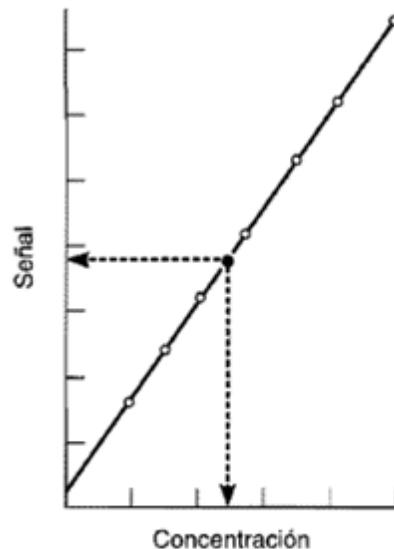


## RELACIÓN ENTRE DOS VARIABLES NUMÉRICAS REGRESIÓN LINEAL SIMPLE. CORRELACIÓN

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.

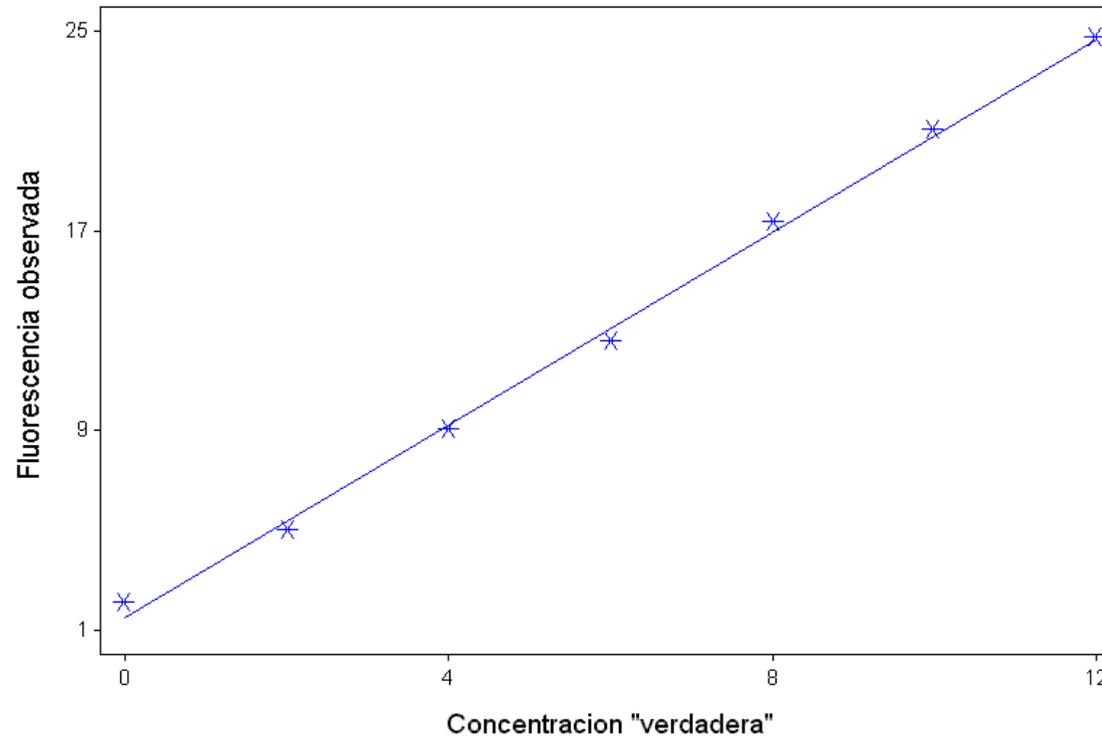
Veamos un ejemplo numérico:

**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 estándar, estudiar la relación entre las mediciones y el "verdadero valor". Esta relación permitirá en el futuro, medir una muestra desconocida y conocer aproximadamente su valor verdadero.

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". Veremos cómo hallar esta recta.

Recordemos que la **ecuación de una recta** es de la forma

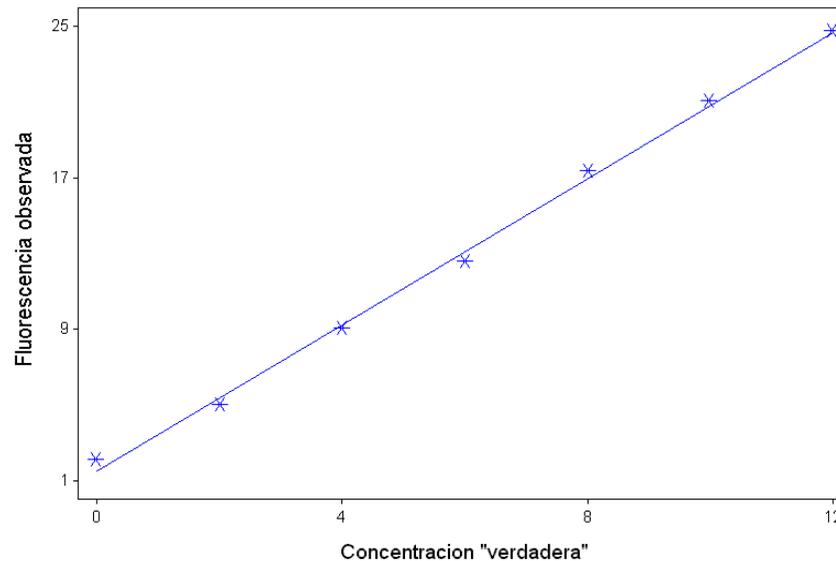
$$y = \alpha + \beta x$$

**Ordenada al origen** ←      → **Pendiente**

## Recta de cuadrados mínimos.

La recta representada en el gráfico anterior es *la recta de cuadrados mínimos*. Esta es la recta 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:

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



PREDICTOR VARIABLES	COEFFICIENT	STD ERROR	STUDENT'S T	P
-----	-----	-----	-----	-----
CONSTANT	<b>1.51786</b>	0.29494	5.15	0.0036
CONCENTRA	<b>1.93036</b>	0.04090	47.20	0.0000

Observando los "coeficientes" de la salida vemos que la recta de mínimos cuadrados 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 fuera 8, la ordenada de acuerdo con esta recta sería

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

Esto no quiere decir que para las 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 hacer inferencias, es decir realizar tests de hipótesis o calcular intervalos de confianza, se necesita suponer un modelo, que llamaremos "*modelo de regresión lineal simple*".

La palabra "simple" se debe a que 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  $i=1, \dots, n$ , que cumplen:

$$y_i = \alpha + \beta x_i + e_i \quad (2)$$

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

- 1)  $E(e_i) = 0$  para todo  $i$
- 2)  $\text{Var}(e_i) = \sigma^2$
- 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 algo más:

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

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

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

### **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 y 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:

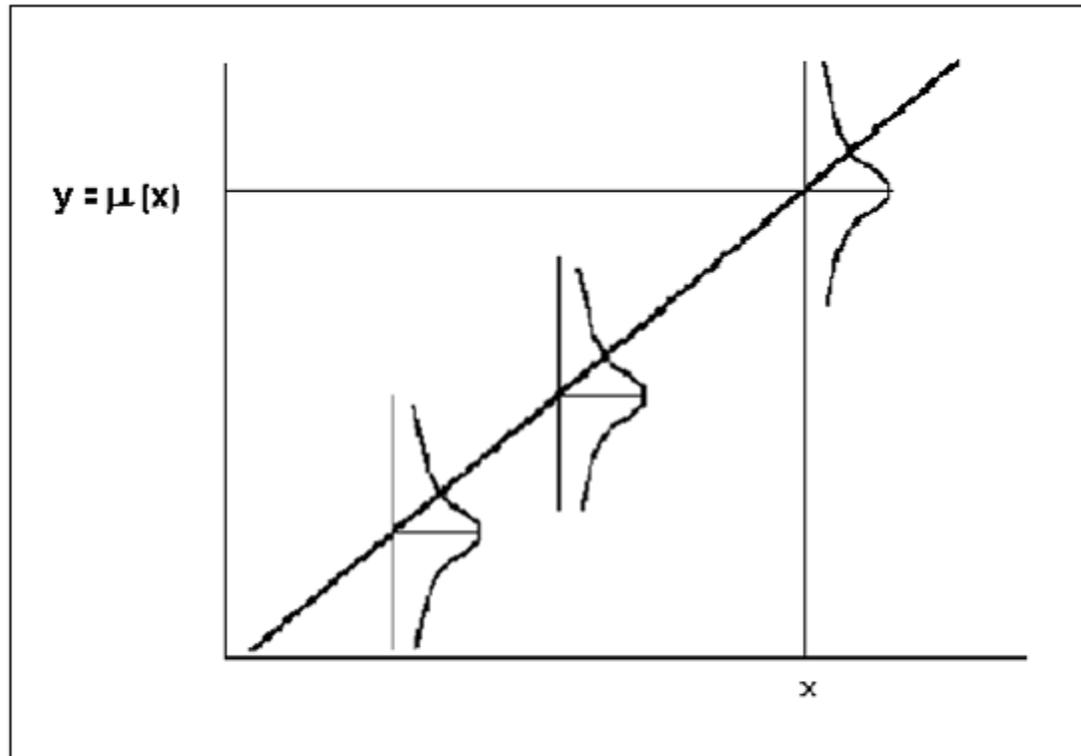
- 1\*)  $E(y_i) = \alpha + \beta x_i$  (para  $i=1, \dots, n$ )
- 2\*)  $\text{Var}(y_i) = \sigma^2$  (para  $i=1, \dots, n$ )
- 3\*)  $y_1, y_2, \dots, y_n$  son v. a. independientes
- 4\*)  $y_i \sim \text{Normal}$

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

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

**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 la respuestas  $y_i$  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 $\alpha$ y $\beta$ 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 mínimos" de  $\alpha$  y  $\beta$ .

¿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

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

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

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  $\sigma^2$ :**  $\sigma^2$  es la varianza de  $e_i$ , es decir  $\sigma^2 = \text{Var}(e_i)$ . 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

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

es un estimador insesgado de  $\sigma^2$ .

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

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

y además

$$\text{cov}(\bar{y}, \hat{\beta}) = 0 \quad (7)$$

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

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

Si la suposición 4) (de normalidad) se cumple, el intervalo

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

es un IC para  $\beta$  con nivel  $1-\alpha$ .

**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:

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

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 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:

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

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.

Veamos todo esto en el ejemplo.

```
concentra=c(0,2,4,6,8,10,12)
fluo<- c(2.1,5,9,12.6,17.3,21,24.7)
salida<- lm(fluo~concentra)
summary(salida)
```

Call:

```
lm(formula = fluo ~ concentra)
```

Residuals:

```
1 2 3 4 5 6 7
```

```
0.58214 -0.37857 -0.23929 -0.50000 0.33929 0.17857 0.01786
```

**Coefficients:**

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.5179	0.2949	5.146	0.00363 **
concentra	1.9304	0.0409	47.197	8.07e-08 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**Residual standard error: 0.4328** on 5 degrees of freedom

**Multiple R-squared: 0.9978**, Adjusted R-squared: 0.9973

F-statistic: 2228 on 1 and 5 DF, p-value: 8.066e-08

En el ejemplo de la fluorescencia, tenemos que

R-SQUARED 0.9978

y, como la pendiente es positiva, es  $r = (0.9978)^{1/2} = 0.9989$ . 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:

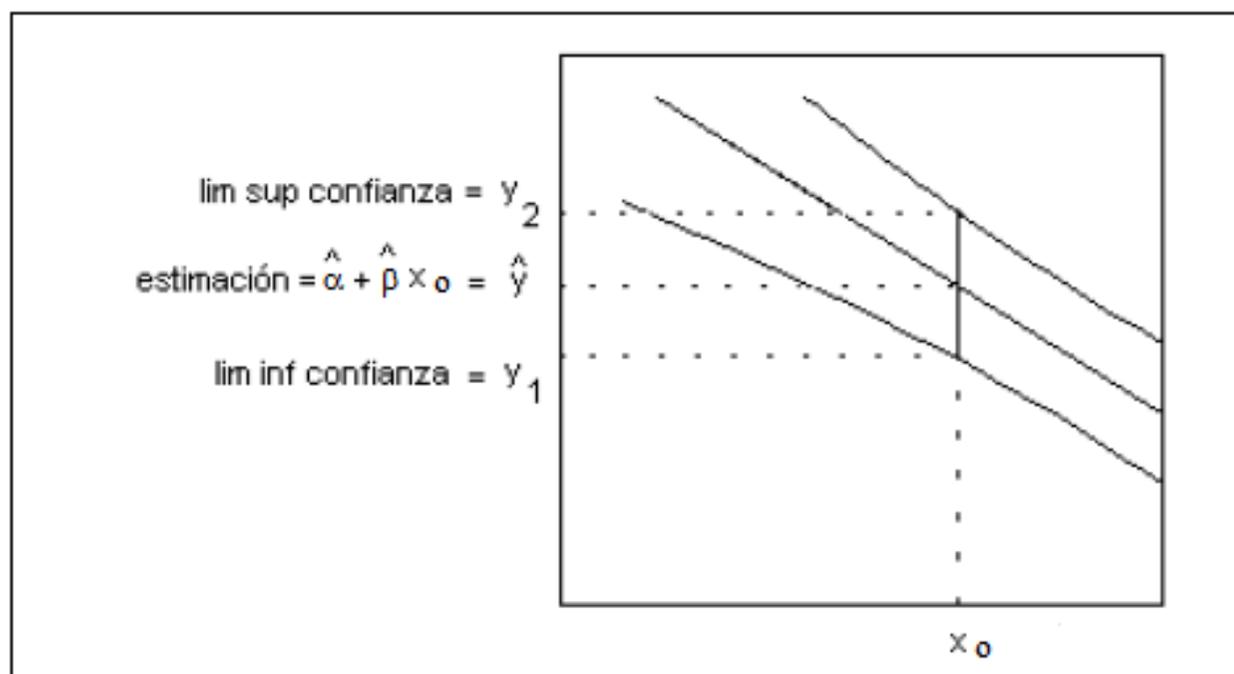
$$\text{Var}(\hat{\alpha} + \hat{\beta} x_0) = \sigma^2 \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \quad (11)$$

y que el intervalo de extremos

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

es un IC con nivel  $1-\alpha$  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  $(x_i, y_i)$  ( $i=1, \dots, n$ ).

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  $1-\alpha$ ) (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  $y_{n+1}$  sea:

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

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

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

y que el intervalo de extremos

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

es un "**intervalo de predicción**" con nivel  $1-\alpha$  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:

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

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

No aparece directamente en la salida el IC para  $\beta$ , pero es fácil obtenerlo usando (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. 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  $\beta$  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:

$$1.52 \pm 0.76$$

**IC para  $\alpha$  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:

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

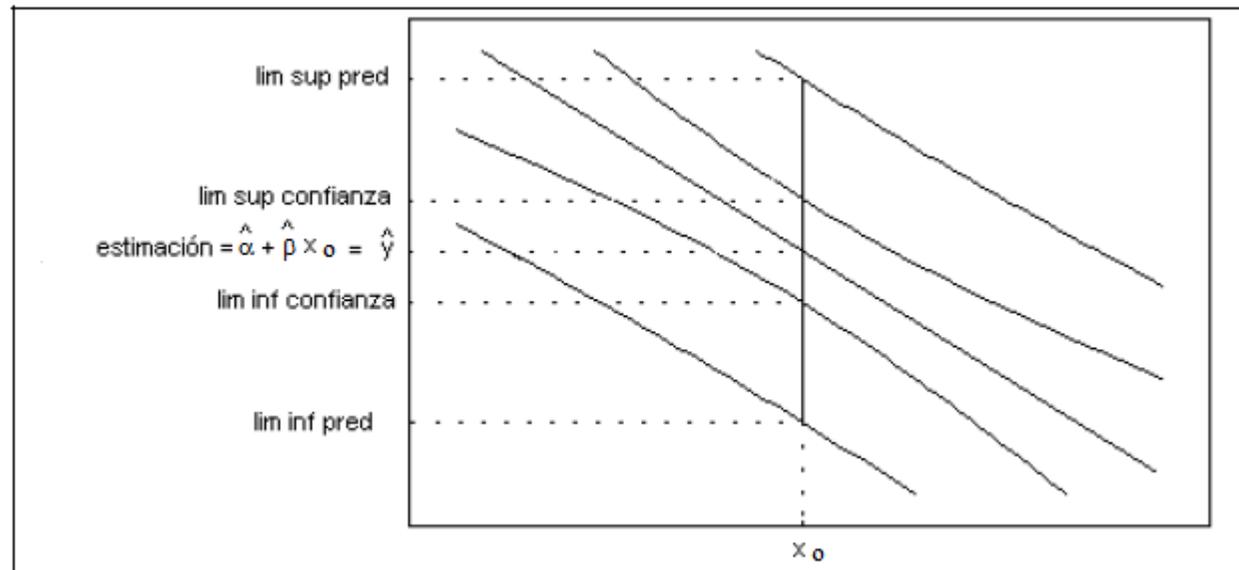
Para obtener el intervalo de predicción para  $y_{n+1}$  hay que usar la expresión (14).

Vemos que el predictor o valor predicho es 16.961 y el intervalo de predicción al 95% es

$$[15.753 ; 18.169].$$

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

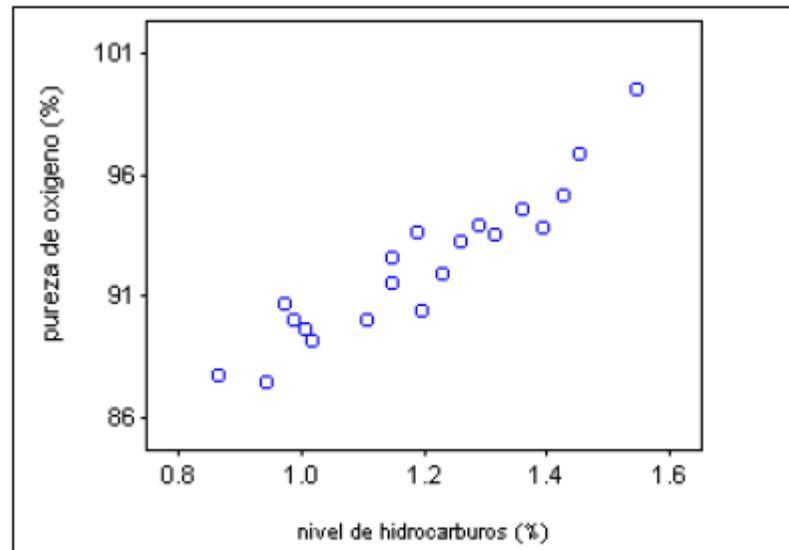
Gráficamente los dos intervalos quedarían así:



Aquí mostramos los resultados en otro ejemplo:

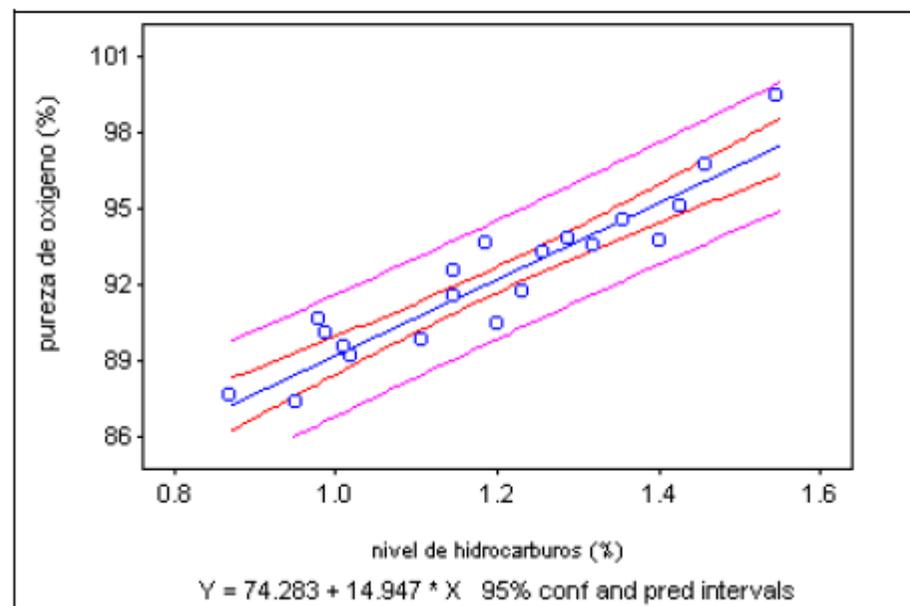
Interesa estudiar la relación entre la pureza de oxígeno (y) producido en un proceso de destilación y el porcentaje de hidrocarburos (x) presentes en el condensador principal de un destilador. Los datos se muestran en la tabla y scatter plot siguientes:

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%



```
destilacion=read.table("C:\\Users\\Ana\\estadisticaQ\\2012\\destilacion.txt",header=T)
attach(destilacion)
salida<- lm(oxigeno~hidrocarburos)
summary(salida)
```

```
Call:
lm(formula = oxigeno ~ hidrocarburos)
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-1.83029	-0.73334	0.04497	0.69969	1.96809

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	74.283	1.593	46.62	< 2e-16 ***
hidrocarburos	14.947	1.317	11.35	1.23e-09 ***

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 1.087 on 18 degrees of freedom
```

```
Multiple R-squared: 0.8774, Adjusted R-squared: 0.8706
```

```
F-statistic: 128.9 on 1 and 18 DF, p-value: 1.227e-09
```

## Intervalos de Confianza

```
predict(salida,interval="confidence",level=0.95)
```

	fit	lwr	upr
1	89.08132	88.31627	89.84637
2	89.52974	88.82434	90.23515
3	91.47292	90.94686	91.99897
4	93.56556	92.99271	94.13842
5	96.10663	95.21561	96.99766
6	94.61189	93.92897	95.29480
7	87.28762	86.25134	88.32390
8	92.66871	92.14969	93.18774
9	97.45191	96.34756	98.55626
10	95.20979	94.44885	95.97072
11	92.07082	91.56012	92.58152
12	91.47292	90.94686	91.99897
13	88.93184	88.14597	89.71772
14	89.38027	88.65549	90.10505
15	90.87502	90.31186	91.43817
16	92.22029	91.70974	92.73084
17	93.11714	92.57687	93.65740
18	94.01399	93.39900	94.62898
19	95.65821	94.83384	96.48258
20	88.48342	87.63273	89.33411

## Intervalos de Predicción

```
predict(salida,interval="prediction",level=0.95)
```

	fit	lwr	upr
1	89.08132	86.67381	91.48882
2	89.52974	87.14052	91.91896
3	91.47292	89.13037	93.81546
4	93.56556	91.21207	95.91906
5	96.10663	93.65619	98.55708
6	94.61189	92.22921	96.99456
7	87.28762	84.78070	89.79454
8	92.66871	90.32774	95.00969
9	97.45191	94.91609	99.98772
10	95.20979	92.80358	97.61599
11	92.07082	89.73167	94.40996
12	91.47292	89.13037	93.81546
13	88.93184	86.51764	91.34605
14	89.38027	86.98526	91.77528
15	90.87502	88.52386	93.22617
16	92.22029	89.88118	94.55940
17	93.11714	90.77136	95.46291
18	94.01399	91.64988	96.37809
19	95.65821	93.23120	98.08522
20	88.48342	86.04735	90.91949

## 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 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  $e_1, e_2, \dots, e_n$ .

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:

$$\hat{x} = \frac{y - \hat{\alpha}}{\hat{\beta}} \tag{15}$$

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

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

Llamando

$$ES(\hat{x}) = \sqrt{\hat{\text{Var}}(\hat{x})} \quad (17)$$

el intervalo

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

es un intervalo de confianza con nivel aproximado  $1-\alpha$  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:

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

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

Quedando (17) y (18) sin cambios.

**Ejemplo:** 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.

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 (16):

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

Luego

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

Aplicando (18) obtenemos que

$$6.21 \pm 2.57 \cdot 0.240$$
$$6.21 \pm 0.62$$

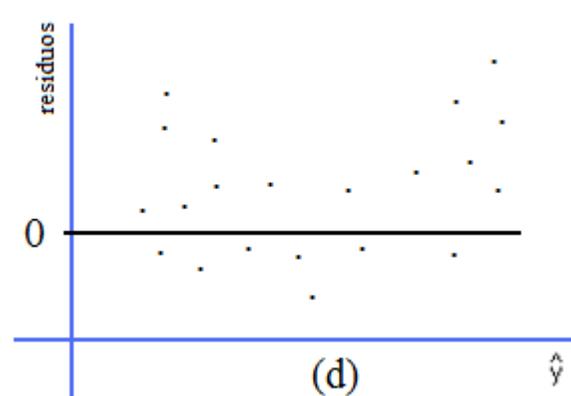
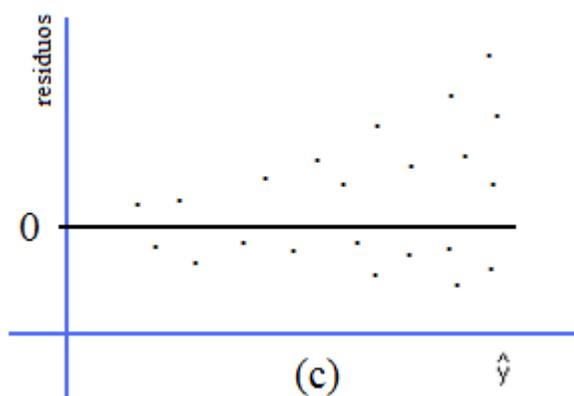
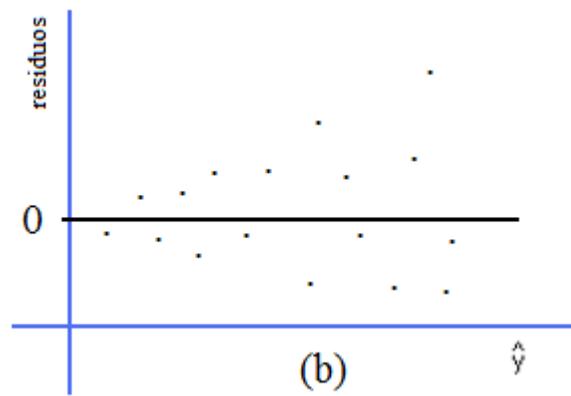
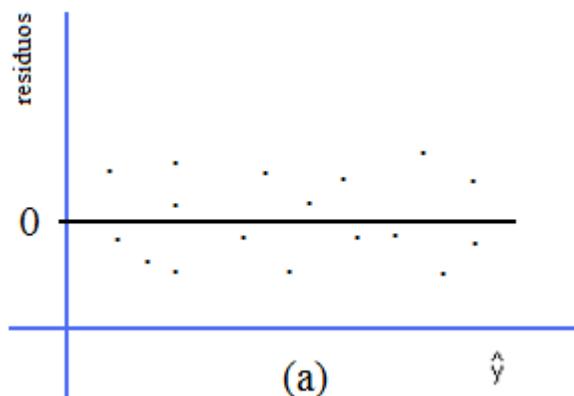
son los límites de confianza al 95% para la concentración de fluoresceína en la nueva muestra observada.

¿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 homoscedasticidad. 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.

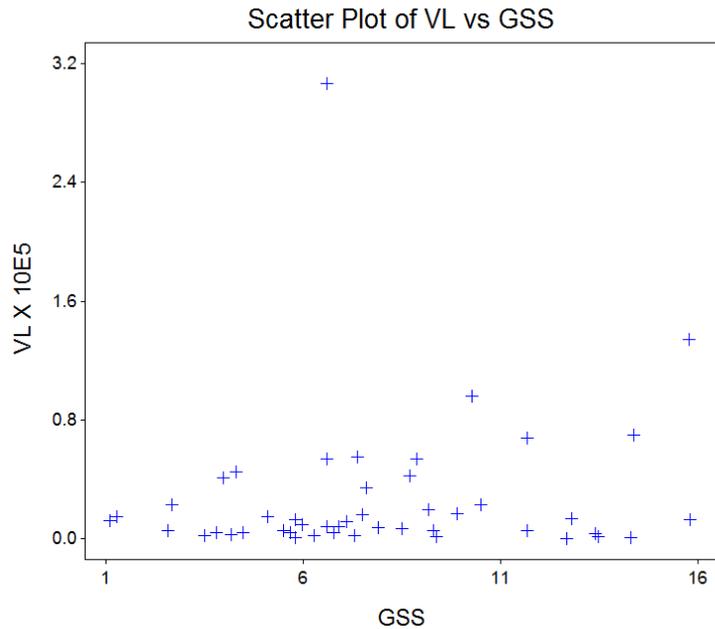


(a)  
Representa la situación esperable si el modelo se cumple: una nube de residuos alrededor del 0 sin estructura.

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

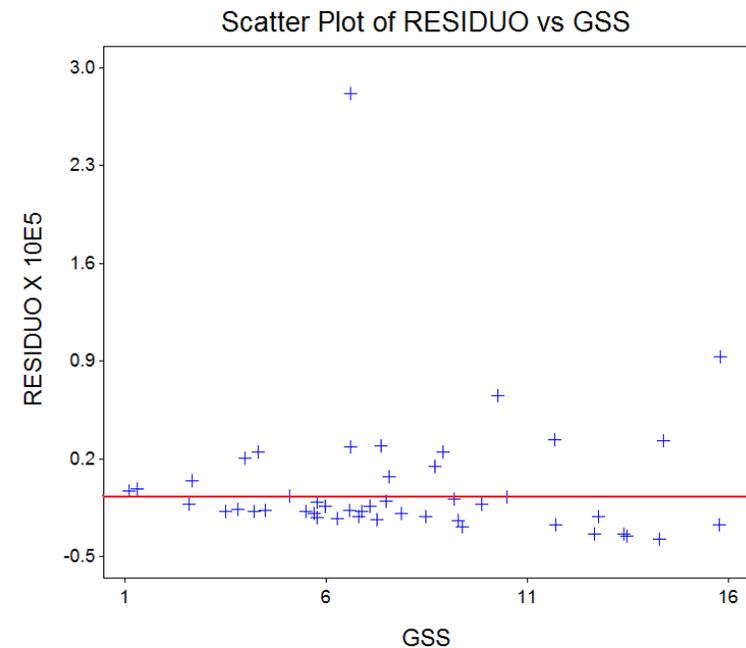
(d) El supuesto de linealidad no se satisface.

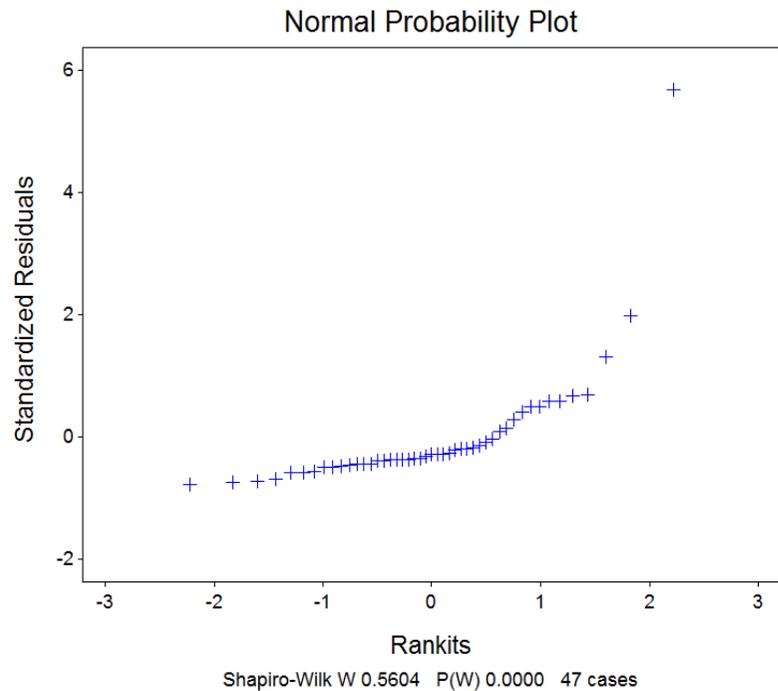
**Ejemplo 1:** Consideramos los datos correspondientes a 47 pacientes de HIV para los que se registraron las variables **VL**: carga viral y **GSS** que corresponde a un score genético del paciente.



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

En este gráfico hemos usado “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.

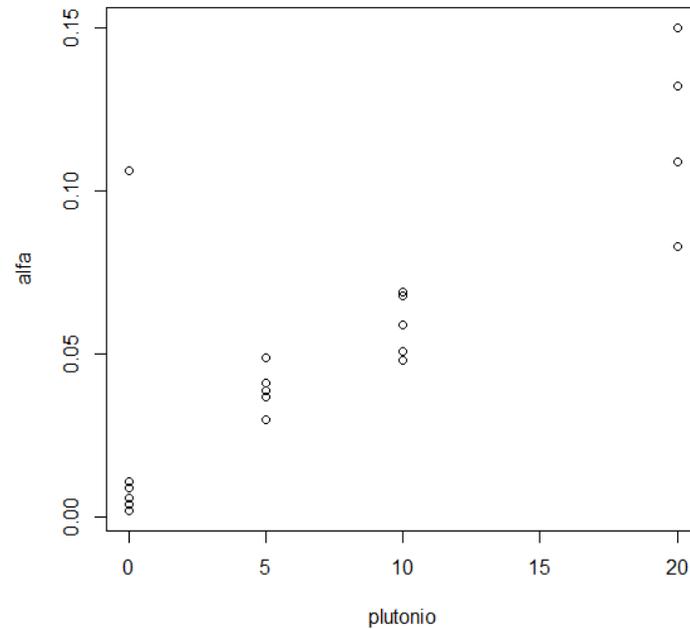
Teniendo en cuenta la varianza de los residuos definimos los residuos standarizados como:

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

**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 dispersión:



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”.

```
radiacion=read.table("C:\\Users\\Ana\\estadisticaQ\\2012\\radiacion.txt",header=T)
attach(radiacion)
salida<- lm(alfa~plutonio)
summary(salida)
```

```
Call:
lm(formula = alfa ~ plutonio)
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-0.031826	-0.010529	-0.005603	0.001878	0.091471

```
Coefficients:
```

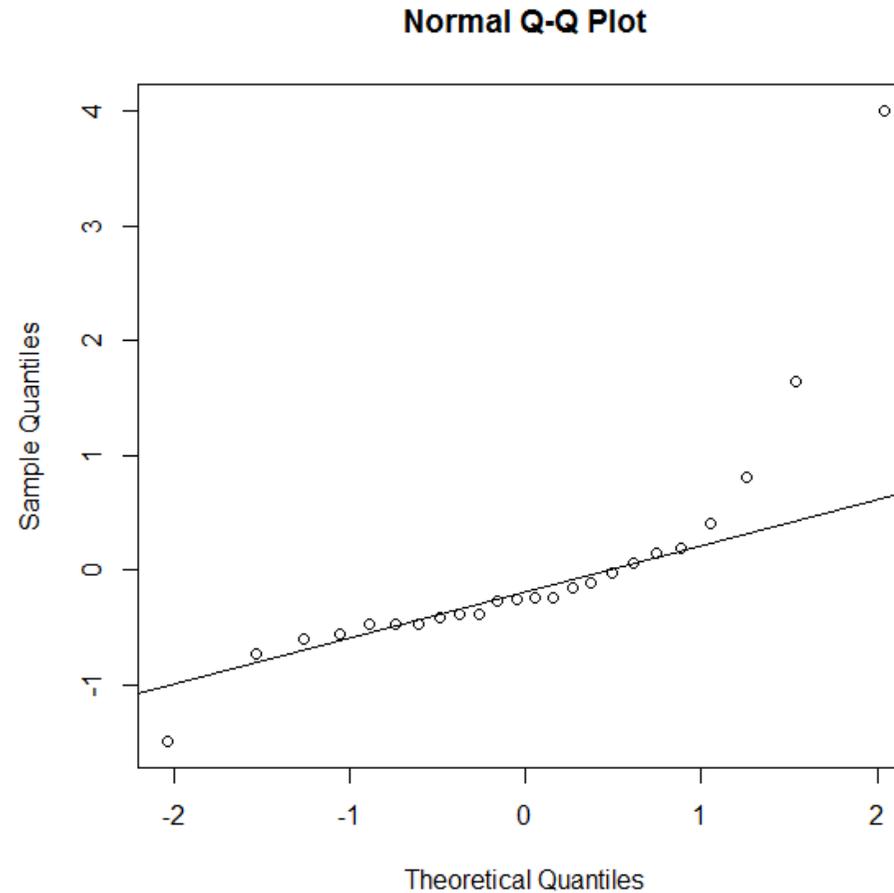
	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	0.0145294	0.0065264	2.226	0.0366	*
plutonio	0.0050148	0.0006778	7.398	2.11e-07	***

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.02371 on 22 degrees of freedom
Multiple R-squared: 0.7133, Adjusted R-squared: 0.7003
F-statistic: 54.74 on 1 and 22 DF, p-value: 2.107e-07
```

```
ls.diag(salida)
res.std<- ls.diag(salida)$std.res
qqnorm(res.std)
qqline(res.std)
```



```
shapiro.test(res.std)
```

Shapiro-Wilk normality test

```
data: res.std
W = 0.691, p-value = 7.666e-06
```

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, resulta:

```
alfa.sout<- alfa[-10]  Sacamos la observación 10!!
plutonio.sout<- plutonio[-10]
salida.sout<- lm(alfa.sout~plutonio.sout)
summary(salida.sout)
```

Call:

```
lm(formula = alfa.sout ~ plutonio.sout)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.034773	-0.004061	-0.001033	0.004939	0.032227

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.0070331	0.0035988	1.954	0.0641 .
plutonio.sout	0.0055370	0.0003659	15.133	9.08e-13 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01257 on 21 degrees of freedom

Multiple R-squared: 0.916, Adjusted R-squared: 0.912

F-statistic: 229 on 1 and 21 DF, p-value: 9.077e-13

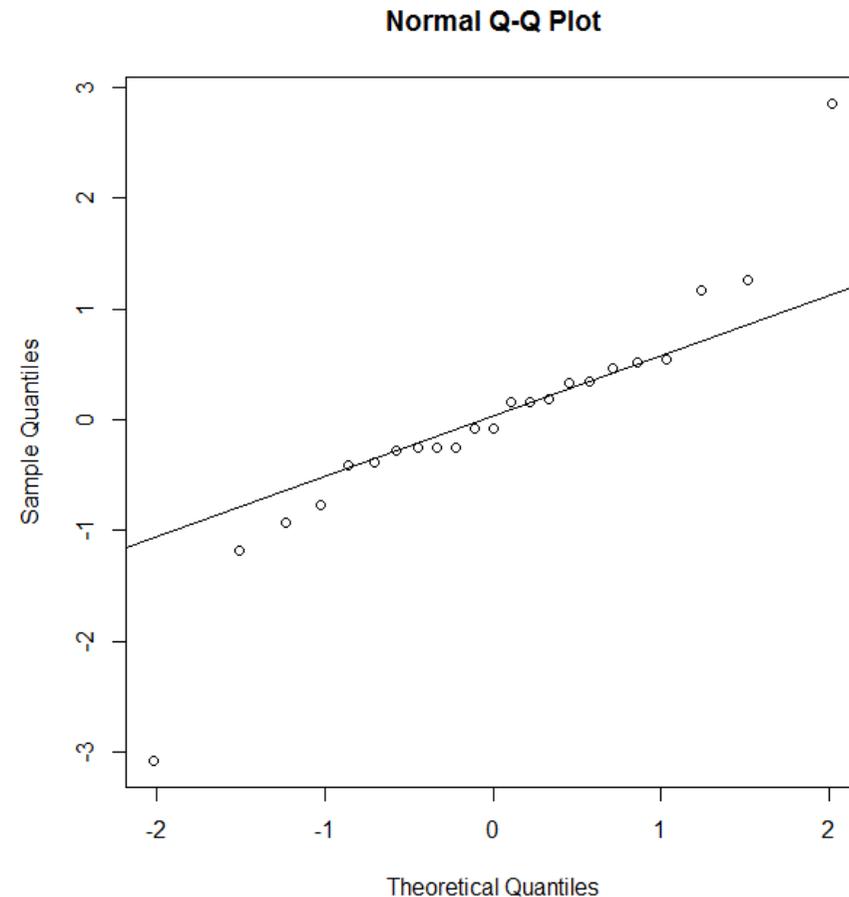
Notemos las diferencias con respecto al análisis anterior en la estimación de la intercept y de su significación, así como las diferencias entre los R-squared.

Analicemos estos residuos.

```
res.std<- ls.diag(salida.sout)$std.res  
qqnorm(res.std)  
qqline(res.std)  
shapiro.test(res.std)
```

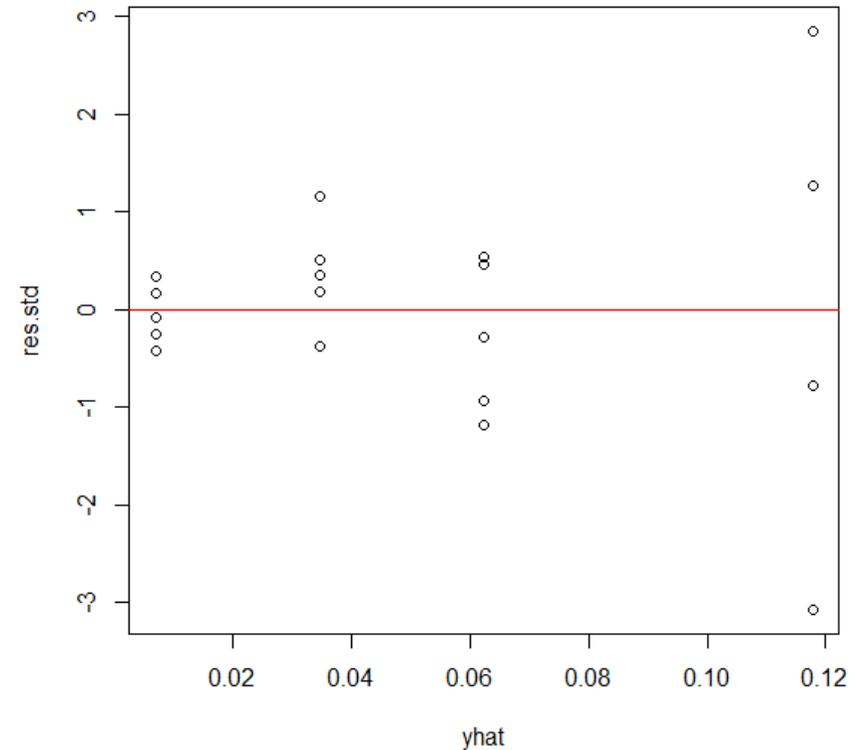
Shapiro-Wilk normality test

```
data: res.std  
W = 0.8985, p-value = 0.02355
```



```
yhat<- salida.sout$fit
plot(yhat,res.std)
abline(h=0,col="red")
```

Todavía persiste un efecto abanico en el gráfico.



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

$$y_i = \alpha + \beta x_i + e_i \quad (\text{para } i=1, \dots, n) \quad (2)$$

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 homocedasticidad, es decir que los errores tienen igual varianza, sino que la varianza cambia con  $x$ , digamos en general que es de la forma:

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

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 desviación standard son proporcionales a  $x$ , o sea

$$\text{a) } \text{Var}(e_i) = \text{cte. } x_i$$

$$\text{b) } \text{Var}(e_i) = \text{cte. } x_i^2$$

Tanto a) como b) pueden escribirse como:

$$\text{Var}(e_i) = \theta v_i \quad (19)$$

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

Supongamos 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 (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

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

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 (20), que equivale a minimizar

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

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.

Muchos programas permiten calcular los estimadores de cuadrados mínimos pesados.

### **Observación:**

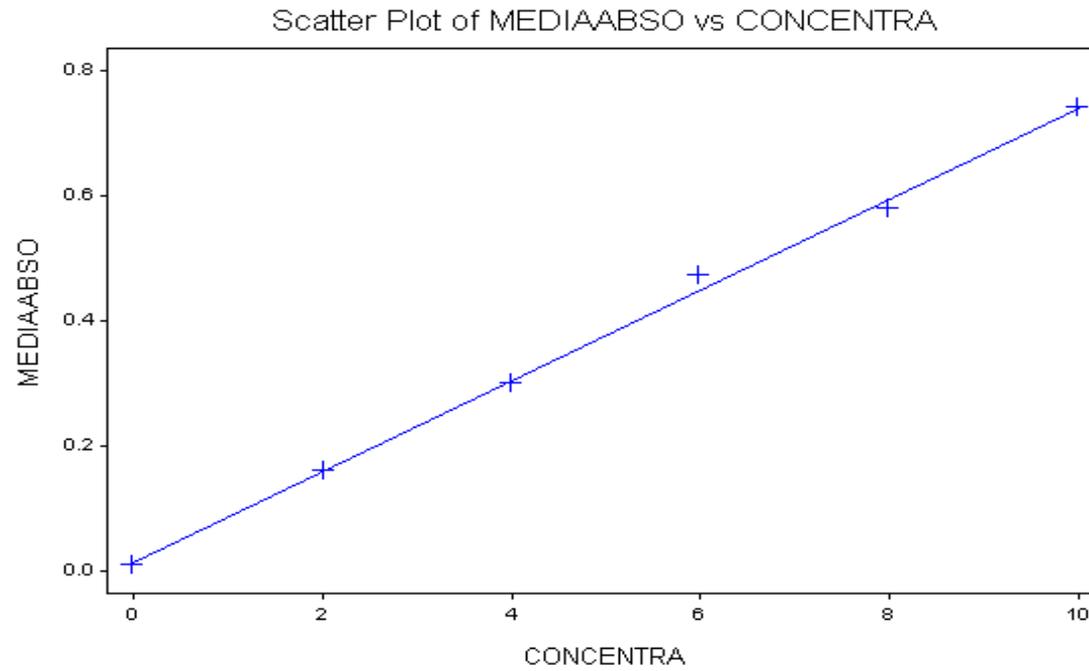
Cuando para cada valor de  $x$  se hacen **varias observaciones** de  $y$ , se pueden estimar las varianzas de los errores por las varianzas muestrales en lugar de hacer suposiciones como en (19).

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:

<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 se ve que a medida que la verdadera concentración aumenta, crece el desvío standard.

Por lo tanto, 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 además la función  $f$  parece ser 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. Calculamos los estimadores de cuadrados mínimos ponderados. Ingresando los datos, calculamos los pesos y obtenemos la siguiente salida:

```
summary(lm(absorbancia~concentra,weights=(1/variacion)))
```

Call:

```
lm(formula = absorbancia ~ concentra, weights = (1/variacion))
```

Residuals:

```
      1      2      3      4      5      6  
-0.08391  0.34904 -0.31238  1.56587 -1.30374 -0.34925
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)  
(Intercept) 0.009084    0.001048   8.671 0.000974 ***  
concentra    0.073760    0.001064  69.330 2.59e-07 ***
```

---

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 1.061 on 4 degrees of freedom

Multiple R-squared: 0.9992, Adjusted R-squared: 0.999

F-statistic: 4807 on 1 and 4 DF, p-value: 2.593e-07

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

```
summary(lm(absorbancia~concentra))
```

Call:

```
lm(formula = absorbancia ~ concentra)
```

Residuals:

```
      1      2      3      4      5      6  
-0.0042857 -0.0003714 -0.0024571  0.0234571 -0.0166286  0.0002857
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)  
(Intercept)  0.013286   0.010559   1.258    0.277  
concentra    0.072543   0.001744  41.602   2e-06 ***
```

---

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 0.01459 on 4 degrees of freedom

Multiple R-squared: 0.9977, Adjusted R-squared: 0.9971

F-statistic: 1731 on 1 and 4 DF, p-value: 1.995e-06

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.