27 Regresión Lineal Simple

27.1 Introducción

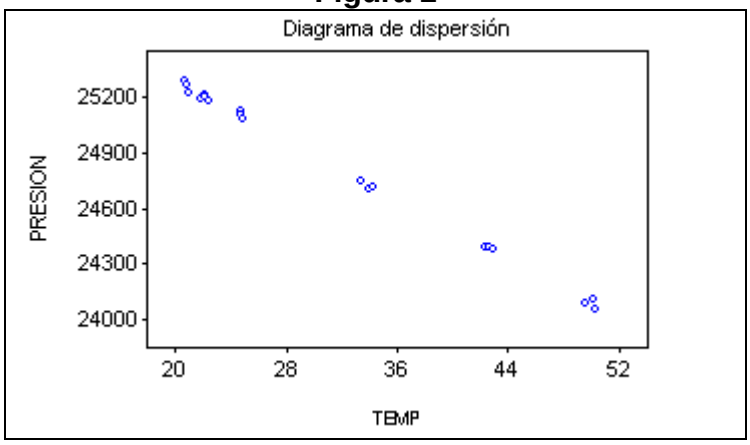
En muchos problemas científicos interesa hallar la relación entre una variable (**Y**), llamada variable de respuesta, ó **variable de salida**, ó variable dependiente y un conjunto de variables (X_1, X_2, ...) llamadas variables explicativas**,** ó variables independientes ó **variables de entrada**.

Consideraremos el caso más simple que consiste en describir la relación entre dos variables continuas mediante una recta. Aún cuando el análisis incluya eventualmente más de una variable explicativa, es habitual mirar inicialmente una variable explicativa por vez.

Ejemplo. Interesa estudiar la relación entre la presión (bar) de transición de Bismuto I-II con la temperatura (^{o} C)

TEMP	PRESION	TEMP	PRESION	TEMP	PRESION	TEMP	PRESION
20.8	25276	22.1	25187	25	25080	42.7	24394
20.9	25256	22.4	25177	33.5	24750	42.9	24384
21	25216	22.5	25177	34	24701	49.7	24077
21.9	25187	24.8	25112	34.2	24716	50.1	24106
22.1	25217	24.8	25093	42.5	24374	50.3	24057

Figura 1



Vemos que la presión de transición de Bismuto I-II, decrece a medida que aumenta la temperatura, observamos una tendencia lineal decreciente aunque los puntos del diagrama de dispersión no están “perfectamente alineados”.

27.2 Puntos sobre una recta

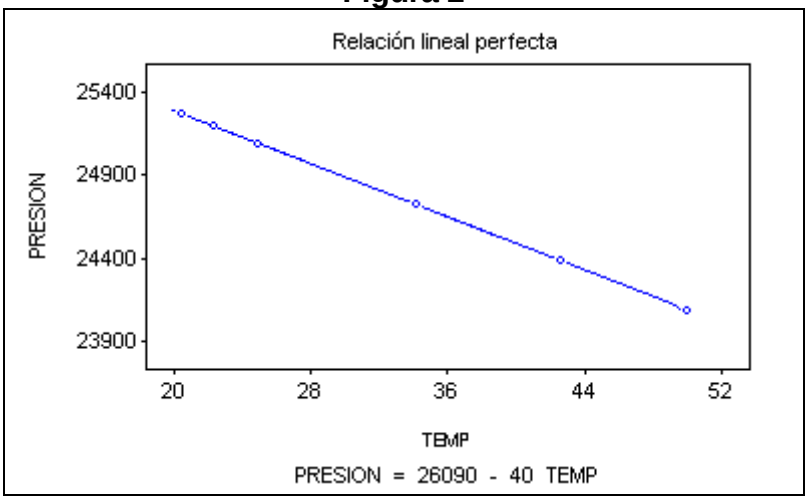
Diremos que la relación entre dos variables *X* e *Y* es “perfectamente lineal”, si todos los pares de valores observados (x_i,y_i) de dichas variables satisfacen la ecuación de una recta:

$$y_i = \alpha + \beta\, x_i \qquad (1)$$

En esta expresión α y β son constantes: α es la ordenada al origen y β la pendiente.

Decimos que *X* es una **variable predictora** de *Y*, ecuación (1). El valor i del subíndice indexa las observaciones: i = 1,2,3,...,n. Para el ejemplo y_1 representa el valor de la presión obtenido para la temperatura x_1 .

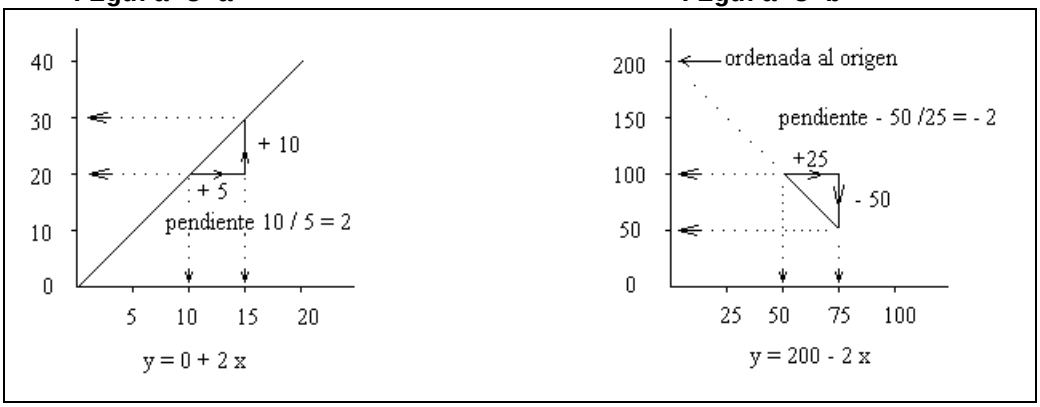
Figura 2



Gráficamente, (1) define una línea recta, donde:

- α (la *ordenada al origen*) es el punto donde la recta corta al eje vertical *y*
- β (la *pendiente*), indica cuantas unidades cambia y cuando x aumenta 1 unidad.
- Si β positivo la recta sube β unidades por cada aumento de *x* en 1 unidad. Si β es negativo la recta cae cuando x aumenta. Si $\beta = 0$ la recta es horizontal.

Figura 3 a **Figura 3 b**



La figura 3 muestra dos ejemplos hipotéticos.

Si la relación entre X e Y es perfectamente lineal y conocemos los valores α y β, la ecuación (1) permite predecir qué valor de *Y* corresponde a cualquier valor de *X*. Más aún, dos pares de datos son suficientes para determinar los parámetros α y β, de la misma manera que **dos puntos y una regla alcanzan para dibujar una línea recta**. La relación entre datos reales es rara vez tan simple.

27.3 Modelo de Regresión Lineal Simple

En forma más realista podríamos plantear que el **valor esperado** (la media poblacional) de ***Y***, más que los valores individuales, **cambia linealmente con *X***:

$$E\left[Y_i / X_i = x_i\right] = \alpha + \beta x_i \,, \qquad (2)$$

donde α es igual a la media poblacional de *Y* cuando *X* = 0. Con un aumento de una unidad en *X* se obtiene un aumento de la media poblacional de *Y* en β unidades. Este tipo de modelos tiene muchas aplicaciones prácticas.

En el caso de la presión y la temperatura el modelo dice que la **media poblacional** de las mediciones de la presión para una temperatura fija está dada por

$$\alpha + \beta \text{ TEMP}$$

Otras cosas, además de *X*, causan que los valores observados de Y_i varíen alrededor de la media de todos los valores de *Y* cuando *X* toma el valor *x, E*[*Y/X*=*x*]. Esas otras cosas son lo que determinan el error (de medición en nuestro ejemplo) ε_i.

$$\varepsilon_i = Y_i - (\alpha + \beta x_i) = Y_i - E[Y_i / X_i = x_i]. \quad (3)$$

El valor de Y_i es igual a la media más un error:

$$Y_i = E[Y_i / X_i = x_i] + \varepsilon_i$$
$$= \alpha + \beta x_i + \varepsilon_i .$$

Por lo tanto, otra forma de expresar el modelo lineal dado en (2) es: los valores de la variable respuesta se encuentran relacionados linealmente con la variable explicativa más un error. Tenemos así el siguiente

Modelo de regresión lineal simple

$$Y_i = \alpha + \beta x_i + \varepsilon_i . \quad (4)$$

Si nos interesa predecir PRESION a partir de TEMP (tabla1), llamaremos a la primera variable **respuesta** y a la segunda variable **explicativa** o **predictora**. La variable respuesta siempre se grafica en el **eje vertical**, o **eje *Y,*** y la variable predictora en el **eje horizontal,** o **eje *X,*** como muestra el **diagrama de dispersión** de la figura 1.

El problema consiste en ajustar una recta que represente al conjunto de datos de la mejor manera, para obtener la predicción de *Y* para cualquier valor de *X*. Hay muchas maneras de evaluar si una recta representa bien al conjunto de datos. El enfoque tradicional consiste en hallar la recta que en promedio tenga la menor **distancia vertical, residuo,** al cuadrado a cada uno de los puntos. Este procedimiento se llama método de **Cuadrados Mínimos** (CM) y lo describiremos en la Sección 14.5.

```
UNWEIGHTED LEAST SQUARES LINEAR REGRESSION OF PRESION

PREDICTOR
VARIABLES    COEFFICIENT    STD ERROR     STUDENT'S T       P
---------    -----------    ---------     -----------    ------
CONSTANT        26079.9       11.9034       2190.97      0.0000
TEMP           -39.8935       0.35843       -111.30      0.0000
```

La **recta de regresión** ($\hat{y} = a + b x$) obtenida por el método de cuadrados mínimos para los datos de la tabla 1 es:

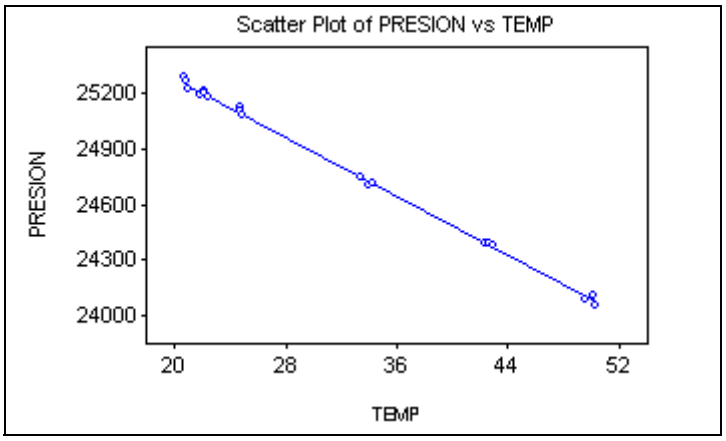
PRESION *ajustada* ($\hat{y}$) = 26079.9 - 39.8935 TEMP (5)

El valor *ajustado* ($\hat{y}$) puede utilizarse de dos maneras distintas:

a) como *estimador* de la media poblacional de Y para cada x fijo, en este caso como estimador de la media de la presión de transición para una temperatura fija.
b) como *predictor* de un valor futuro de Y para un valor fijo de x.

La diferencia entre a) y b) se encuentra únicamente en la varianza de $\hat{y}$.

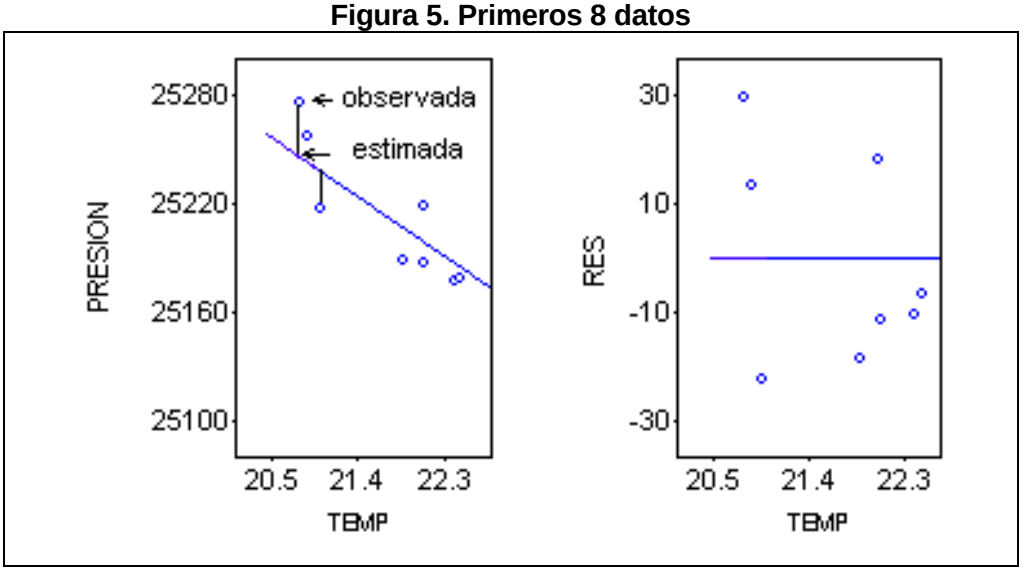
Figura 4. Diagrama de dispersión junto con la recta ajustada.



27. 4 Residuos

El **residuo** de un punto a una recta en un diagrama de dispersión es la **distancia vertical** del punto a dicha recta. La figura 5 muestra el diagrama de dispersión de los datos junto con la recta ajustada y el diagrama de dispersión de los residuos vs. la temperatura para los primeros 8 datos .

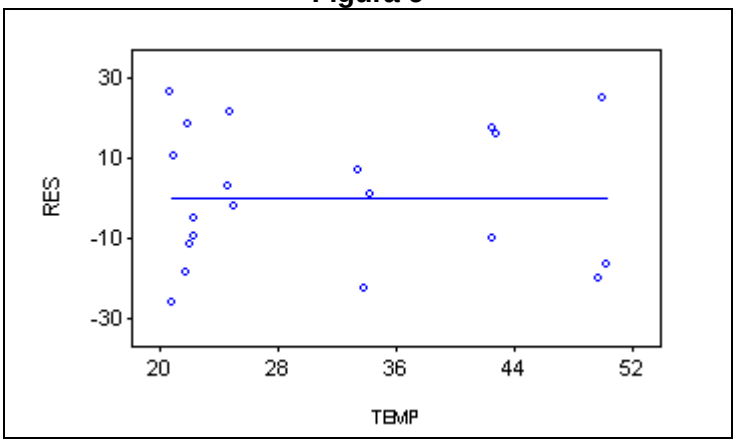
Figura 5. Primeros 8 datos



Algunos residuos son positivos, la presión observada está por encima de la recta, y otros son negativos, la presión observada está por debajo de la recta. La suma de todos los residuos es cero.

La figura 6 muestra el diagrama de dispersión de los residuos vs. la temperatura del conjunto de datos completo.

Figura 6



Para la primera medición TEMP = 20.8, PRESION = 25276. El residuo =25.8799 es positivo y se obtiene como

residuo = valor de *Y observado* - valor de *Y estimado* por la recta
= y - $\hat{y}$ = PRESION -{26079.9 - 39.8935 TEMP }
= 25276 - {26079.9 - 39.8935 * 20.8 }
= 25.8799

Para la tercera medición TEMP= 21, PRESION = 25216. El residuo= -26.1413 es negativo.

27.5 El Método de Cuadrados Mínimos

La **suma de los cuadrados de los residuos** (**RSS**) da una medida de la "bondad de ajuste" de la recta. Cuanto más pequeño es ese número tanto mejor es el ajuste.

Hemos observado valores de dos variables, *X* (TEMP) e *Y* (PRESION), y hemos realizado una "regresión de *Y* sobre *X*", obteniendo una recta que da un valor "ajustado" estimado de *Y* ($\hat{y}$, y "sombrero") para cada valor de la variable *X*.

Un estudio comienza por un modelo lineal (4) porque existe una teoría que lo sugiere o porque se desea comenzar de manera simple. En cualquiera de los dos casos, nos interesa obtener los mejores estimadores de los parámetros α y β. Si llamamos *a* y *b* a nuestros estimadores, la ecuación de la recta estimada es:

$$\hat{y}_i = a + b\, x_i \ ,$$

donde $\hat{y}_i$ (y_i "sombrero") indica el valor ajustado (o predicho) de la variable *Y* para el caso i (es el valor de la ordenada para x_i sobre la recta ajustada) (ver figura 5).

Los residuos e_i, la contraparte muestral de los errores (ε_i), son las diferencias entre el valor observado y el valor predicho:

$$e_i = y_i - \hat{y}_i$$
$$= y_i - (a + b\, x_i).$$

Los residuos miden el error de predicción. Como hemos visto, si el valor observado es mayor que el valor predicho ($y_i > \hat{y}_i$) el residuo es positivo; en caso contrario es negativo. Con una predicción perfecta ($y_i = \hat{y}_i$) resulta un residuo nulo. La suma de los cuadrados de los residuos (RSS) refleja la precisión y exactitud global de nuestras predicciones:

$$\text{RSS} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 . = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \qquad (6)$$

Cuanto más cerca estén los valores observados de los predichos tanto menor será RSS.

El **método de Cuadrados Mínimos** (CM) consiste en elegir a y b de manera que la suma de cuadrados de los residuos (RSS) sea lo más pequeña posible.

¿Cómo hallamos a y b ?

$$\begin{aligned} \frac{\partial\left(\sum_{i=1}^{n}(y_i - a - bx_i)^2\right)}{\partial a} &= -2\sum_{i=1}^{n}(y_i - a - bx_i) = 0 \\ \frac{\partial\left(\sum_{i=1}^{n}(y_i - a - bx_i)^2\right)}{\partial b} &= -2\sum_{i=1}^{n} x_i(y_i - a - bx_i) = 0 \end{aligned} \qquad (7)$$

Las ecuaciones anteriores definen los estimadores de los parámetros α y β resultan de derivar (6) con respecto a a y a b. Se trata de dos ecuaciones lineales con dos incógnitas cuyas soluciones son

$$\begin{aligned} b &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ a &= \bar{y} - b\bar{x} \end{aligned} \qquad (8)$$

Observaciones

- De la primera ecuación de (7) tenemos que la suma de los residuos es 0.
- La segunda ecuación de (8), nos dice que la recta de cuadrados mínimos pasa por $(\bar{x}, \bar{y})$, ya que $\bar{y} = a + b\bar{x}$.

Podemos pensar al método de cuadrados mínimos como fijando un punto, dado por el promedio de los valores de la variable explicativa (x´s) y el promedio de los valores de la variable respuesta (y's) y luego girando la recta que pasa por ese punto elegimos la que deja en promedio, en forma pareja, tantos valores observados por arriba como por abajo.

Ninguna otra recta tendrá, para el mismo conjunto de datos, una RSS tan baja como la obtenida por CM. En este sentido, el método de mínimos cuadrados brinda la solución que mejor ajusta a ese conjunto de datos.

Advertencia: en general no pueden realizarse predicciones fuera del rango de valores observados de la variable independiente.

¿Qué significa la ecuación de la recta (PRESION *estimada* ($\hat{y}$) = `26079.9 - 39.8935 TEMP`) ajustada?:

- La estimación de la variable PRESION, obtenida a partir de la ecuación de regresión ajustada, es el **valor predicho** de PRESION.
- Para cualquier valor de la variable TEMP un aumento en un grado de la temperatura produce una reducción de 39.8935(bar) en la **presión media** ("verdadera") de transición de Bismuto I-II.

El método de CM permite estimar una recta a partir de un conjunto de datos. Si estos datos son una muestra "adecuada" de una población, la recta nos permite extender resultados a dicha población. Ciertas características de los datos podrían invalidar los resultados del método.

27.6 Supuestos

Antes de utilizar el análisis de regresión y considerar medidas de incerteza o dispersión, es necesario conocer los supuestos en los que se basa el método. Veremos primero cuáles son esos supuestos y luego qué procedimientos pueden utilizarse para validarlos.

27.6.1 Descripción de los supuestos

Supuesto a: Normalidad de los errores.
Para cada valor x, de la variable predictora X, la variable respuesta Y debe tener distribución Normal

Por ejemplo, si se cumple este supuesto, la presión de transición (Y) es una variable aleatoria Normal con media μ_x que depende de x (temperatura).

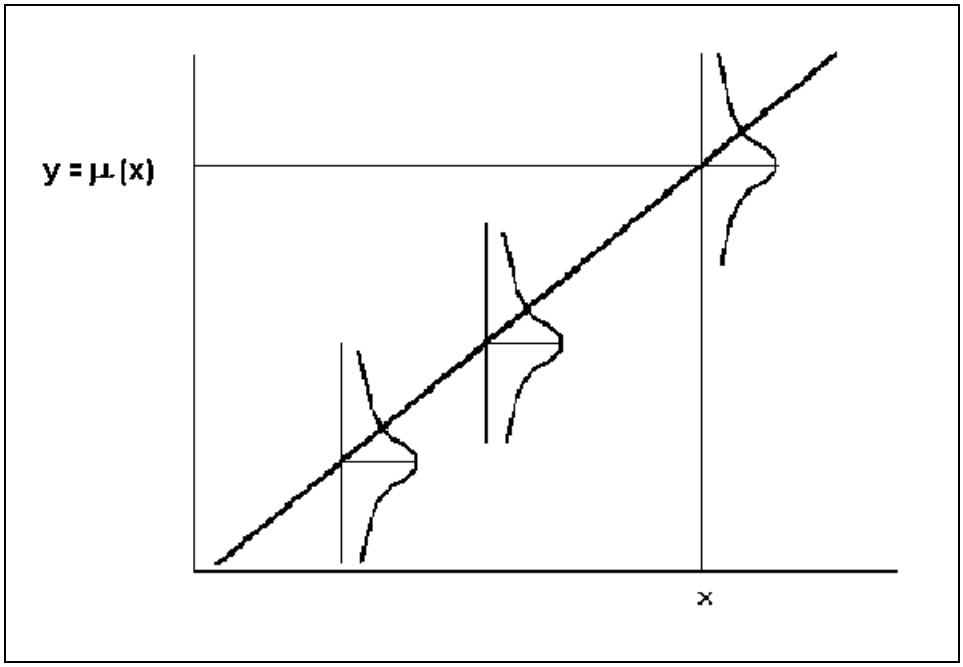
Supuesto b: Linealidad
La media de la variable *Y* varía linealmente con *X*.
Si pasar de 21 a 22 ^{0}C no fuera lo mismo que pasar 41 a 42 ^{0}C respecto del cambio de la presión de transición, este supuesto no se cumpliría.

Supuesto c: Homoscedasticidad
La variabilidad de *Y*, que es medida por su varianza (σ^2), o por su desvío estándar (σ), debe ser la misma para cada valor x de la variable *X.*

Este supuesto no se cumpliría por ejemplo si a medida que aumenta (o disminuye) la temperatura los valores de la presión de transición de Bismuto I-II estuvieran comprendidos en un rango más amplio.

No sabemos si los supuestos se satisfacen, ni conocemos los verdaderos valores de los parámetros α y β.

Figura 7. Supuestos de Normalidad, linealidad y homoscedasticidad



La figura 7 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 .

Supuesto d: Independencia de los errores

Hemos visto que cuando dos variables son independientes su correlación es cero, en general la recíproca no es cierta pero bajo el supuesto de normalidad el supuesto de independencia de los errores se reduce a que no estén correlacionados ($\text{corr}(\varepsilon_i, \varepsilon_j) = 0 \quad \forall i \neq j$).

Por ejemplo, si las presiones de transición fueron obtenidas en un orden secuencial con la temperatura, podría ocurrir que los errores fueran mayores en temperaturas más bajas que en temperaturas más altas invalidando el supuesto de independencia de los errores.

27.6.2 Validación de los Supuestos

La validación de los supuestos se realiza en base a los datos y a los residuos de los mismos respecto de la recta ajustada. El diagrama de dispersión de los datos permite obtener una impresión sobre el supuesto de linealidad y homoscedasticidad. El análisis posterior de residuos permitirá confirmar la impresión inicial y validar los supuestos de Normalidad e independencia.

Veamos algunas estructuras que suelen verse en los diagramas de dispersión de los residuos.

Figura 8

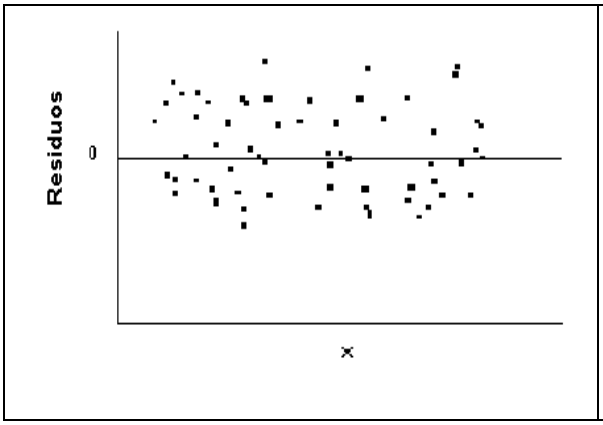
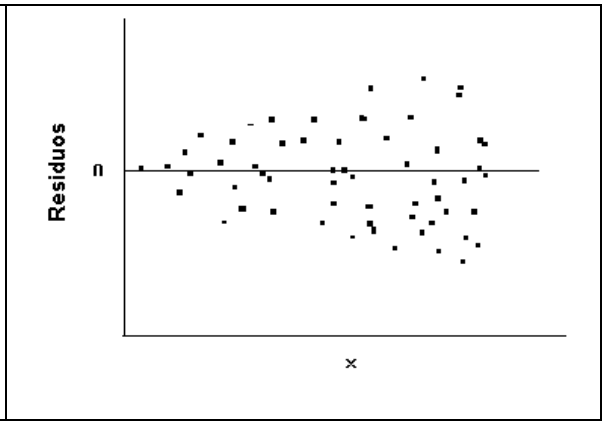
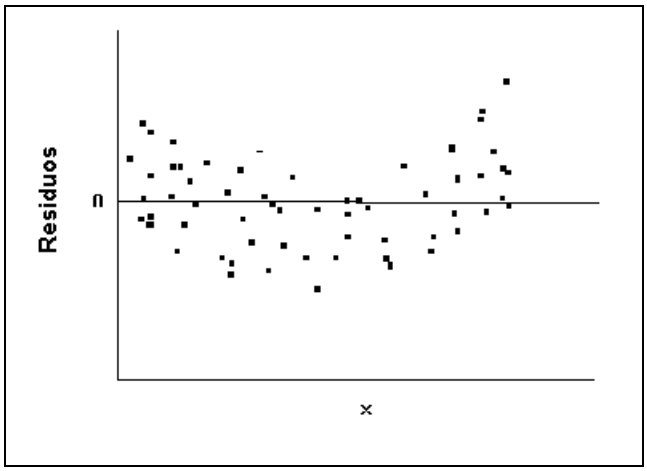


Figura 9. No se satisface el supuesto de homoscedasticidad



Si graficáramos los residuos contra los valores de *X* los puntos deberían estar distribuidos en forma de nube alrededor del valor 0 del residuo, para todos los valores de *X*, como se muestra en la figura 8. En la figura 9 los datos no satisfacen el supuesto c, ya que los residuos tienen variabilidad creciente a medida que *X* crece. En la figura 10 se aprecia una estructura curva en los residuos invalidando la linealidad.

Figura 10. No se satisface el supuesto de linealidad.

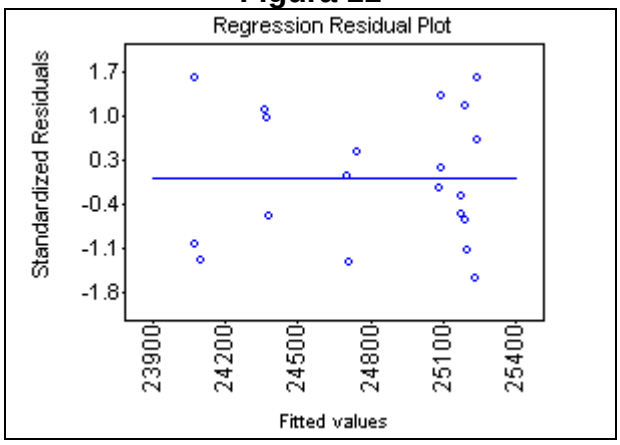


Cuando los supuestos a, b, c y d se satisfacen, los errores no están correlacionados y tienen una distribución Normal con media 0 y varianza constante.

Ejemplo. Retomemos el ajuste lineal de la presión (bar) de transición de Bismuto I-II con la temperatura (^{o} C).

El Statistix muestra un diagrama de dispersión similar, los “residuos estandarizados” en función de los “valores ajustados”. Como los valores ajustados son de la forma a + b x, el eje horizontal es simplemente un cambio de unidades respecto de los de las figuras 8, 9 y 10. Pero, ¿qué son los residuos estandarizados del eje vertical?

Figura 11



La figura 11 muestra el diagrama de dispersión (scatter plot) de los residuos estandarizados correspondientes al ejemplo de la presión de transición, graficados en función de los valores ajustados. El gráfico tiene un aspecto satisfactorio.

¿Qué son los Residuos Estandarizados?

En el método de cuadrados mínimos los valores de la variable explicativa alejados de su media tienden a acercar la recta hacia ellos, esto es llamado "efecto palanca". Como consecuencia, los residuos tienen una tendencia a ser menores para valores de x extremos. La varianza de los residuos será más chica **si x_i está lejos de su promedio**

$$\mathrm{Var}(e_i) = \sigma^2 (1 - h_{ii})$$

y el valor ajustado ($\hat{y}_i$) estará cerca del valor observado por efecto palanca.

El **residuo estandarizado** definido por:

$$r_{si} = \frac{y_i - \hat{y}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}} = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_{ii}}} \qquad (9)$$

o sea, el residuo dividido por su error estándar, donde

$$\hat{\sigma} = \sqrt{\frac{RSS}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y})^2}{n-2}} \qquad (10)$$

es un estimador de σ, el desvío estándar de los errores, y h_{ii} es una medida, de la distancia de cada valor de la variable explicativa (x_i) a la media muestral $\bar{x}$, llamada **palanca,** o **"leverage"**:

aumenta a medida que x_i se aleja del promedio $\bar{x}$.

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{k=1}^{n}(x_k - \bar{x})^2}$$

Suma fija

(11)

Recordemos: un gráfico de probabilidad Normal de un conjunto de datos indica que estos tienen una distribución aproximadamente normal cuando los puntos caen aproximadamente sobre una recta.

Ejemplo. Retomemos nuevamente el ajuste lineal de la presión (bar) de transición de Bismuto I-II con la temperatura (^{o} C).

La figura 12 muestra el gráfico de probabilidad Normal de los residuos correspondientes al mismo ejemplo. Vemos que es razonablemente lineal aunque presenta colas un poco más livianas que lo esperable bajo Normalidad. El valor del estadístico de Shapiro-Wilk es 0.9423 y su valor-p= 0.2652 > 0.20 (cuanto más alto es el p-valor mayor es la evidencia a favor de la hip. nula de Normalidad de los errores). No se rechaza el supuesto.

Figura 12

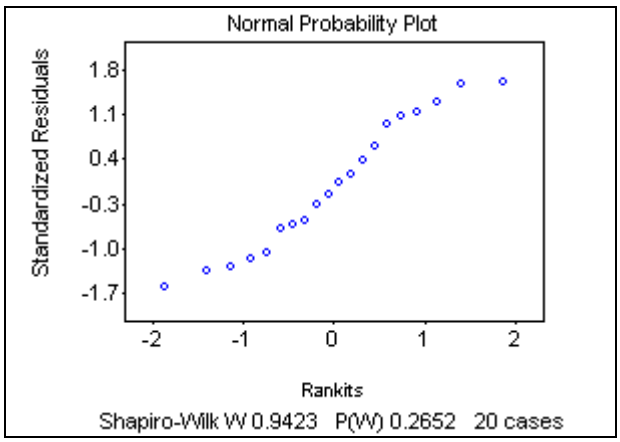


Figura 13

Con respecto al supuesto d, es útil realizar un diagrama de dispersión de los residuos en función del orden en que fueron realizadas las mediciones, figura 13. Los tests para estudiar la auto-correlación de los errores (ε) se basan en examinar los residuos (e). Usualmente esto es realizado mediante el test de Durbin-Watson:

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n}(e_i)^2} . \qquad (12)$$

```
DURBIN-WATSON TEST FOR AUTOCORRELATION

DURBIN-WATSON STATISTIC  2.3061

P-VALUES, USING DURBIN-WATSON'S BETA APPROXIMATION:
  P (POSITIVE CORR) = 0.6750,  P (NEGATIVE CORR) = 0.3250

EXPECTED VALUE OF DURBIN-WATSON STATISTIC     2.1065
EXACT VARIANCE OF DURBIN-WATSON STATISTIC    0.17682
```

Si los sucesivos residuos son no correlacionados entonces $d \cong 2$. Valores de d alejados de 2 indican la presencia de cierto tipo de auto-correlación, la denominada auto-correlación a lag 1. Un p-valor alto asociado a este estadístico, por ejemplo mayor que 0.10, indica que la diferencia de d con 2 no es estadísticamente significativa, que puede atribuirse al azar, y por lo tanto no se rechaza la hipótesis nula de residuos no correlacionados. La aplicación del test de Durbin-Watson sólo tiene sentido cuando la varaiable explicativa está ordenada en el tiempo o en el espacio, o sea por ejemplo, cuando el índice i representa el tiempo. Por otra parte, y debido a que el test fue diseñado para detectar cierto tipo especial de auto-correlación, puede ocurrir que con este estadístico no se detecten auto-correlaciones cuando en realidad el supuesto no se satisface.

Concluimos que se satisfacen razonablemente los supuestos del Método de Cuadrados Mínimos.

27.7.1 Evaluación de la asociación: Asociación positiva, asociación negativa

Dado un conjunto de pares de datos (que corresponden a valores observados de dos variables aleatorias) decimos que están asociados en forma **positiva** cuando los valores que están por **encima del promedio** en una componente del par tienden a ocurrir mayoritariamente con valores por **encima del promedio** de la otra. Recíprocamente la

asociación es **negativa** cuando valores por **encima del promedio** de una componente suelen estar acompañados por valores por **debajo del promedio** de la otra.

Si se ajusta una recta por cuadrados mínimos en el diagrama de dispersión de un conjunto de pares de datos, el signo de la pendiente, b, de la recta ajustada indica si la asociación es positiva o negativa.

Sin embargo, la pendiente no mide directamente la fuerza (o grado) de la asociación. Esto se debe a que el valor absoluto de la pendiente está intrínsecamente vinculado a las unidades en las que se han expresado las mediciones. Podemos obtener valores tan grandes o tan chicos como queramos con sólo elegir las unidades adecuadamente.

Las medidas de asociación que consideramos a continuación no varían con cambios en las unidades de medición.

27.7.2 Correlación de Pearson

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}. \qquad (13)$$

El coeficiente de correlación de Pearson mide el grado de asociación *lineal* de un conjunto de datos $(x_1, y_1), \ldots, (x_n, y_n)$, de tamaño n, correspondiente a observaciones de dos variables continuas *X* e *Y*

Como la pendiente de la recta ajustada por cuadrados mínimos está dada por

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

resulta que el coeficiente de correlación se puede expresar de la siguiente manera

$$r = \frac{s_X}{s_y} b \qquad \text{ó} \qquad b = r \frac{s_Y}{s_X}. \qquad (14)$$

Si los desvíos estándar s_x y s_y son iguales o cuando los valores de ambas variables (X e Y) han sido estandarizados, de manera que sus medias muestrales son cero y su desvíos estándar 1, entonces la recta de regresión tiene pendiente b = r y pasa por el origen. Es por esta razón que el coeficiente de correlación de Pearson es también llamado *coeficiente de regresión estandarizado*.

27.7.3. Propiedades del Coeficiente de correlación

- La correlación, a diferencia de la pendiente b, trata a los valores x´s e y´s en forma simétrica. El valor del coeficiente de correlación muestral r , no depende de las

unidades en que se miden las variables y su valor está siempre entre -1 y 1.

- A mayor valor absoluto de r, mayor el grado de asociación lineal.
- r tiene el mismo signo que b.
- Cuando r = 0 también b = 0, no hay una tendencia lineal creciente o decreciente en la relación entre los valores x´s e y´s.
- Los valores extremos, *r* = 1 y *r* = -1, ocurren únicamente cuando los puntos en un diagrama de dispersión caen exactamente **sobre** una recta. Esto corresponde a asociaciones positivas ó negativas perfectas. En este caso, el error de predicción es cero al utilizar la recta ajustada $\hat{y}$ = a + b x, para predecir el valor de Y.
- Valores de *r* cercanos a 1 ó -1 indican que los puntos yacen **cerca** de una recta.
- Valores de r positivos indican que la mayoría de los desvíos $x_i - \bar{x}$ e $y_i - \bar{y}$ tendrán el mismo signo, es decir, hay una asociación positiva entre las variables.
- Valores de r negativos indican que la mayoría de los desvíos $x_i - \bar{x}$ e $y_i - \bar{y}$ tendrán signos opuestos, es decir, hay una asociación negativa entre las variables.

La figura 14 muestra cómo los valores de r se acercan a cero, i.e. se alejan del 1 ó -1, a medida que decrece el grado de asociación lineal entre las variables.

Figura 14. Comportamiento del coeficiente de correlación a medida que decrece el grado de asociación lineal.

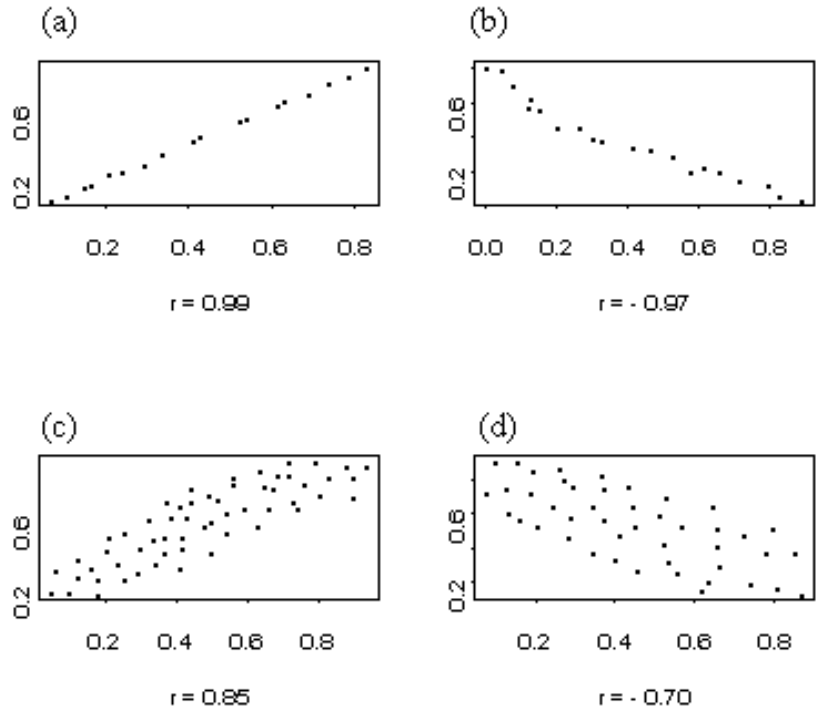
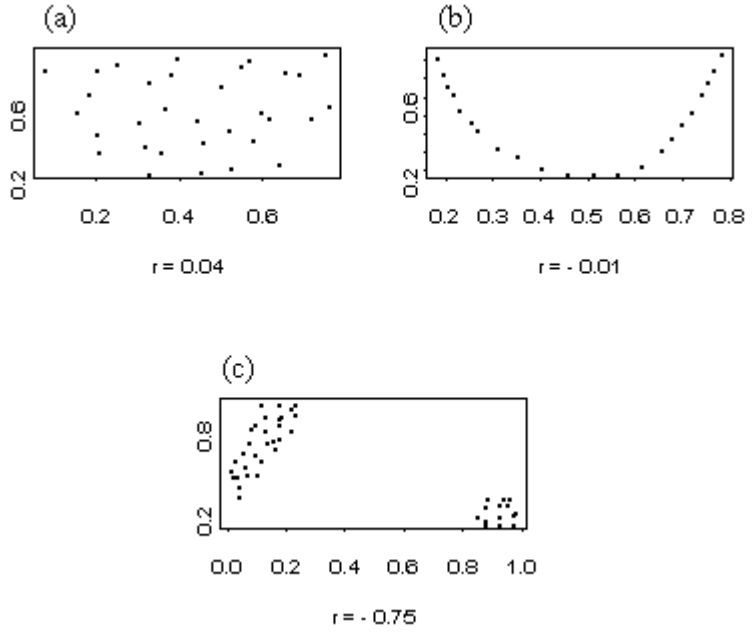


Figura 15. Comportamiento del coeficiente de correlación ante diferentes tipos de asociaciones no lineales.



La figura 15. (a) muestra una falta total de asociación entre las dos variables y un coeficiente de correlación cercano a cero, en cambio la (b), que también tiene un r cercano a cero, muestra una clara relación funcional entre las variables y la (c) muestra dos grupos uno con asociación positiva y otro con asociación nula, sin embargo el coeficiente de correlación, -0.75, indica una asociación negativa.

27.7.4 Coeficiente de determinación

El coeficiente de determinación es una medida de la proporción en que se reduce el error de predicción de una variable respuesta (Y) cuando se predice Y utilizando los valores de una variable X en la ecuación de predicción, $\hat{y}$ = a + b x, con respecto al error de predicción que se obtendría sin usarla.

El **coeficiente de determinación R^2** se define como

$$R^2 = (TSS - RSS) / TSS = 1 - RSS / TSS \qquad (15)$$

- TSS, es la suma de cuadrados total. Es decir, la suma de los cuadrados de las desviaciones de cada respuesta observada a la media:

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2 . \qquad (16)$$

  Mide el error de predicción si no se tiene en cuenta la variable explicativa y se predice la respuesta por $\bar{y}$, la media muestral de las respuestas observadas.
- RSS, es la suma de cuadrados de los residuos:

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 . \qquad (17)$$

Es una medida del error de predicción que se comete cuando la variable respuesta se predice por $\hat{y}$ = a + b***x***, la teniendo en cuenta la variable explicativa.

Cuando hay una fuerte relación lineal entre X e Y, el modelo ajustado provee predictores $\hat{y}$ mucho mejores que $\bar{y}$, en el sentido que la suma de cuadrados de los errores de predicción es mucho menor.

Como el numerador de R^2 , TSS - RSS es igual a $\Sigma (\hat{Y}_i - \bar{Y})^2$, R^2 puede expresarse como

$$R^2 = \frac{\sum \left(\hat{Y}_i - \bar{Y}\right)^2}{\sum \left(Y_i - \bar{Y}\right)^2} .$$

El numerador se denomina suma de cuadrados explicada por el modelo (REGRESSION sum of squares en el STATISTIX, y ESS en otros programas). Por lo tanto R^2 mide la proporción de la variación total explicada por la regresión.

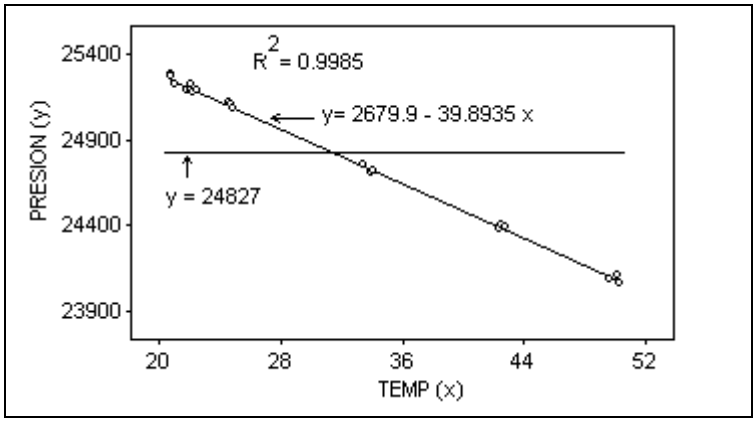
La correlación al cuadrado coincide con el coeficiente de determinación **únicamente en regresión lineal simple**:

$$r^2 = R^2 \qquad (18)$$

Consideremos el diagrama de dispersión del presión de transición versus la temperatura que aparece en la figura 16. La línea horizontal representa la media de los valores de PRESION, 24827, de las 20 observaciones. Los valores observados de Y varían, como lo indican las desviaciones verticales de los puntos a la línea horizontal.

La otra recta es la de mínimos cuadrados. Los desvíos verticales de los puntos a esta recta son, en muchísimo menores. Sobre esta recta, cuando *x* cambia *y* cambia, de manera que esta relación lineal explica una parte de la variación de *Y*. Como R^2 = 0.9985 decimos que la recta de regresión explica el 99% de la variación total observada en la presión de transición.

Figura 16



- El coeficiente de determinación no depende de las unidades en que se expresan los datos y toma valores entre cero y uno.
- Vale 0 cuando la regresión no explica nada; en ese caso, la suma de cuadrados total es igual a la suma de cuadrados de los residuos.
- Vale 1 cuando la variabilidad observada de la respuesta es explicada totalmente por la regresión; en ese caso, la suma de los cuadrados de los residuos es cero.

Ejemplo. Retomemos nuevamente el ajuste lineal de la presión (bar) de transición de Bismuto I-II con la temperatura (^{o} C).

La tabla 2 muestra la salida del Statistix para la regresión de la presión de transición sobre la temperatura.

Tabla 2

```
UNWEIGHTED LEAST SQUARES LINEAR REGRESSION OF PRESION

PREDICTOR
VARIABLES     COEFFICIENT     STD ERROR         STUDENT'S T        P
---------     -----------     ---------         -----------     ------
CONSTANT (a)     26079.9      11.9034            2190.97        0.0000
TEMP     (b)    -39.8935      0.35843 ⇐ s(b)    -111.30        0.0000

R-SQUARED            0.9985      RESID. MEAN SQUARE (MSE)  298.854 ⇐ σ̂²
ADJUSTED R-SQUARED   0.9985      STANDARD DEVIATION        17.2874 ⇐ σ̂

SOURCE         DF        SS            MS          F         P
----------    ---    ----------    ----------    -----     ------
REGRESSION      1      3702187       3702187    12387.96    0.0000
RESIDUAL       18      5379.36       298.854 ⇐σ̂²

TOTAL          19      3707567

CASES INCLUDED 20   MISSING CASES 0
```

La columna encabezada **COEFFICIENT** presenta la ordenada al origen (CONSTANT = ***a*** = 26079.9) y la pendiente (***b*** = -39.8935).

La columna encabezada **STD ERROR** presenta los correspondientes errores estándar.

El cociente entre el valor del coeficiente y su error estándar aparece en la columna **STUDENT'S T**. Sus p-valores, presentados en la columna encabezada **P** , permiten testear si los coeficientes son significativamente distintos de cero. En este caso ambos p-valores (0.0000 y 0.0000) son menores que 0.01 y decimos que los coeficientes *a* y *b* son altamente significativos.

El coeficiente de la variable TEMP, es estadísticamente significativo distinto de cero, entonces decimos que **la variable** temperatura **explica la respuesta**, presión de transición

Otros estadísticos incluidos en la tabla 2 son:

R-SQUARED: Coeficiente de determinación $R^2 = 1 - \frac{5379.36}{3707567} = 0.9985$.

RESID. MEAN SQUARE ERROR (MSE), $\hat{\sigma}^2$: Cociente entre la suma de cuadrados de los residuos y sus grados de libertad, RSS / DF = 5379.36 / 18 = 298.854

STANDARD DEVIATION es la raíz cuadrada del MSE. Es un estimador del desvío σ. En nuestro ejemplo $\hat{\sigma} = \sqrt{298.854} = 17.2874$.

La salida incluye también una tabla, denominada tabla de análisis de la varianza correspondiente al ajuste, la cuál contiene las sumas de cuadrados, bajo el encabezado "**SS**", los grados de libertad de cada suma de cuadrados ("**DF**"), el cociente entre la suma de cuadrados y sus correspondientes grados de libertad ("**MS**") y finalmente el estadístico **F** y su **p**-valor asociado.

TOTAL SS Es la suma de cuadrados total, TSS = 3707567 (g.l. = 19)

RESIDUAL SS Es la suma de cuadrados residual, RSS = 5379.36 (g.l. = 18)

REGRESSION SS Es la suma de cuadrados explicada por el modelo, TSS - RSS = 3702187

REGRESSION MS (TSS - RSS) / 1 =3702187 / 1 = 3702187
RESIDUAL MS: RSS / 18 = 5379.36 / 18 = 298.854 $\Leftarrow \hat{\sigma}^2$

ESTADISTICO F Es una medida global de la bondad de la regresión y se calcula como el cociente de los dos últimos valores definidos, es decir

$$F = \frac{(TSS - RSS)/1}{RSS/18} = \frac{3702187}{298.854} = 12387.96$$

El correspondiente valor-p = 0.0000 y por lo tanto decimos que la regresión es altamente significativa.

Observación : En la **regresión lineal simple únicamente**, el p-valor del estadístico F coincide con el del test para decidir si la pendiente es estadísticamente distinta de cero. Esto no ocurre cuando se incluyen más variables explicativas al modelo.

28 Intervalos de confianza para la pendiente y la ordenada al origen

Podemos calcular un intervalo de confianza para la pendiente de la recta ajustada y también realizar un test para decidir si es significativamente distinta de cero. Una pendiente cero querría decir que no hay relación lineal entre *Y* y *X*.

Recordemos que la pendiente de la recta ajustada es:

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

Luego Var (b) = $\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ (considerando que las x's son fijas y sin error)

y el **desvío estándar estimado de la pendiente** es:

$$s(b) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \qquad (19)$$

donde $\hat{\sigma}^2$ es un estimador de σ^2, calculado como suma de los cuadrados de los residuos dividida por *n*-2.

Rara vez será necesario utilizar (19) para el cálculo ya que s(b) es un valor que muestra el Statistix automáticamente al realizar un ajuste por cuadrados mínimos. En este caso, *s(b)* = 0.35843 (Vea la tabla 2, **STD ERROR**). Sin embargo es interesante detenerse en su expresión ; vemos que:

- intervienen todas las observaciones,
- a medida que aumenta el número de observaciones (*n*), *s*(b) se hace cada vez más pequeño,
- para un mismo número de observaciones, cuanto más dispersos estén los valores x_i tanto más pequeño será *s*(b).

Los extremos de un intervalo de confianza de nivel (1-α) 100% para la pendiente son:

$$b \pm t_{n-2,\,\alpha/2} * s(b)\,, \qquad (20a)$$

Conclusión: si queremos que la estimación de la pendiente de la recta sea lo más precisa posible, debemos elegir un tamaño de la muestra grande y los valores de la variable explicativa lo más espaciados que se pueda dentro del rango de interés.

Tomando α = 0.05, en nuestro ejemplo *n* -2 = 18, b = -39.8935, *s(b)* = 0.35843 y $t_{18,\,0.025}$ = 2.10, resulta el intervalo del 95% de confianza (-40.646, -39.141)de las pendientes compatibles con los datos.
Como el cero no pertenece al intervalo de confianza obtenido, se rechaza la hipótesis de pendiente nula al nivel 0.05.

Un **intervalo de confianza** de nivel (1-α) 100% para la ordenada al origen es de la forma

$$a \pm t_{n-2,\,\alpha/2} * s(a)$$

donde

$$s(a) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

En el ejemplo (tabla 2) a= `26079.9` y `s(a) = 11.9034`

29. Intervalos de confianza e intervalos de predicción.

Hemos dicho (sección 27.3) que la recta ajustada puede utilizarse de dos maneras distintas

a) para *estimar* de la media poblacional de Y para cada x fijo.
b) para *predecir* un valor futuro de Y para un valor fijo de x.

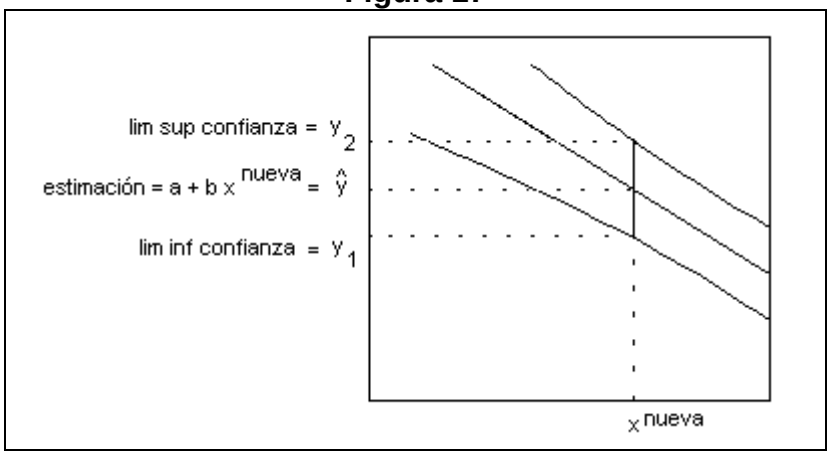
Agregaremos algunas más

c) para *estimar* de la media poblacional de Y para varios valores de x diferentes.
d) para *predecir* varios valores futuros de Y para cada uno con un valor fijo de x diferente.
e) realizar predicciones del valor de X que dio lugar al nuevo valor observado de Y. Esto se llama *predicción inversa.*

Los intervalos que resultan de a) - d) están todos centrados en $\hat{y}$, difieren únicamente en su amplitud debido a la diferencia en las varianzas.

29.1 Intervalos de confianza para la respuesta media

Figura 17



Debe tenerse en cuenta la incerteza de la recta ajustada. Para ello se construye una *banda* alrededor de la recta de regresión ajustada, tal que para cada valor fijo de x (x^{nueva}), el intervalo determinado por la banda y una recta vertical a la abscisa en x^{nueva}, sea un intervalo de confianza del (1-α) 100%:

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

Si llamamos $s(a + b\,x^{nueva})$ a $\hat{\sigma}\;\sqrt{\frac{1}{n} + \frac{(x^{nueva} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$

tenemos que los límites de un intervalo de confianza para la media de la variable Y dado el valor x^{nueva} son

$$a + b\,x^{nueva} \pm t_{n-2,\,\alpha/2}\;s(a + b\,x^{nueva}) \qquad (21)$$

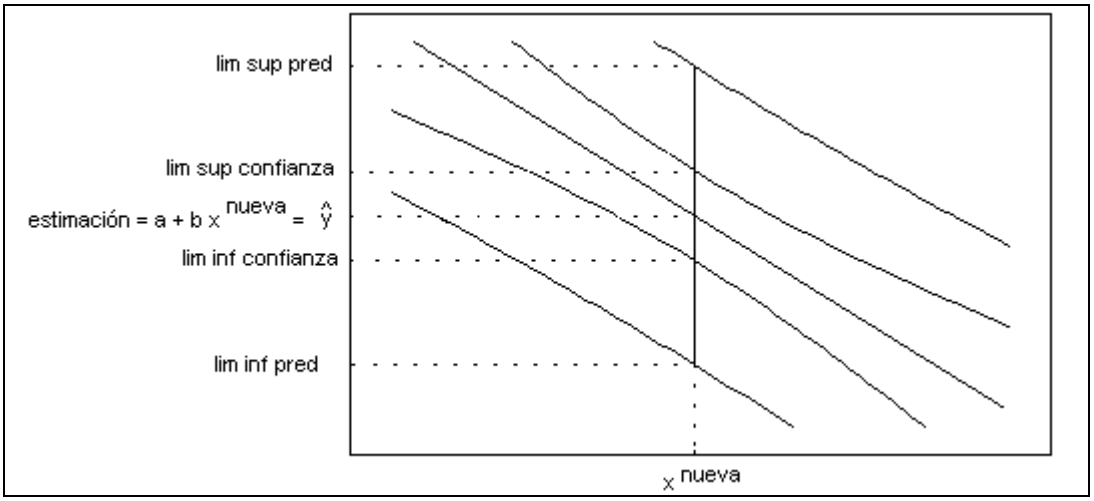
En particular si α = 0.05, el 95% de confianza significa que el intervalo es uno de una familia de intervalos, tal que 95 de cada 100 contienen la verdadera media de *Y* para

ese valor fijo de x (x^{nueva}) ; 5 no. Podemos confiar en que el que tenemos es uno de esos 95.

29.2 Intervalos de predicción para una observación futura

Para un mismo valor (*x*) de la variable explicativa, un intervalo de predicción refleja, además de la variabilidad debida a que la recta estimada no representa exactamente la media verdadera de la variable respuesta para ese valor de X, la **variabilidad** individual de la variable respuesta alrededor de la media verdadera y es por esa razón es de mayor amplitud que el intervalo de confianza.

Figura 18. Intervalos de confianza junto con los intervalos de predicción para una observación futura



La expresión general de los límites de predicción del (1-α) 100 % para una observación futura (y^{nueva}) para el valor x^{nueva} de la variable explicativa es:

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

$$a + b\,x^{nueva} \pm t_{n-2,\alpha/2}\ \ s\,(\text{pred } Y) \qquad , \qquad (22)$$

La única diferencia entre el intervalo de confianza (21) y el de predicción (22) es que aparece un 1 dentro de la raíz. Esta diferencia hace que la longitud de los intervalos de confianza pueda hacerse tan pequeña como se quiera, con tal de tomar suficientes observaciones, mientras que la longitud de los intervalos de predicción nunca pueda ser menor que $2\ t_{n-2,\ \alpha/2}\ \hat{\sigma}$

Si la cantidad de observaciones es grande la raíz que aparece en la expresión (22) es aproximadamente igual a 1 y la longitud del intervalo de predicción de nivel 0.95, resulta cerca de 4*s.* Por lo tanto, si estamos interesados en predicción, $4\hat{\sigma}$ es un excelente indicio de la calidad del ajuste, y como consecuencia, de la incerteza de las predicciones.

Ejemplo: Interesa estudiar la relación entre la pureza del oxigeno (Y) producido en un proceso de destilación y el porcentaje de hidrocarburos (X) presentes en el

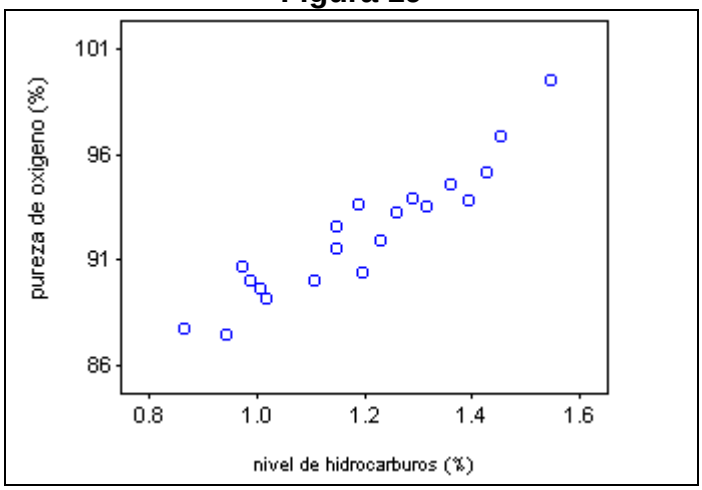
condensador principal de la unidad de destilación. No se conoce un modelo determinístico funcional que relacione la pureza del oxígeno con los niveles de hidrocarburo.

Tabla 3. Niveles de oxigeno e hidrocarburos

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

El diagrama de dispersión de la figura 19 muestra que a pesar de que ninguna curva simple pasará por todos los puntos hay una tendencia lineal creciente de manera que es razonable suponer que la media de la pureza de oxígeno esté relacionada linealmente con el nivel de hidrocarburos.

Figura 19

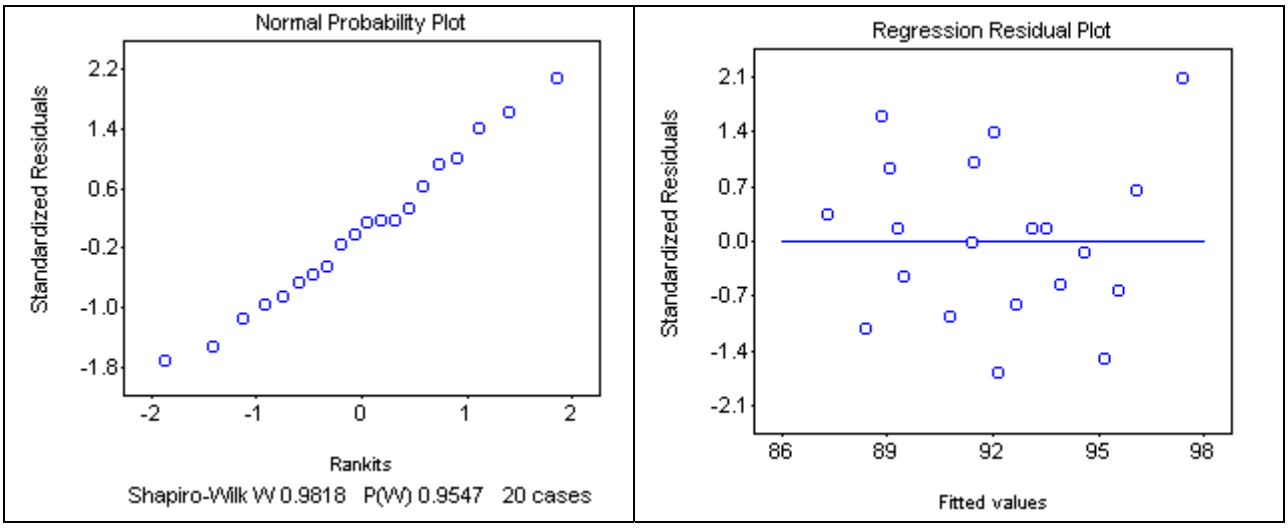


La tabla 4 los coeficientes y sus errores estándar, resultantes de un ajuste de cuadrados mínimos a los datos de la pureza de oxígeno. La variable X (% de hidrocarburos) es estadísticamente significativa.

Tabla 4

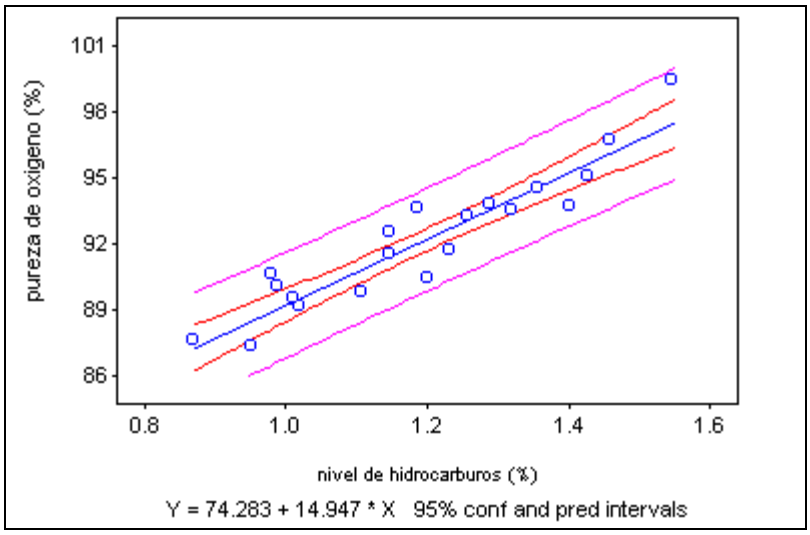
```
PREDICTOR
VARIABLES     COEFFICIENT     STD ERROR      STUDENT'S T        P
---------     -----------     ---------      -----------     ------
CONSTANT         74.2833        1.59347         46.62        0.0000
X                14.9475        1.31676         11.35        0.0000

R-SQUARED      0.8774        RESID. MEAN SQUARE (MSE)     1.18055
ADJUSTED R-SQUARED  0.8706   STANDARD DEVIATION           1.08653
```

Figura 20



Los gráficos de la figura 20 nos permiten concluir que los datos no presentan alejamientos de los supuestos de Normalidad y homoscedasticidad.

Figura 21. Recta ajustada junto con las bandas de confianza y de predicción del 95%



La banda interna de la figura 21 es la banda de confianza (21). El intervalo más angosto, (91.650 , 92.671), se encuentra en el nivel promedio de hidrocarburos ($\bar{x}$ = 1.196 %). Los intervalos se van ensanchando a medida que aumenta la distancia a dicho valor promedio. Un alto porcentaje de valores observados cae fuera de la banda de confianza. Esto pone de manifiesto que dichas bandas están formadas por intervalos de confianza para la respuesta media, nada dicen respecto de los valores de la variable de interés.

La longitud de estos intervalos decrece con el aumento del tamaño de la muestra y/o de la dispersión de los valores de la variable independiente.

Siguiendo con el ejemplo, en el nivel promedio de hidrocarburos (1.196 %), el intervalo de predicción es (89.821 ; 94.500). Los intervalos de predicción (22) del 95% también se ensanchan con la distancia al nivel promedio de hidrocarburos, aunque esto no se ve fácilmente de la figura.

29.3 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:

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

29.4 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.

29.4.1 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:

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

$$a + b\, x_k^{nueva} \pm t_{n-2,\ \alpha/K2}\, s(a + bx_k^{nueva}) \qquad 1 \le k \le K \qquad (24)$$

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%

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

$$a + b\,x^{nueva} \pm \sqrt{2f_{2,n-2,\alpha}}\ s(a + b\,x^{nueva}) \qquad (25)$$

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.

29.4.2 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

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

$$a + b\,x_k^{nueva} \pm t_{n-2,\alpha/2K}\ s\,(pred\ Y) \qquad 1 \le k \le K \qquad (26)$$

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

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

$$a + b\,x_k^{nueva} \pm \sqrt{Kf_{k,n-2,\alpha}}\ s\,(pred\ Y) \qquad 1 \le k \le K \qquad (27)$$

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.

29.5 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:

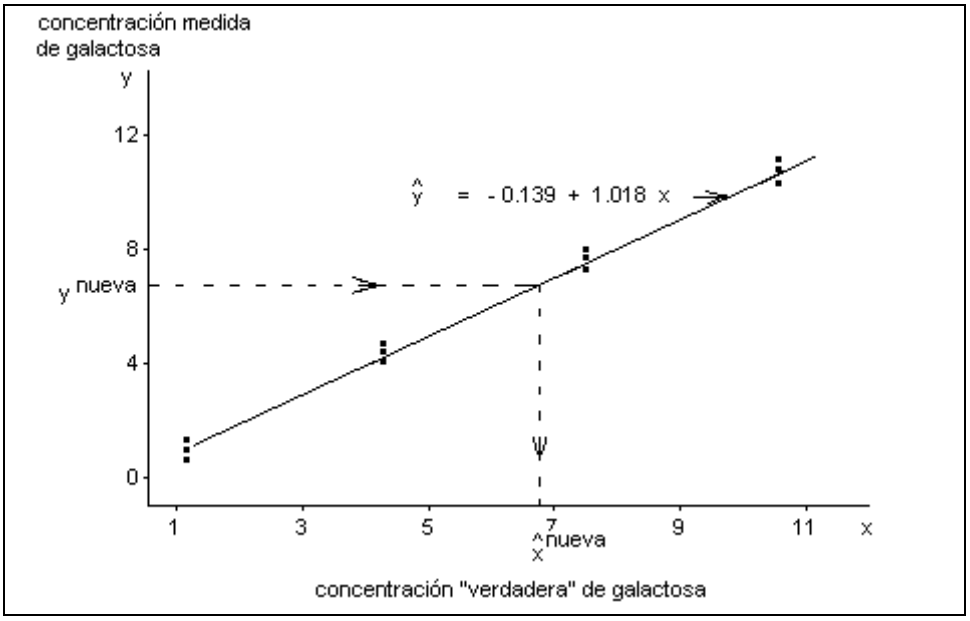
$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

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:

$$\hat{y} = a + bx \tag{28}$$

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} (método barato) y se quiere estimar el nivel x^{nueva} (método caro), es natural obtener un estimador de x^{nueva} despejando de (28)

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

29.5.1 Límites de estimación resultantes de una predicción inversa

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

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

donde $s(\text{pred}\,\hat{x}^{\text{nueva}})$ =

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

=

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

Volviendo al ejemplo

Tabla 5. Medidas resumen de los datos de galactosa

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

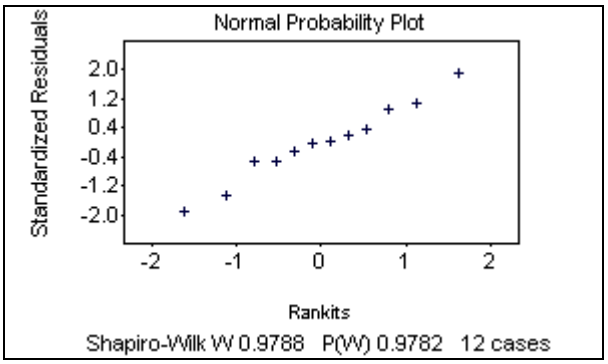
Tabla 6 Resultados del ajuste por cuadrados mínimos a los datos de galactosa

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			

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 5 y 6 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

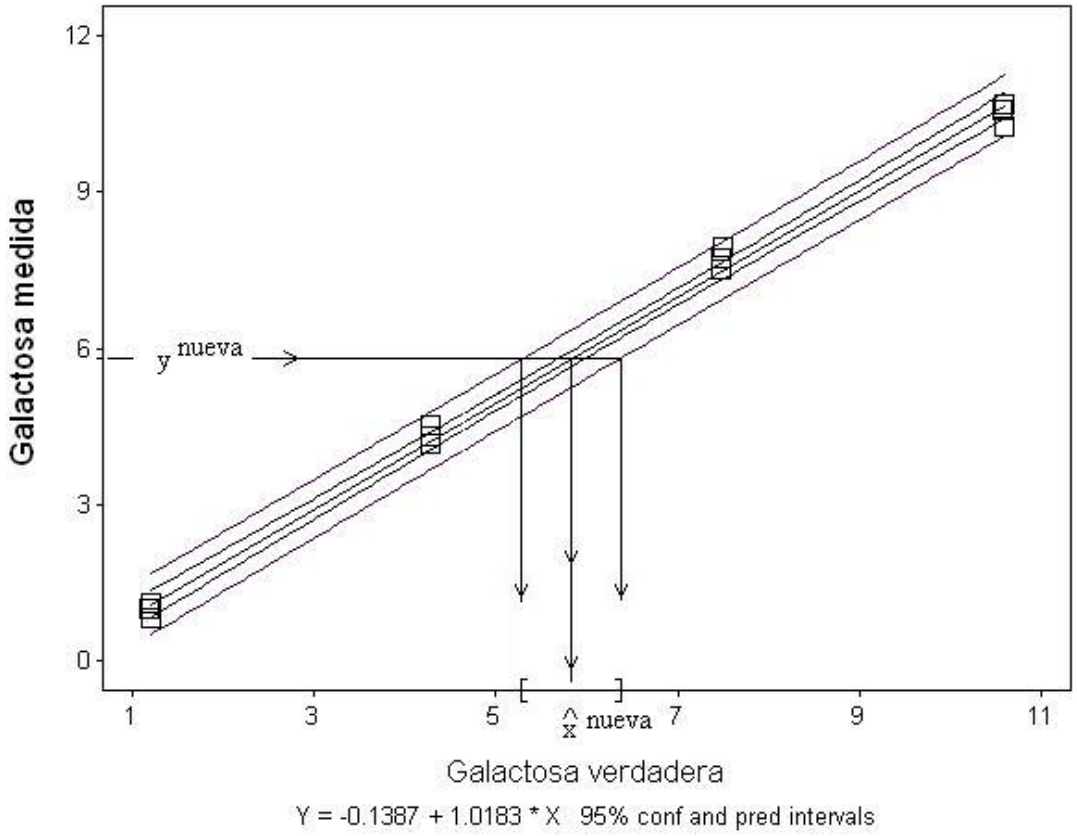
$$s(\text{pred } \hat{x}^{\text{nueva}}) = \frac{\hat{\sigma}}{b}\sqrt{1 + \frac{1}{n} + \frac{(\hat{x}^{\text{nueva}} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \frac{0.23637}{1.01835}\sqrt{1 + \frac{1}{12} + \frac{(5.85 - 5.90)^2}{147.895}}$$

$= 0.2416$

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

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) y está representado en la siguiente figura.



29.5.2 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 $y_1^{nueva},\dots,y_m^{nueva}$ para un mismo valor desconocido x^{nueva} y estimarlo utilizando $\bar{y}_m^{nueva}=\frac{\sum_{i=1}^{m} y_i^{nueva}}{m}$:

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

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

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

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

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

Volvamos al ejemplo

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)

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

=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} $1 \le i \le m$).
- 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.