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?