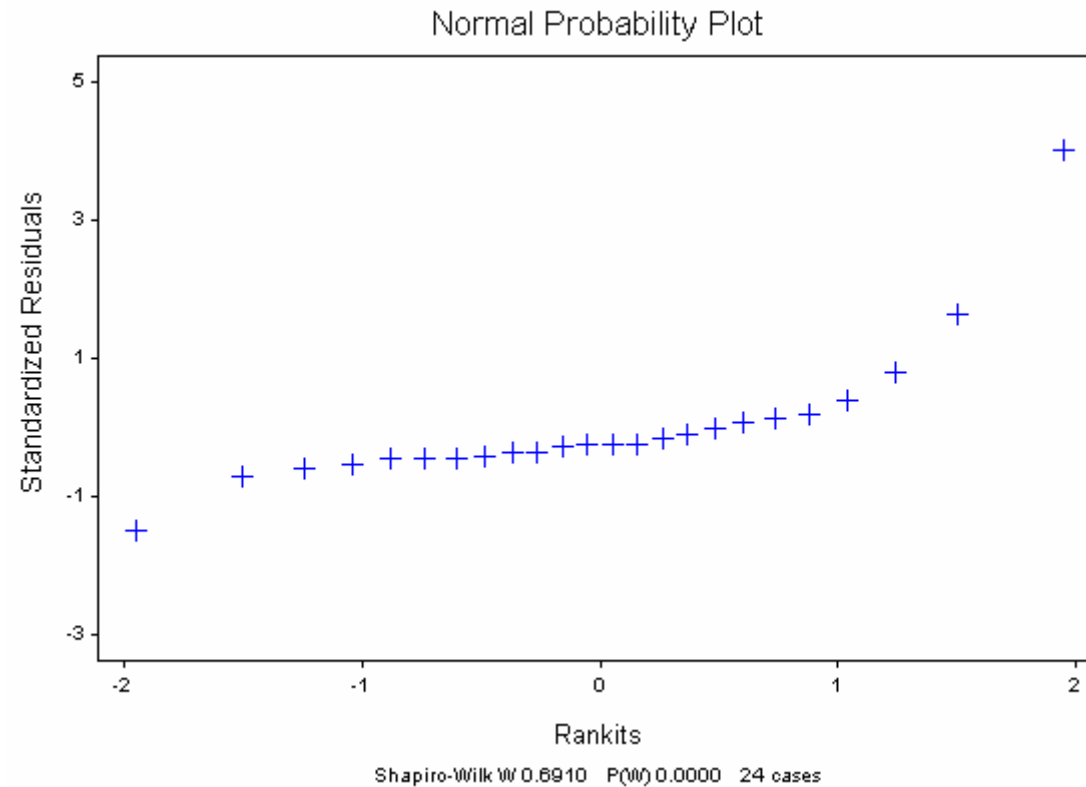
PREDICTOR VARIABLES	COEFFICIENT	STD ERROR	STUDENT'S T	P
-----	-----	-----	-----	-----
CONSTANT	0.01453	0.00653	2.23	0.0366
PLUTONIO	0.00501	6.778E-04	7.40	0.0000

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

Con esta salida a la vista, se marca "Results", "Plots", "Normal Probability Plot", el programa hace el siguiente gráfico para estudiar 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 excluimos el dato atípico y volvemos a estimar los parámetros de la regresión y hacer gráficos con los residuos.

UNWEIGHTED LEAST SQUARES LINEAR REGRESSION OF PARTALFA

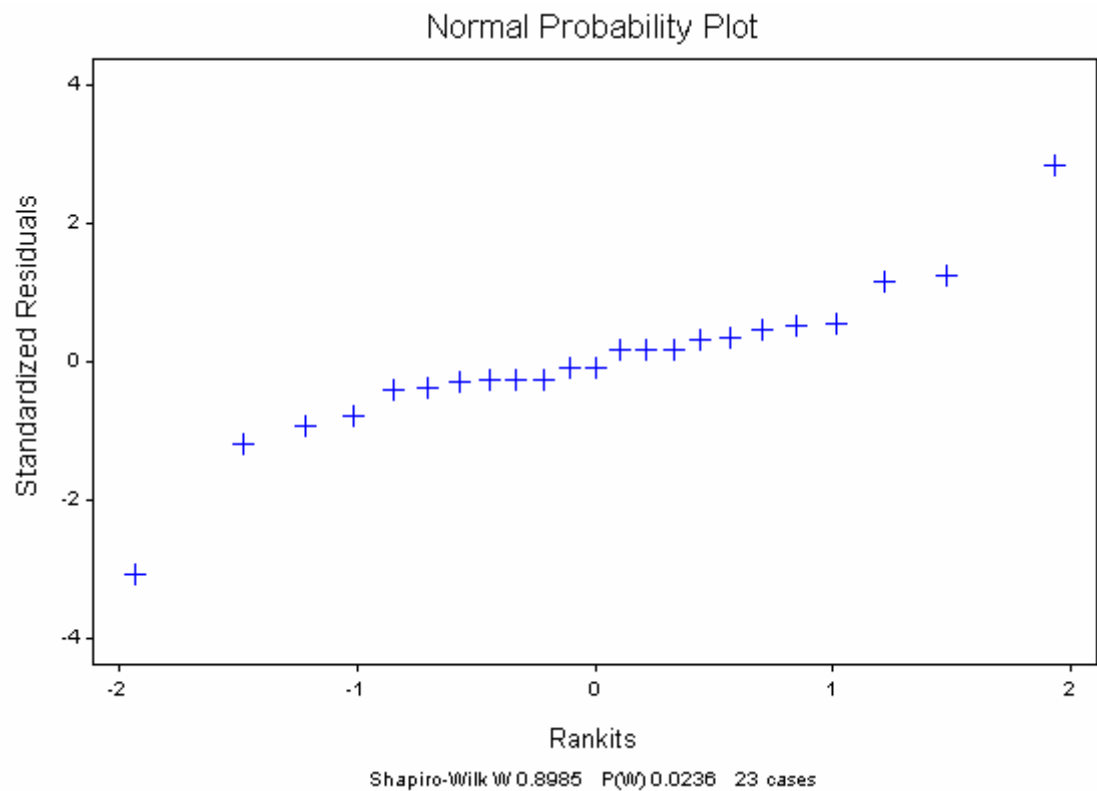
PREDICTOR VARIABLES	COEFFICIENT	STD ERROR	STUDENT'S T	P
-----	-----	-----	-----	-----
CONSTANT	0.00703	0.00360	1.95	0.0641
PLUTONIO	0.00554	3.659E-04	15.13	0.0000

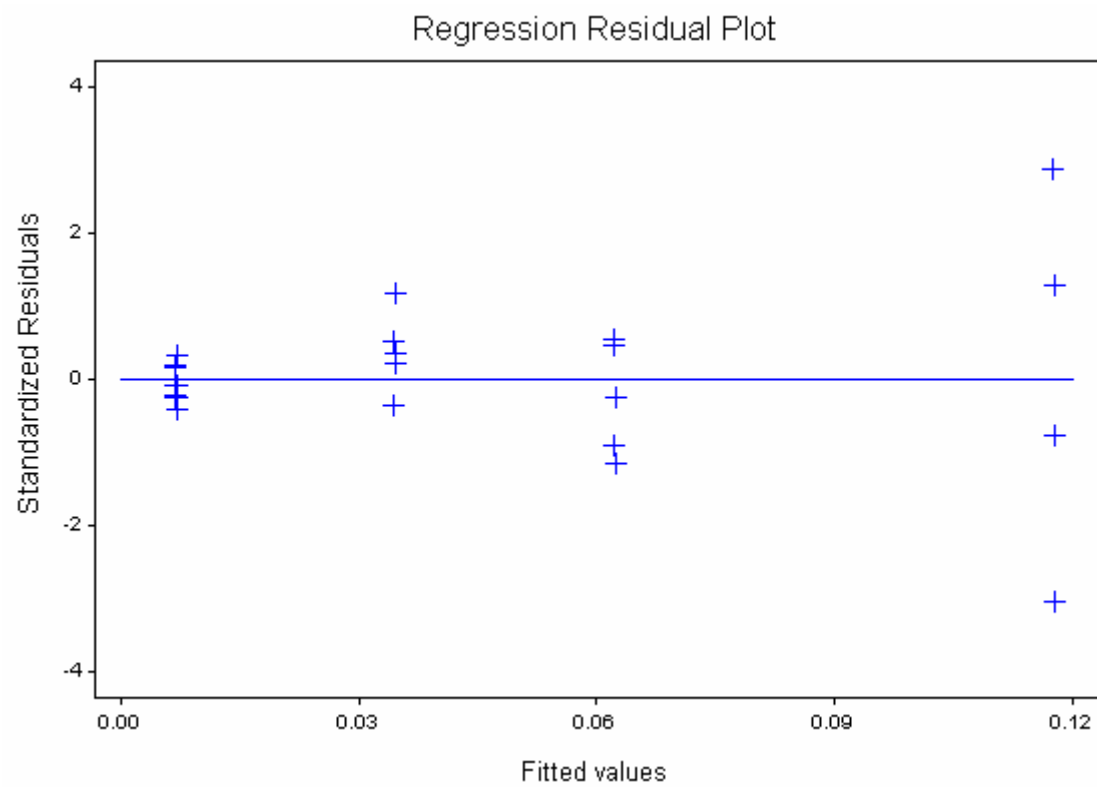
R-SQUARED	0.9160	RESID. MEAN SQUARE (MSE)	1.580E-04
ADJUSTED R-SQUARED	0.9120	STANDARD DEVIATION	0.01257

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

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





Esto indica 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  $i=1, \dots, n$ , que cumplen:

$$Y_i = \alpha + \beta x_i + e_i \quad (\text{para } i=1, \dots, n) \quad (32)$$

donde  $e_1, e_2, \dots, e_n$  son variables aleatorias tales que

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

En algunos problemas, se sabe de antemano o se observa en los datos que no se cumple la suposición de que los errores tienen igual varianza ( supuesto 2)), sino que la varianza cambia con  $x$ , digamos en general que es de la forma:

$$\text{Var}(e_i) = f(x_i) \quad (50)$$

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 lo que se quiere medir ( $x_i$ ). Si tenemos la suerte de conocer de antemano algo sobre la relación que hay entre la varianza y el valor de  $x$ , o proponemos esta relación observando los datos, la solución es simple. Las relaciones más usadas son que la varianza o la desviación standard son proporcionales a  $x$ , o sea

$$\begin{array}{ll} \text{a) } \text{Var}(e_i) = \text{cte. } x_i & \text{b) } \text{DS}(e_i) = \text{cte. } x_i \end{array} \quad (51)$$

Tanto a) como b) pueden escribirse como:

$$\text{Var}(e_i) = \theta v_i$$

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

Supongamos ahora que se cumple el modelo

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

con las suposiciones 1) y 3), pero cambiando 2) por  $\text{Var}(e_i) = \theta v_i$ . Entonces si dividimos por  $\sqrt{v_i}$  ambos miembros de (32) 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

$$Y_i^* = \alpha \frac{1}{\sqrt{v_i}} + \beta x_i^* + e_i^* \quad (\text{para } i=1, \dots, n) \quad (52)$$

donde ahora  $e_i^*$  cumple las suposiciones 1) a 3) del modelo lineal “clásico”. Luego para estimar los parámetros  $\alpha$  y  $\beta$  se aplica cuadrados mínimos en (52), que equivale a minimizar

$$\Sigma (1/v_i) ( y_i - (a + b x_i) )^2 \quad (53)$$

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

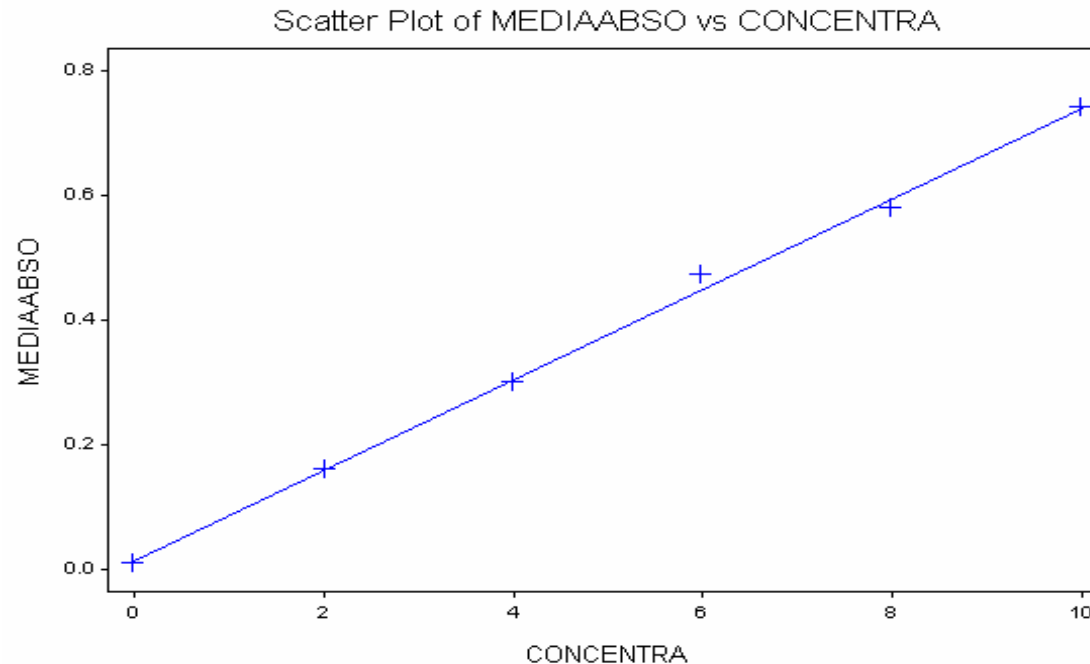
El peso de cada observación es inversamente proporcional a su varianza, lo que es intuitivamente razonable. Statistix permite calcular los estimadores de cuadrados mínimos pesados.

Cuando para cada valor de  $x$  se hacen **varias observaciones** de  $y$ , se puede estimar las varianzas de los errores por las varianzas muestrales en lugar de hacer suposiciones como en (51). 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 el libro de Miller<sup>1</sup> se presenta el siguiente problema. En un experimento de calibración se analizaron soluciones standard con concentración conocida. Cada solución fue medida 10 veces. Se muestran las medias y las DS de las absorbancias observadas:

<b>Concentración</b>	0	2	4	6	8	10
<b>Absorbancia</b>						
<b>Promedio</b>	0.009	0.158	0.301	0.472	0.577	0.739
<b>DS</b>	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 (esto no se ve en el gráfico, porque no tenemos los datos originales del experimento, verdad?) se ve que a medida que la verdadera concentración aumenta, crece la DS. Así que es insostenible la suposición 2) del modelo de regresión en este ejemplo y es evidente que

$$\text{Var}(e_i) = f(x_i) = v_i$$

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 cuadrado 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  $\alpha$  y  $\beta$  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.

**Transformaciones:** Otra forma de solucionar la falla en la suposición 2) que suele ser al mismo tiempo útil para lograr que los residuos tengan una distribución más próxima a la normal es aplicar transformaciones a los datos. Se suele probar con la transformación logarítmica. Esta es la transformación más usada en química experimental.

En el “Curso de Introducción al conocimiento científico experimental” de la Dra. Celia Croto, cap 17 (<http://www.quimicaviva.qb.fcen.uba.ar/contratapa/aprendiendo/capitulo17.htm>) se explica:

“En la ciencia experimental se aplican los logaritmos en distintas situaciones, una de ellas es cuando se analizan los resultados de un experimento en el que se trabaja con diluciones seriadas. Las más comunes son las diluciones en base 10, pero en serología, por ejemplo, se usan las diluciones en base 2. Cuando una sustancia se diluye en forma seriada de modo de obtener diluciones de la misma de 1/10; 1/100; 1/1000; 1/10.000 y 1/100.000, son las potencias de  $10^n$ . Si fueran diluciones en base dos, tendríamos 1/2; 1/4, 1/8, 1/16, 1/32, 1/64 y así utilizando las potencias de  $2^n$ .”

También se suele usar la transformación raíz cuadrada o elegir una transformación en una familia de transformaciones que incluye al logaritmo, la raíz cuadrada y a otras (método de Box y Cox). Existen métodos para elegir una transformación, dentro de una familia. No trataremos estos métodos en este curso. Se pueden ver por ejemplo en el capítulo 3 del libro de Neter y otros<sup>2</sup>.



## **Referencias:**

1. Miller y Miller. Estadística para Química Analítica. Addison Wesley.
2. Neter, Kutner, Nachtsheim y Wasserman. Applied Linear Statistical Models. Mc Graw Hill.

## Metodo Gráfico: QQ-Plot (o Plor Normal)

En un par de ejes coordenados se representan los percentiles teóricos de la distribución Normal y en el otro eje los percentiles de la muestra.

- Si la muestra realmente proviene de una población Normal, el gráfico debería estar, salvo fluctuaciones aleatorias, sobre una recta ( $y=x$ ).
- Las distintas curvaturas que pueden oresentar estos gráficos, nos sugieren que tipo de distribución puede tener la población de origen.

El SX realiza estos gráficos y además nos provee el valor del estadístico de Shapiro-Wilk,  $W$ , que mide la asociación entre los percentiles poblacionales y muestrales. Este test es recomendado por muchos autores para muestras de tamaño medio. Si tenemos una muestra aleatoria  $u_1, \dots, u_n$ , tales que  $u_i \sim F$ , la hipótesis nula del test es que los datos provienen de una muestra con distribución normal, es decir

$$H_0: F \sim \text{Normal}$$

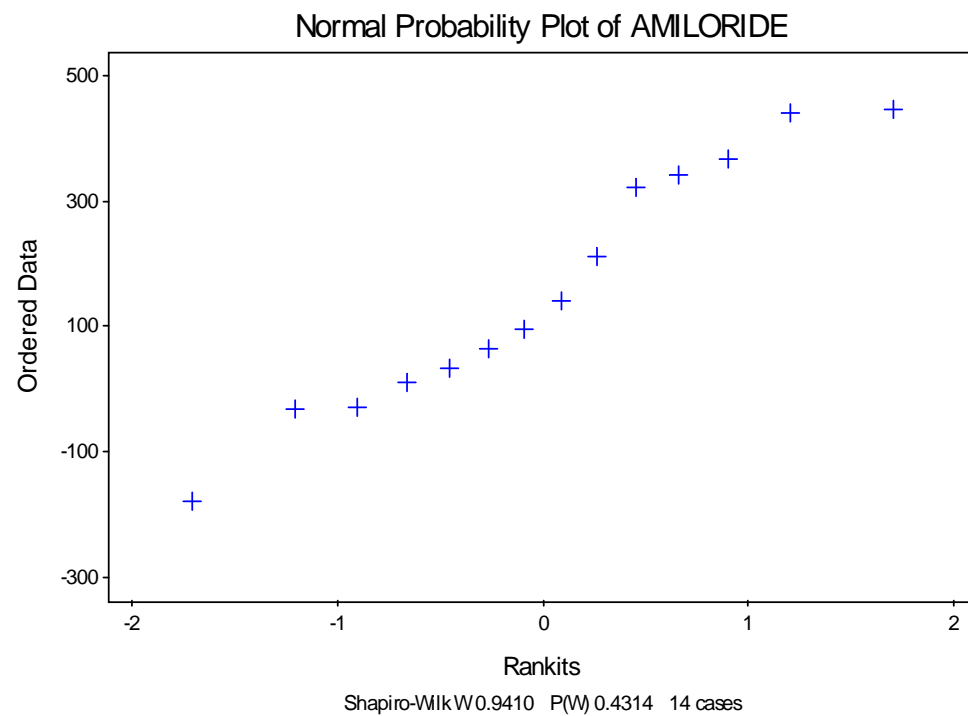
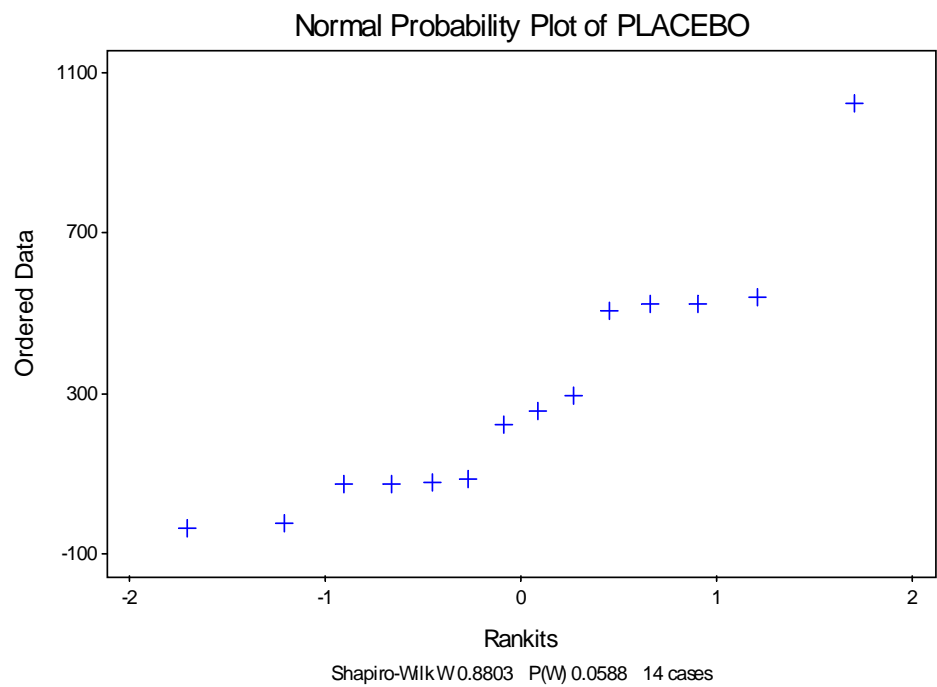
La distribución teórica de  $W$  también está tabulada, así que podemos comparar el valor obtenido con los de la tabla correspondiente.

SX nos devuelve el valor del estadístico  $W$  y un valor  $P(W)$  que nos indica si el  $W$  observado es grande o pequeño.

Veamos un ejemplo:

Los siguientes datos corresponden a la reducción en la FVC en pacientes con fibrosis quística después de 25 semanas de tratamiento cuando fueron tratados con la droga amiloride y con placebo:

Placebo	Amiloride
224	213
80	95
75	33
541	440
74	-32
85	-28
293	445
-23	-178
525	367
-38	140
508	323
255	10
525	65
1023	343

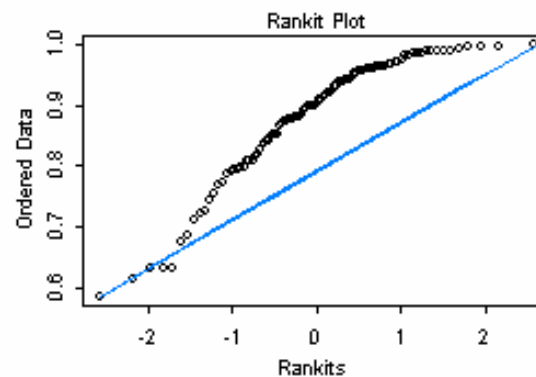
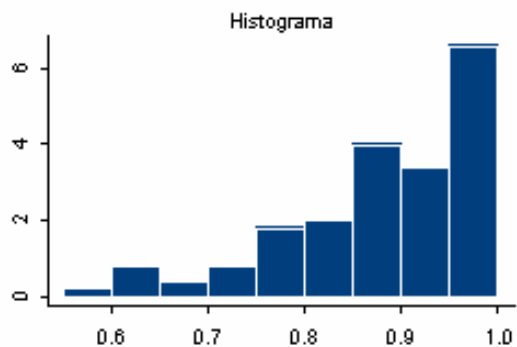


Los valores de  $W$  y de  $P(W)$  obtenidos son:

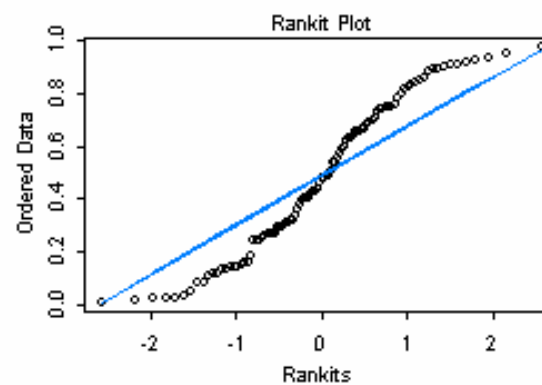
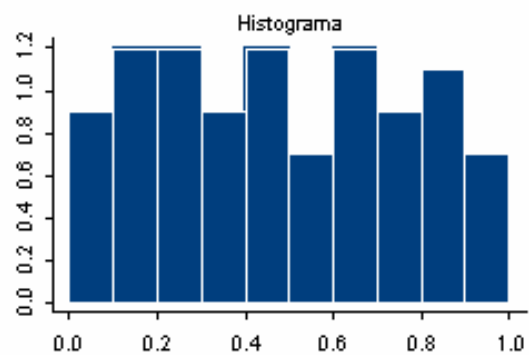
	$W$	$P(W)$
Amiloride	0.9410	0.4314
Placebo	0.8803	0.0588

Por lo tanto no es razonable que las diferencias de FVC de los placebos sean normales, mientras que este supuesto no parece contradecirse en el caso de los tratados con Amiloride.

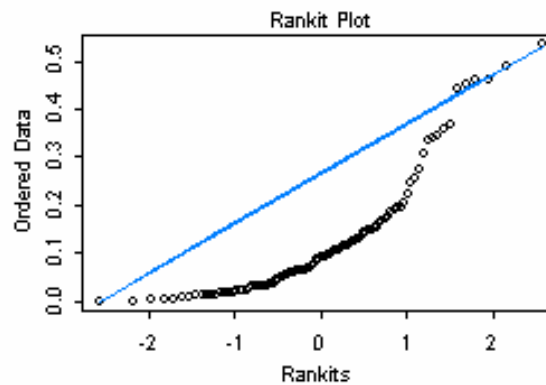
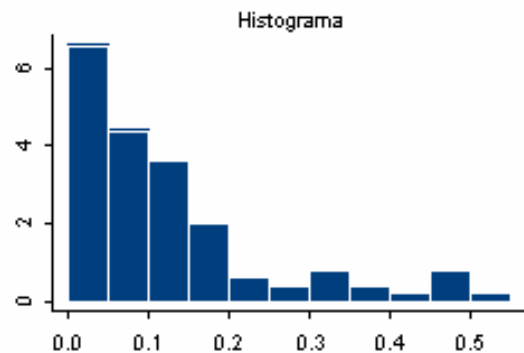
Asimétrica a Izquierda



Simétrica con Colas Livianas



Asimétrica a Derecha



Simétrica con Colas Pesadas

