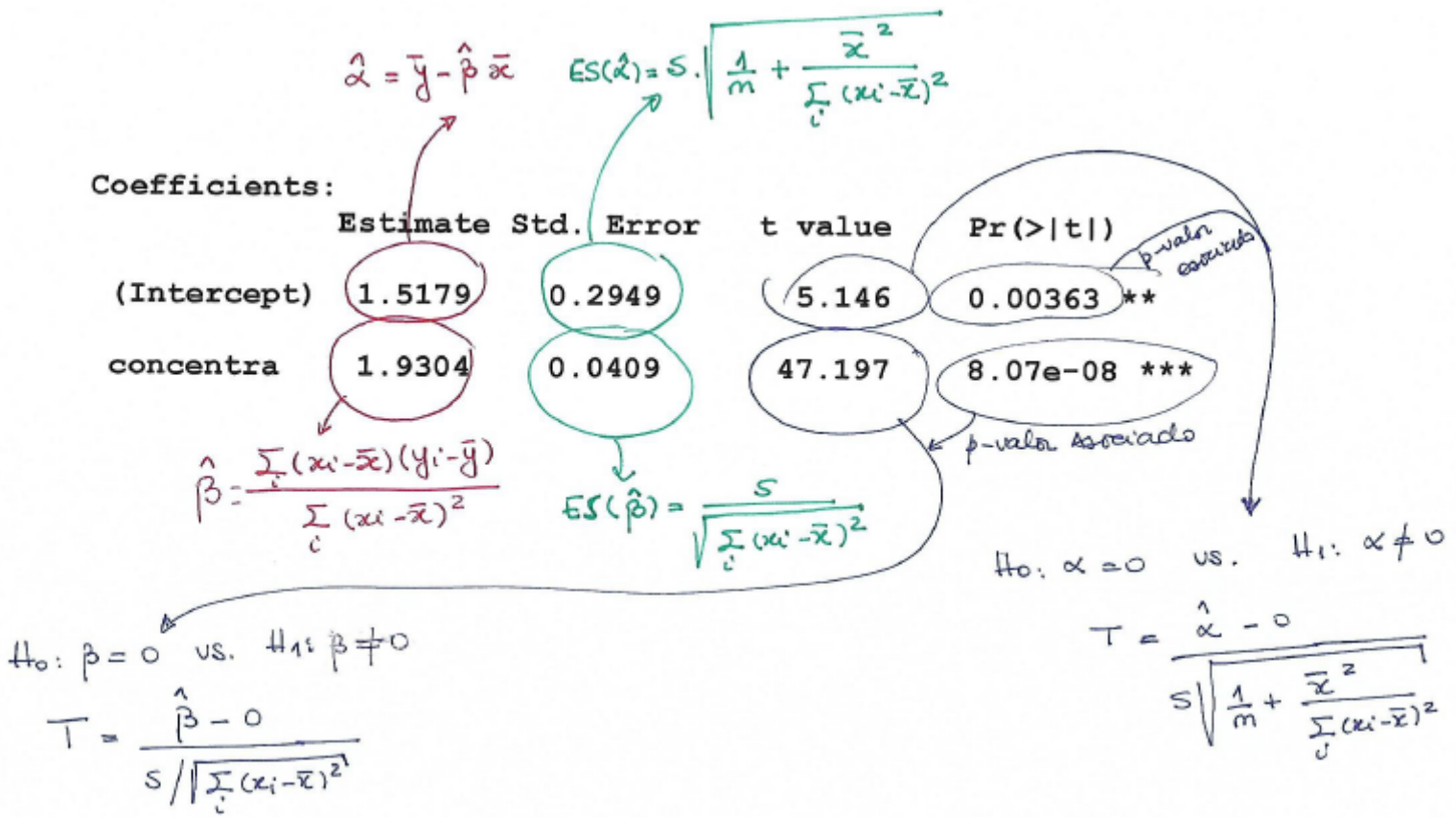
```
concentra=c(0,2,4,6,8,10,12)
fluo<- c(2.1,5,9,12.6,17.3,21,24.7)
salida<- lm(fluo~concentra)
summary(salida)

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
```


$\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x}$
$ES(\hat{\alpha}) = S \cdot \sqrt{\frac{1}{m} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}}$
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 ***
p-valor asociado
p-valor asociado
$\hat{\beta} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$
$ES(\hat{\beta}) = \frac{S}{\sqrt{\sum_i (x_i - \bar{x})^2}}$
$H_0: \beta = 0$ vs. $H_1: \beta \neq 0$
$T = \frac{\hat{\beta} - 0}{S / \sqrt{\sum_i (x_i - \bar{x})^2}}$
$H_0: \alpha = 0$ vs. $H_1: \alpha \neq 0$
$T = \frac{\hat{\alpha} - 0}{S\sqrt{\frac{1}{m} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}}}$

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?

Asumimos que en x_0 se cumplen las condiciones del modelo. 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:

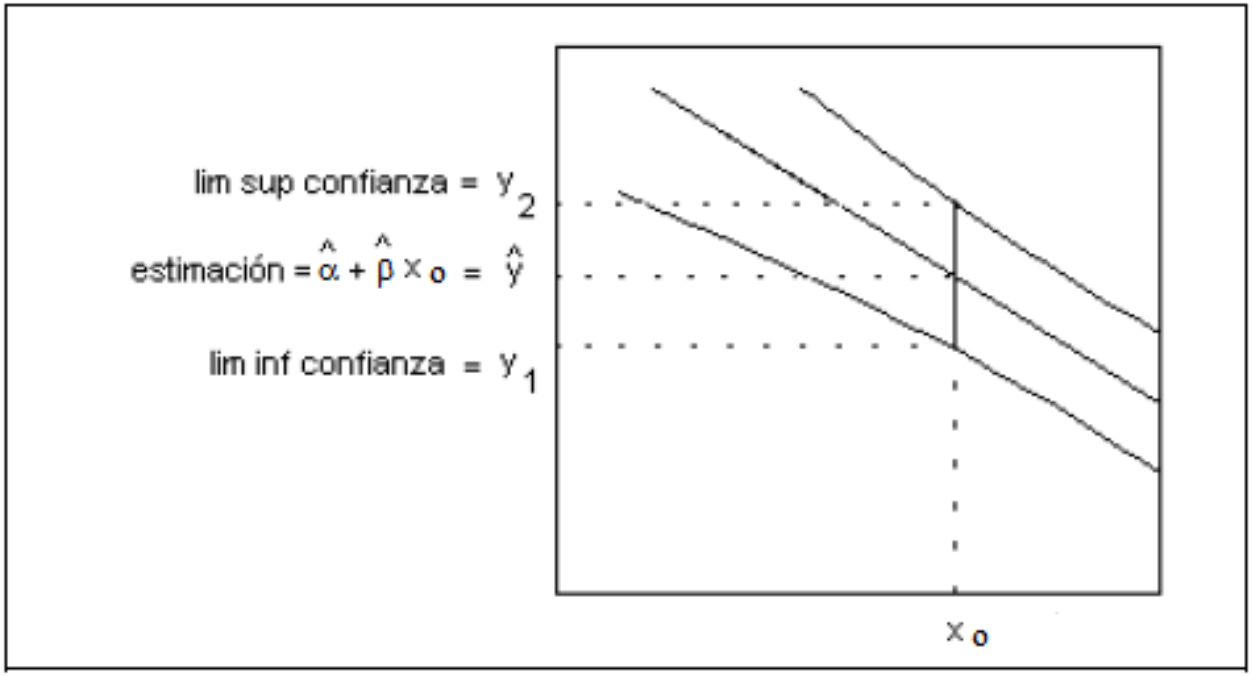
$$\mathrm{Var}(\hat{\alpha} + \hat{\beta}\, x_0) \quad = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] \qquad (11)$$

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} , es decir queremos "predecir" y_{n+1} y dar un intervalo que contenga a y_{n+1} con una probabilidad 0.95 (o 1-α) (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_{n+1} 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:

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

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

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. 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:

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

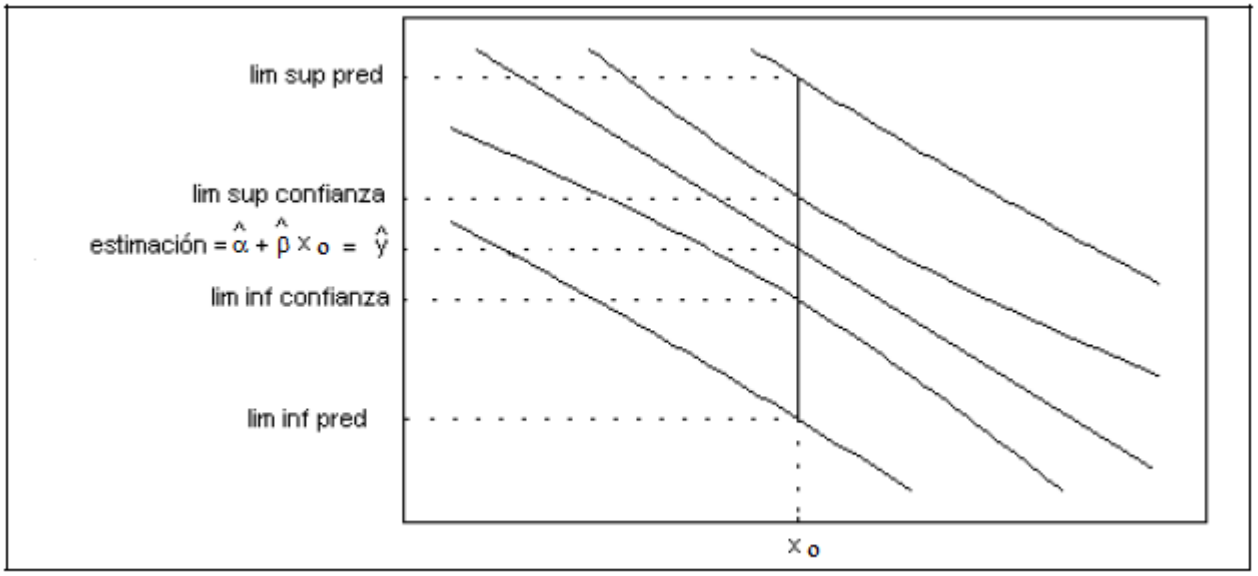
Para obtener el intervalo de predicción para y_{n+1} hay que usar la expresión (14).

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?*

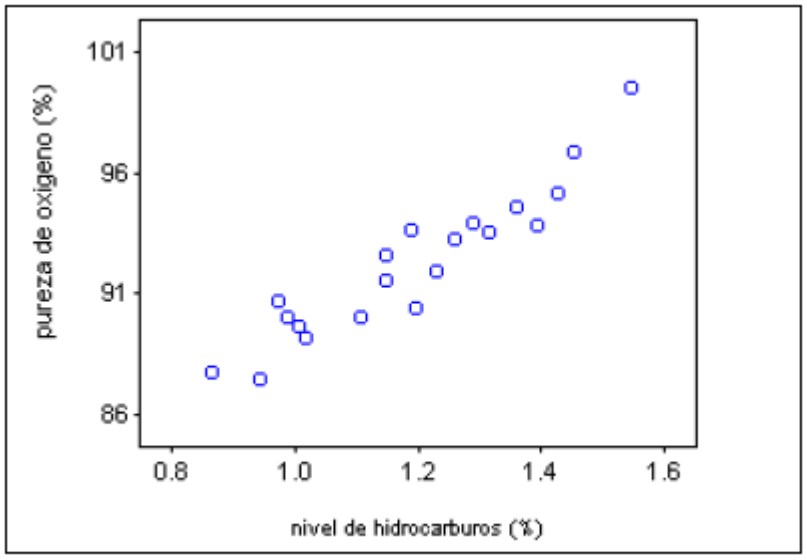
Gráficamente los dos intervalos quedarían así:



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:

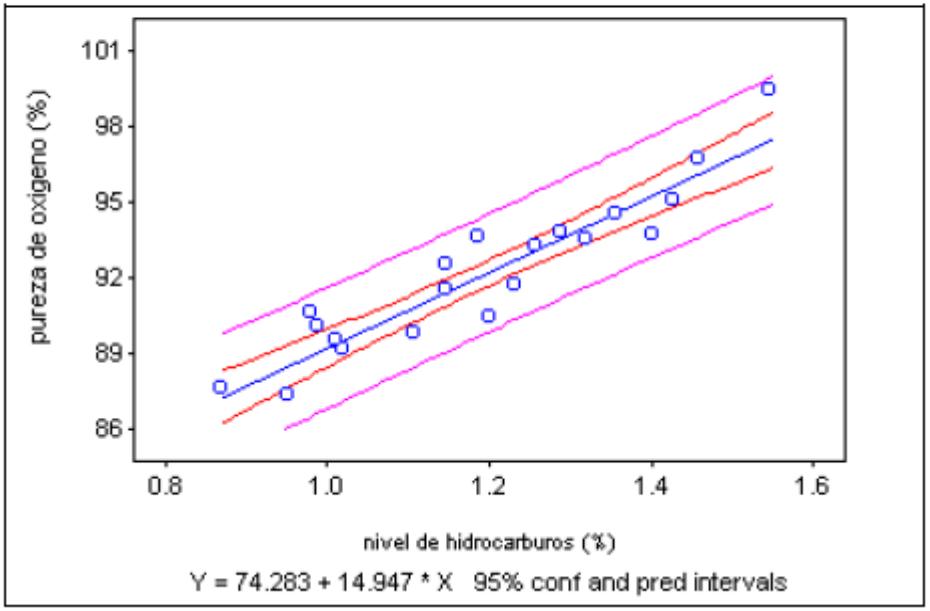
x(%)	y(%)	x(%)	y(%)	x(%)	y(%)	x(%)	y(%)
0.99	90.01	1.36	94.45	1.19	93.54	1.2	90.39
1.02	89.05	0.87	87.59	1.15	92.52	1.26	93.25
1.15	91.43	1.23	91.77	0.98	90.56	1.32	93.41
1.29	93.74	1.55	99.42	1.01	89.54	1.43	94.98
1.46	96.73	1.4	93.65	1.11	89.85	0.95	87.33



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,...,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

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

Llamando

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

el intervalo

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

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: 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	**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?