17.3 Intervalos de predicción para el promedio de m observaciones futuras

Para reducir la incerteza de las predicciones no alcanza con aumentar indefinidamente el tamaño de la muestra en la que se basa el ajuste. Sin embargo es posible reducir la longitud del intervalo de predicción del *promedio de m* observaciones nuevas ($\bar{y}^{nuevas} = \sum_{i=1}^{m} y_i^{nueva} \Big/ m$) cuyo intervalo de predicción del (1-α)100% tendrá la siguiente expresión general para el valor x^{nueva} de la variable explicativa:

$$a + b\, x^{\text{nueva}} \pm t_{n-2,\alpha/2}\, \sigma \sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(x^{\text{nueva}} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}, \qquad (23)$$

17.4 Intervalos simultáneos

Muchas veces interesa construir intervalos de confianza o de predicción para varios valores nuevos de la variable x (para varios niveles de x) simultáneamente. Esto ocurre, por ejemplo, cuando una misma recta ajustada es utilizada varias veces para estimar la media de la variable respuesta o para predecir valores futuros para distintos valores de x.

Aunque cada uno de los intervalos de confianza construidos, utilizando la expresión (21), tenga nivel de confianza (1-α)100%, no se puede garantizar ese nivel global. Esto es similar al problema de obtener intervalos de confianza simultáneos en el análisis de la varianza.

17.4.1 K intervalos de confianza simultáneos para varias respuestas medias

Presentaremos dos tipos de intervalos que se obtienen modificando levemente los intervalos dados por (21) de manera que se puede asegurar un nivel global 1- α para el cual todos los intervalos son correctos.

Procedimiento de Bonferroni. Si interesan construir K intervalos simultáneos los limites de confianza están dados por:

$$a + b\, x_k^{\text{nueva}} \pm t_{n-2,\,\alpha/K2}\, \sigma \sqrt{\frac{1}{n} + \frac{(x_k^{\text{nueva}} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \qquad (24)$$

$$a + b\, x_k^{\text{nueva}} \pm t_{n-2,\,\alpha/K2}\, s(a + b x_k^{\text{nueva}}) \qquad 1 \le k \le K$$

Esta última expresión es similar a la de (21) salvo en que $t_{n-2,\,\alpha/2}$ se ha cambiado por $t_{n-2,\,\alpha\,/\,K2}$ para obtener un nivel global de por lo menos 1-α, los grados de libertad no cambian porque provienen de $\hat{\sigma}$ = RSS/(n-2).

Procedimiento de Hotelling-Scheffé. Este procedimiento está basado en una banda de confianza para toda la recta de regresión, de manera que podemos utilizar los límites de confianza dados por esta banda para todos los x's de interés y el nivel de confianza global será por lo menos (1-α) 100%

$$a + b\,x^{\text{nueva}} \pm \sqrt{2f_{2,n-2,\alpha}}\;\hat{\sigma}\;\sqrt{\frac{1}{n} + \frac{(x^{\text{nueva}} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \qquad (25)$$

$$a + b\,x^{\text{nueva}} \pm \sqrt{2f_{2,n-2,\alpha}}\;s(a + b\,x^{\text{nueva}})$$

El valor crítico ($f_{2,n-2,\alpha}$) para hallar el intervalo de confianza de nivel aproximado (1-α) 100% , corresponde a una distribución F con 2 y n-2 grados de libertad.

¿Cuál de las dos familias de intervalos deberíamos elegir? Debemos comparar $t_{n-2,\alpha/K2}$ y $\sqrt{2f_{2,n-2,\alpha}}$. El intervalo de confianza más preciso será el que provenga del menor de estos valores. En general cuando la cantidad de intervalos simultáneos que interesan es pequeña el procedimiento de Bonferroni será el mejor. En el caso que interesen muchos intervalos el procedimiento de Hotelling-Scheffé podrá dar intervalos de menor longitud pues el valor crítico no depende de la cantidad de intervalos que interese construir simultáneamente.

17.4.2 K intervalos de predicción simultáneos, de nivel global aproximado 1-α, para nuevas observaciones

Los **límites de predicción de Bonferroni** para *K observaciones* futuras de Y obtenidas en *K* niveles diferentes de las x´s son

$$a + b\,x_k^{\text{nueva}} \pm t_{n-2,\alpha/2K}\;\hat{\sigma}\;\sqrt{1 + \frac{1}{n} + \frac{(x_k^{\text{nueva}} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \qquad (26)$$

$$a + b\,x_k^{\text{nueva}} \pm t_{n-2,\alpha/2K}\;s\,(\text{pred } Y) \qquad 1 \le k \le K$$

Intervalos de Hotelling-Scheffé para K observaciones futuras de Y obtenidas en K niveles diferentes de las x´s

$$a + b\, x_k^{nueva} \pm \sqrt{Kf_{k,n-2,\alpha}}\ \sigma \sqrt{1 + \frac{1}{n} + \frac{(x_k^{nueva} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \qquad (27)$$

$$a + b\, x_k^{nueva} \pm \sqrt{Kf_{k,n-2,\alpha}}\ s(\text{pred } Y) \qquad 1 \le k \le K$$

Comentarios

- Los intervalos de predicción simultáneos de nivel 1-α para K observaciones nuevas de Y para K niveles diferentes de x tienen mayor longitud que los correspondientes a una única predicción. Cuando la cantidad de predicciones no es grande esta diferencia es moderada.
- Para los dos procedimientos *la longitud de los intervalos de predicción aumenta a medida que* K *aumenta*. Cuando construimos intervalos de confianza simultáneos la longitud de los intervalos de Hotelling-Scheffé no aumentaba.

17.5 Predicción inversa. Problema de Calibración

En algunas ocasiones, un modelo de regresión de Y sobre X es utilizado para realizar estimaciones del valor de X que dio lugar al nuevo valor observado de Y. Este procedimiento es llamado predicción inversa.

Ejemplo: se desarrolla un método rápido y económico para medir la concentración de azúcar (galactosa) en sangre. Supongamos que las mediciones de la concentración de galactosa se relacionan linealmente con la concentración verdadera (obtenida mediante un método preciso y exacto, costoso y lento). Esto es que se satisface el modelo

"concentración medida" = α + β "concentración verdadera" + error

es decir:

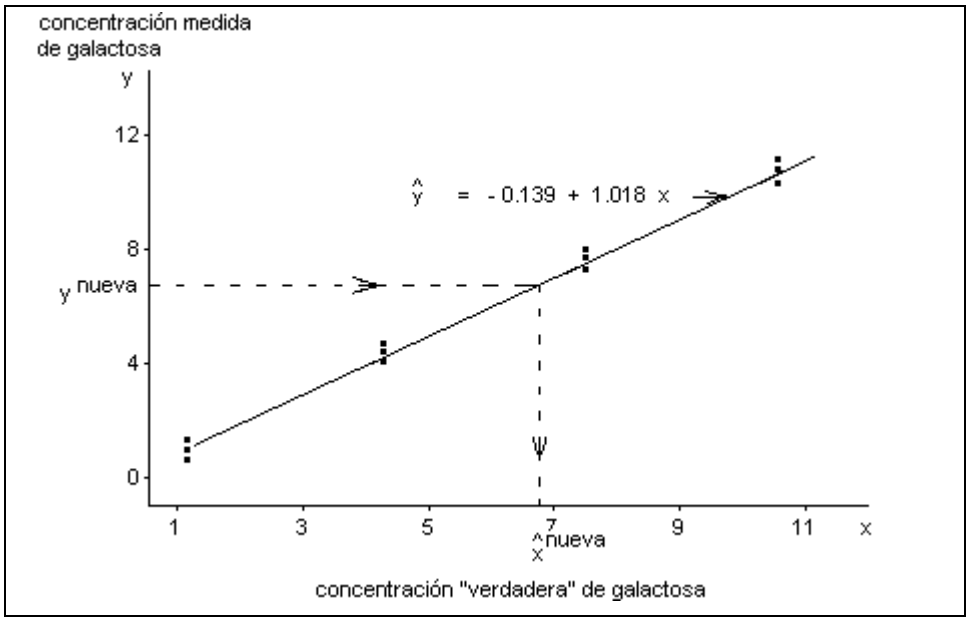
$$Y_i = \alpha + \beta\, x_i + \varepsilon_i$$

Con el objetivo de calibrar el método rápido tomaremos varias muestras con concentraciones x_i "verdaderas" (determinadas con el método preciso y exacto) y obtendremos los valores de Y_i mediante el método rápido. Utilizaremos los datos para ajustar una recta:

$$\hat{y} = a + bx \qquad (28)$$

por el método de cuadrados mínimos.

Figura 22. Diagrama de dispersión y recta ajustada a las concentraciones medidas de 12 muestras que tienen concentración conocida de galactosa (X), tres muestras para cada uno de cuatro niveles diferentes.



Supongamos que se tiene una nueva observación y^{nueva} y se quiere estimar el nivel x^{nueva} que dio lugar a dicha observación, es natural obtener un estimador de x^{nueva} despejando de (28)

$$\hat{x}^{nueva} = \frac{y^{nueva} - a}{b} \qquad b \neq 0 \tag{29}$$

17.5.1 Límites de estimación resultantes de una predicción inversa

La expresión general de los límites de estimación *aproximados* del (1-α) 100 % para $\hat{x}^{nueva}$ basado en una observación y^{nueva} es

$$\hat{x}^{nueva} \pm t_{n-2,\alpha/2}\, s(\text{pred}\,\hat{x}^{nueva}) \tag{30}$$

donde $s(\text{pred}\,\hat{x}^{nueva}) = \frac{\hat{\sigma}}{b}\sqrt{1+\frac{1}{n}+\frac{(\hat{x}^{nueva}-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$ (30 a)

$$= \frac{\hat{\sigma}}{b}\sqrt{1+\frac{1}{n}+\frac{(y^{nueva}-\bar{y})^2}{b^2\sum_{i=1}^{n}(x_i-\bar{x})^2}} \tag{30 b}$$

Volviendo al ejemplo

Tabla 4. Medidas resumen de los datos de galactosa

DESCRIPTIVE STATISTICS

VARIABLE	N	MEAN	SD	VARIANCE
X	12	**5.9000**	3.6668	**13.445** = $\sum_{i=1}^{12}(x_i-\bar{x})^2/11$
Y	12	5.8696	3.7409	13.994

Tabla 5 Resultados del ajuste por cuadrados mínimos a los datos de galactosa

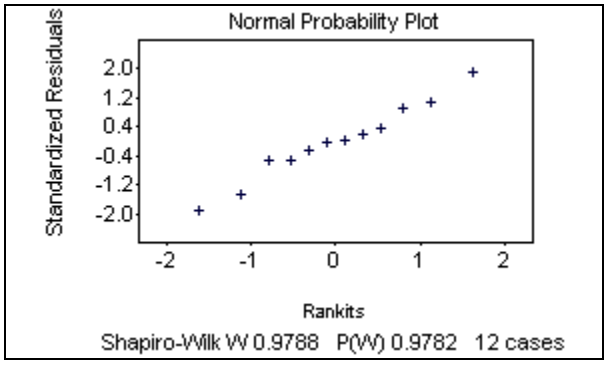
PREDICTOR VARIABLES	COEFFICIENT	STD ERROR	STUDENT'S T	P
CONSTANT	**-0.13866 (a)**	0.13344	-1.04	0.3232
X	**1.01835 (b)**	0.01944	52.40	0.0000

R-SQUARED	0.9964	RESID. MEAN SQUARE (MSE)	0.05587
ADJUSTED R-SQUARED	0.9960	STANDARD DEVIATION	**0.23637** ⇐ $\hat{\sigma}$

SOURCE	DF	SS	MS	F	P
REGRESSION	1	153.376	153.376	2745.29	0.0000
RESIDUAL	10	0.55869	0.05587		
TOTAL	11	153.935			

CASES INCLUDED 12 MISSING CASES 0

Vemos que la pendiente estimada (a) 1.01835 es estadísticamente significativa distinta de cero y que R^2= 0.9964 es decir que la recta ajustada explica el 99.64% de la variabilidad observada en las mediciones de la concentración de galactosa. El coeficiente de correlación es R = 0.9982



Supongamos que interesa utilizar la relación lineal entre la galactosa verdadera y la galactosa medida, para estimar el nivel verdadero de galactosa (x^{nueva}) mediante un intervalo de confianza del 95%, para un paciente para el cual el procedimiento rápido dio 5.82 (y^{nueva}).

Los resultados de las tablas 4 y 5 nos facilitan los cálculos.

$$\hat{x}^{nueva} = \frac{y^{nueva} - a}{b} = \frac{5.82 - (-0.13866)}{1.01835} = 5.85$$

La varianza muestral de los valores x_i's iniciales que determinan la recta ajustada es

$13.445 = \sum_{i=1}^{12}(x_i - \bar{x})^2 / 11 \quad \Rightarrow \quad \sum_{i=1}^{12}(x_i - \bar{x})^2 =$ 13.445 * 11 = 147.895 y $\bar{x}$=5.90

Luego de (30 a) tenemos

$$s(\text{pred } \hat{x}^{\text{nueva}}) = \frac{\hat{\sigma}}{b}\sqrt{1 + \frac{1}{n} + \frac{(\hat{x}^{\text{nueva}} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \frac{0.23637}{1.01835}\sqrt{1 + \frac{1}{12} + \frac{(5.85 - 5.90)^2}{147.895}}$$

$$= 0.2416$$

$t_{n-2,\, \alpha/2} = t_{10,\, 0.025} = 2.23$

Los límites de confianza del 95% resultan de aplicar la expresión (30)

$5.85 \pm 2.23 * 0.2416 = 5.85 \pm 0.54$

Por lo tanto el intervalo de confianza del 95% para la "verdadera" concentración de galactosa es (5.31; 6.39).

17.5.2 Límites de estimación, tomando el promedio de mediciones repetidas

Con el objetivo de reducir la longitud del intervalo de confianza para **una** x^{nueva} es recomendable tomar *m* observaciones nuevas $y_1^{nueva}, \cdots, y_m^{nueva}$ para un mismo valor desconocido x^{nueva} y estimarlo utilizando $\bar{y}_m^{nueva} = \frac{\sum_{i=1}^{m} y_i^{nueva}}{m}$:

$$\hat{x}^{nueva} = \frac{\bar{y}_m^{nueva} - a}{b} \qquad b \neq 0 \qquad (31)$$

La expresión general de los límites de confianza aproximados del (1-α) 100 % para $\hat{x}^{\text{nueva}}$ basados en la media muestral de *m* observaciones nuevas ($\bar{y}_m^{nueva}$) es

$$\hat{x}^{\text{nueva}} \pm t_{n-2,\alpha/2}\, s(\text{pred}\,\hat{x}^{\text{nueva}}) \qquad (32)$$

donde ahora $s(\text{pred}\,\hat{x}^{\text{nueva}}) = \frac{\hat{\sigma}}{b}\sqrt{\frac{1}{m}+\frac{1}{n}+\frac{(\hat{x}^{\text{nueva}}-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$ (32 a)

$$= \frac{\hat{\sigma}}{b}\sqrt{\frac{1}{m}+\frac{1}{n}+\frac{(\bar{y}_m^{nueva}-\bar{y})^2}{b^2\sum_{i=1}^{n}(x_i-\bar{x})^2}} \qquad (32\ b)$$

<u>Volvamos al ejemplo</u>

Supongamos ahora que se han realizado *m* = 10 determinaciones de galactosa, con el método rápido, sobre la misma muestra obteniéndose una media de 5.82 ($\bar{y}_{10}^{nueva}$) obtenemos la misma estimación para la "verdadera" concentración de galactosa,

$$\hat{x}^{nueva} = \frac{\bar{y}_{10}^{nueva} - a}{b} = \frac{5.82-(-0.13866)}{1.01835} = 5.85$$

pero ahora de (32 a)

$$s(\text{pred}\,\hat{x}^{\text{nueva}}) = \frac{\hat{\sigma}}{b}\sqrt{\frac{1}{m}+\frac{1}{n}+\frac{(\hat{x}^{\text{nueva}}-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} = \frac{0.23637}{1.01835}\sqrt{\frac{1}{10}+\frac{1}{12}+\frac{(5.85-5.90)^2}{147.895}} =$$

=0.099

Por lo tanto los límites de confianza con nivel aproximado 95% para la concentración verdadera de galactosa son ahora

$5.85 \pm 2.23*0.099 = 5.85 \pm 0.22$

Por lo tanto el intervalo de confianza para la concentración verdadera es (5.63; 6.07); la longitud se ha reducido a menos de la mitad.

Comentarios

- La predicción inversa, es también conocida como un problema de calibración debido a que se puede aplicar cuando *n* mediciones (y) -económicas, rápidas y aproximadas- son relacionadas con *n* mediciones precisas (x) -habitualmente costosas y que requieren mucho tiempo-. El modelo de regresión ajustado es utilizado para estimar una nueva medición precisa x^{nueva} utilizando una o más mediciones rápidas (y_i^{nueva} $1 \leq i \leq m$).
- Aunque el modelo de regresión que hemos ajustado requiere mediciones de las x's sin error, en la práctica se lo puede utilizar cuando la varianza de las x's es despreciable con respecto a la varianza de las y's.
- Los intervalos de confianza aproximados (dados por (30) y (32)) requieren que la cantidad

$$\frac{(t_{n-2,\alpha/2})^2 \hat{\sigma}^2}{b^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

sea pequeña, digamos, menor a 0.1. Para el ejemplo esta cantidad es

$$\frac{2.33^2 0.24^2}{1.01835^2 147.895} = 0.0020$$

de manera que la aproximación resulta adecuada.

- ***K intervalos simultáneos*** para valores x_k^{nueva} $1 \leq k \leq K$ de ***nivel global*** aproximadamente 1-α basados en K nuevas observaciones y_k^{nueva} $1 \leq k \leq K$, se obtienen reemplazando en (30) el valor crítico $t_{n-2,\alpha/2}$ por $t_{n-2,\alpha/2K}$ o por $(K f_{K,n-2,\alpha})^{1/2}$ (Bonferroni y Scheffé respectivamente).

- ***K intervalos simultáneos*** para valores x_k^{nueva} $1 \leq k \leq K$ de ***nivel global*** aproximadamente 1-α basados en K nuevas medias muestrales $\bar{y}_{k,m}^{nueva} = \sum_{j=1}^{m} y_{kj}^{nueva} / m$ $1 \leq j \leq m$, $1 \leq k \leq K$, se obtienen reemplazando en (32) el valor crítico $t_{n-2,\alpha/2}$ por $t_{n-2,\alpha/2K}$ o por $(K f_{K,n-2,\alpha})^{1/2}$ (Bonferroni y Scheffé respectivamente)

Estas modificaciones son especialmente útiles cuando la misma recta de calibración es utilizada muchas veces.