5 – REGRESIÓN LINEAL SIMPLE

5.1 – Introducción

En muchos problemas existe una relación entre dos o más variables, y resulta de interés estudiar la naturaleza de esa relación. El ***análisis de regresión*** es la técnica estadística para el modelado y la investigación de la relación entre dos o más variables. Veamos un ejemplo.
Los resortes se usan en aplicaciones por su capacidad para alargarse (contraerse) bajo carga. La rigidez de un resorte se mide con la *constante del resorte*, que es la longitud del resorte que se alargará por unidad de la fuerza o de la carga. Para asegurarse de que un resorte dado funciona adecuadamente es necesario calcular la constante de resorte con exactitud y precisión.
En este experimento hipotético un resorte se cuelga verticalmente con un extremo fijo, y los pesos se cuelgan uno tras otro del otro extremo. Después de colgar cada peso se mide la longitud del resorte. Sean $x_1, x_2, \ldots, x_n$ los pesos, y sea l_i la longitud del resorte bajo la carga x_i.
La ley de Hooke establece que

$$l_i = \beta_0 + \beta_1 x_i$$

donde β_0 representa la longitud del resorte cuando no tiene carga y β_1 es la constante del resorte.
Sea y_i la longitud ***medida*** del resorte bajo la carga x_i. Debido al error de medición y_i será diferente de la longitud verdadera l_i. Se escribe como

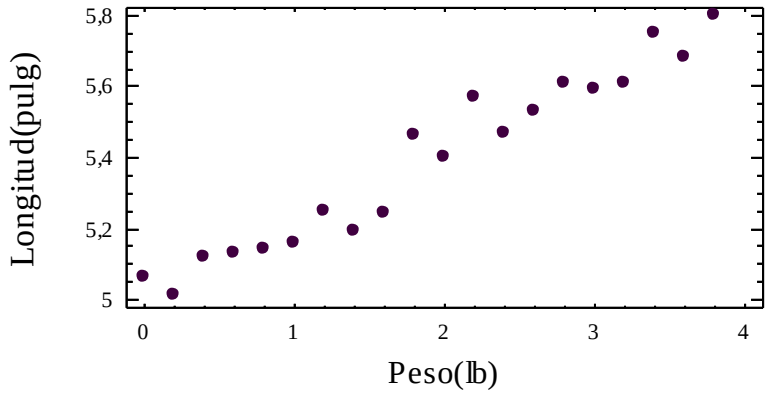
$$y_i = l_i + \varepsilon_i$$

donde ε_i es el error en la *i*-ésima medición. Al combinar ambas ecuaciones se obtiene

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad (10.1)$$

En la ecuación (10.1), y_i es la ***variable dependiente***, x_i es la ***variable independiente***, β_0 y β_1 son los ***coeficientes de regresión***, y ε_i se denomina ***error.*** A la ecuación (10.1) se la llama ***modelo de regresión lineal simple***.
La tabla siguiente presenta los resultados del experimento y la figura el ***diagrama de dispersión*** de *y* contra *x*.

Peso (lb)	***Longitud medida (pulg)***	***Peso (lb)***	***Longitud medida (pulg)***
x	***y***	***x***	***y***
0,0	5,06	2,0	5,40
0,2	5,01	2,2	5,57
0,4	5,12	2,4	5,47
0,6	5,13	2,6	5,53
0,8	5,14	2,8	5,61
1,0	5,16	3,0	5,59
1,2	5,25	3,2	5,61
1,4	5,19	3,4	5,75
1,6	5,24	3,6	5,68
1,8	5,46	3,8	5,80



La idea es utilizar estos datos para ***estimar*** los coeficientes de regresión. Si no hubiese error en la medición, los puntos se encontrarían en una línea recta con pendiente β_1 y ordenada al origen β_0, y estas cantidades serían fáciles de determinar. La idea es entonces que los puntos están dispersos de manera aleatoria alrededor de una recta que es la recta de regresión lineal $l = \beta_0 + \beta_1 x$.

En general podemos decir que al fijar el valor de *x* observamos el valor de la variable *Y*. Si bien *x* es fijo, el valor de Y está afectado por el ***error aleatorio*** ε. Por lo tanto ε ***determina las propiedades de Y***. Escribimos en general

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

donde *x* es, por ahora, una variable no aleatoria, ε ***es la v.a. del error*** y ***asumimos que***

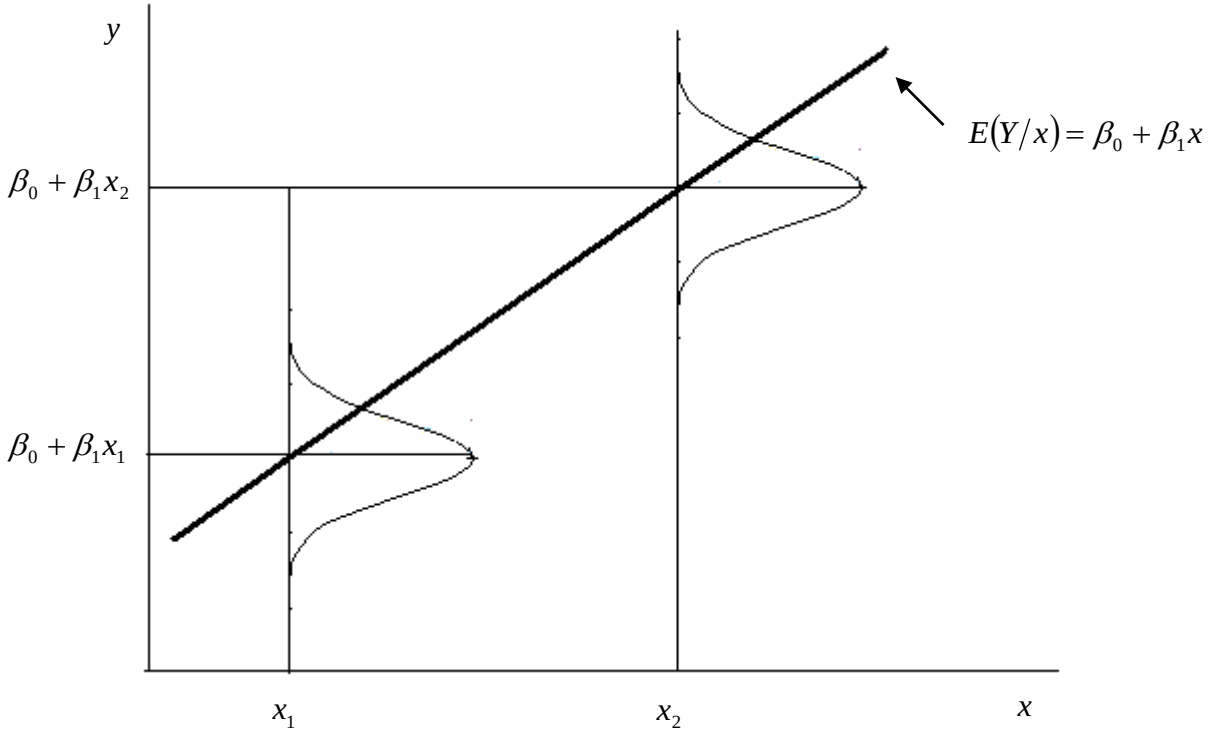
$$E(\varepsilon) = 0 \qquad \text{y} \qquad V(\varepsilon) = \sigma^2$$

Entonces Y es una variable aleatoria tal que

$$E(Y/x) = E(\beta_0 + \beta_1 x + \varepsilon) = \beta_0 + \beta_1 x + E(\varepsilon) = \beta_0 + \beta_1 x$$

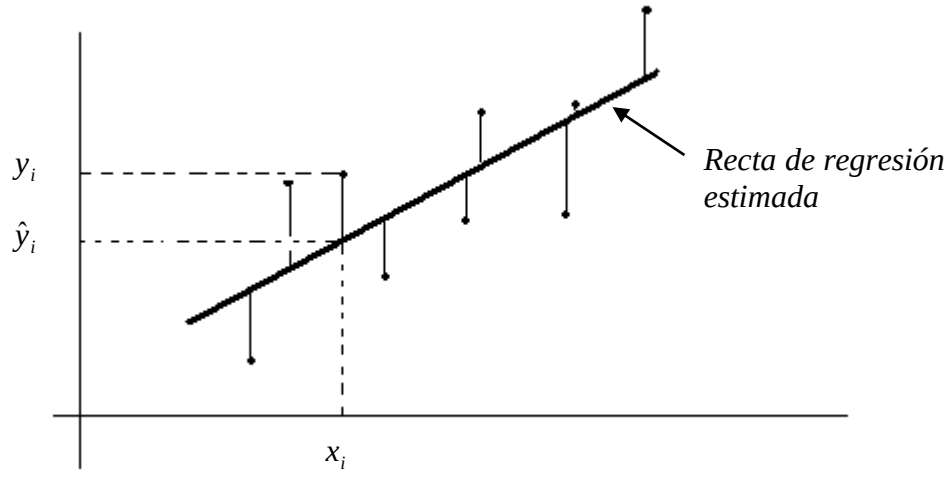
$$V(Y/x) = V(\beta_0 + \beta_1 x + \varepsilon) = V(\varepsilon) = \sigma^2$$

En consecuencia, el modelo de regresión verdadero $E(Y/x) = \beta_0 + \beta_1 x$ es una recta de valores promedio.
Notar que lo anterior implica que existe una distribución de valores de *Y* para cada *x*, y que la varianza de esta distribución es la misma para cada *x*. La siguiente figura ilustra esta situación
Notar que se utilizó una distribución normal para describir la variación aleatoria en ε. Por lo tanto la distribución de *Y* también será normal. La varianza σ^2 determina la variabilidad en las observaciones *Y*. por lo tanto, cuando σ^2 es pequeño, los valores observados de *Y* caen cerca de la línea, y cuando σ^2 es grande, los valores observados de *Y* pueden desviarse considerablemente de la línea. Dado que σ^2 es constante, la variabilidad en *Y* para cualquier valor de *x* es la misma.



5.2 – Regresión lineal simple- Estimación de parámetros

Para estimar los coeficientes de regresión se utiliza el ***método de mínimos cuadrados.*** Supongamos que se tienen *n* pares de observaciones $(x_1, y_1);(x_2, y_2);....;(x_n, y_n)$. Realizamos una gráfica representativa de los datos y una recta como posible recta de regresión Anotamos a la ***recta de regresión estimada*** con $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$



Las estimaciones de β_0 y β_1 deben dar como resultado una línea que en algún sentido se "ajuste mejor" a los datos. El método de mínimos cuadrados consiste en estimar β_0 y β_1 de manera tal que se minimice la suma de los cuadrados de las desviaciones verticales mostradas en la figura anterior.
La suma de los cuadrados de las desviaciones de las observaciones con respecto a la recta de regresión es

$$L = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$$

Los estimadores de mínimos cuadrados de β_0 y β_1, que anotamos $\hat{\beta}_0$ y $\hat{\beta}_1$, deben satisfacer las siguientes ecuaciones

$$\begin{cases} \dfrac{\partial L}{\partial \beta_0} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0 \\ \dfrac{\partial L}{\partial \beta_1} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) x_i = 0 \end{cases} \qquad (10.2)$$

Después de simplificar las expresiones anteriores, se llega a

$$\begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases} \qquad (10.3)$$

Las ecuaciones (10.3) reciben el nombre de ***ecuaciones normales de mínimos cuadrados***.
La solución de estas ecuaciones dan como resultado las ***estimaciones de mínimos cuadrados*** $\hat{\beta}_0$ y $\hat{\beta}_1$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad (10.4)$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \dfrac{\left(\sum_{i=1}^{n} x_i \right) \left(\sum_{i=1}^{n} y_i \right)}{n}}{\sum_{i=1}^{n} x_i^2 - \dfrac{\left(\sum_{i=1}^{n} x_i \right)^2}{n}} \qquad (10.5)$$

donde $\bar{y} = \dfrac{\sum_{i=1}^{n} y_i}{n}$ y $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$

Las diferencias $e_i = y_i - \hat{y}_i$ con $i = 1,...,n$ se llaman ***residuos***. El residuo e_i describe el error en el ajuste del modelo en la *i*-ésima observación y_i.
Para agilizar la notación son útiles los siguientes símbolos

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} \qquad (10.6)$$

$$S_{xy} = \sum_{i=1}^{n} y_i (x_i - \bar{x}) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n} \qquad (10.7)$$

Entonces con esta notación podemos escribir $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$

Ejemplo:
Ajustamos un modelo de regresión lineal a los datos del ejemplo anterior. La estimación de la constante del resorte es $\hat{\beta}_1$ y $\hat{\beta}_0$ la estimación de la longitud sin carga.
De la tabla obtenemos

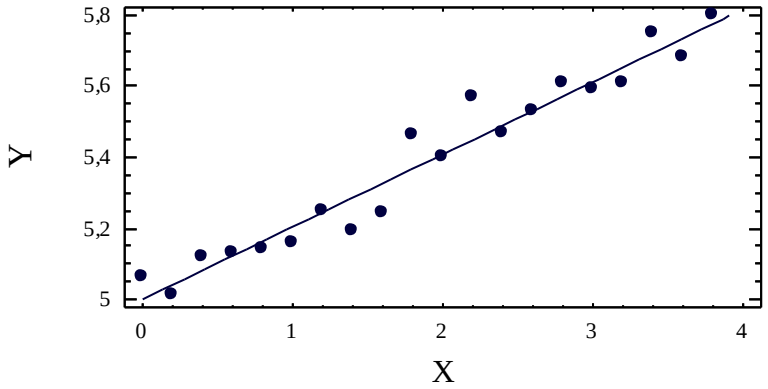
$\bar{x} = 1.9 \quad \bar{y} = 5.3885$
$S_{xx} = 26.6 \quad S_{xy} = 5.4430$

Entonces $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{5.4430}{26.6} = 0.2046$ y $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 5.3885 - 0.2046 \times 1.9 = 4.9997$

La ecuación de la recta estimada es

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad \Rightarrow \quad \hat{y} = 4.9997 - 0.2046x$

La figura siguiente muestra el gráfico de dispersión con la recta de regresión estimada



Podemos utilizar la recta de regresión estimada para predecir la longitud del resorte bajo una carga determinada, por ejemplo con una carga de 1.3 lb:

$\hat{y} = 4.9997 - 0.2046(1.3) = 5.27\text{pulg.}$

Podemos también estimar la longitud del resorte bajo una carga de 1.4 lb:

$\hat{y} = 4.9997 - 0.2046(1.4) = 5.29\text{pulg.}$

Notar que la longitud medida para una carga de 1.4 lb es 5.19 pulg., pero la estimación de mínimos cuadrados de 5.29 pulg. Está basada en todos los datos y es más precisa (tiene menor incertidumbre). Más adelante calcularemos la varianza de estos estimadores.

Observaciones:

1- Las estimaciones de mínimos cuadrados $\hat{\beta}_1$ y $\hat{\beta}_0$ son valores de variables aleatorias y dicho valor varía con las muestras. Los coeficientes de regresión β_0 y β_1 son constantes desconocidas que estimamos con $\hat{\beta}_1$ y $\hat{\beta}_0$.

2- Los residuos e_i no son lo mismo que los errores ε_i . Cada residuo es la diferencia $e_i = y_i - \hat{y}_i$ entre el valor observado y el valor ajustado, y se pueden calcular a partir de los datos. Los errores ε_i representan la diferencia entre los valores medidos y_i y los valores $\beta_0 + \beta_1 x_i$. Como los valores verdaderos de β_0 y β_1 no se conocen entonces, los errores no se pueden calcular.

3- ¿Qué sucede si se quisiera estimar la longitud del resorte bajo una carga de 100 lb? La estimación de mínimos cuadrados es $\hat{y} = 4.9997 - 0.2046(100) = 25.46\text{pulg.}$ Pero esta estimación no es confiable, pues ninguno de los pesos en el conjunto de datos es tan grande. Es probable que el resorte se deformara, por lo que la ley de Hooke no valdría. Para muchas variables las relaciones lineales valen dentro de cierto rango, pero no fuera de él. Si se quiere saber cómo respondería el resorte a una carga de 100 lb se deben incluir pesos de 100 lb o mayores en el conjunto de datos. Por lo tanto ***no hay que extrapolar una recta ajustada fuera del rango de los datos***. La relación lineal puede no ser válida ahí.

5.3 – Propiedades de los estimadores de mínimos cuadrados y estimación de σ^2

Los ***estimadores*** de β_1 y β_0 los anotamos

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} \qquad \hat{\beta}_1 = \frac{S_{xY}}{S_{xx}} = \frac{\sum_{i=1}^{n} Y_i (x_i - \bar{x})}{S_{xx}} \qquad (10.8)$$

Como $\hat{\beta}_1$ y $\hat{\beta}_0$ son estimadores de β_1 y β_0 respectivamente, son variables aleatorias, por lo tanto podemos calcular su esperanza y varianza. Como estamos asumiendo que *x* no es v.a. entonces $\hat{\beta}_1$ y $\hat{\beta}_0$ son funciones de la v.a. *Y*.

Recordemos que el modelo es $Y = \beta_0 + \beta_1 x + \varepsilon$, si medimos *n* veces la variable *Y* tenemos

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

donde ***asumimos*** $E(\varepsilon_i) = 0$; $V(\varepsilon_i) = \sigma^2$ $i = 1,2,\ldots,n$ y $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ ***independientes***

Por lo tanto

$$E(Y/x_i) = E(\beta_0 + \beta_1 x_i + \varepsilon) = \beta_0 + \beta_1 x_i + E(\varepsilon) = \beta_0 + \beta_1 x_i$$

$$V(Y/x_i) = V(\beta_0 + \beta_1 x_i + \varepsilon_i) = V(\varepsilon_i) = \sigma^2$$

Consideramos $\hat{\beta}_1 = \dfrac{S_{xY}}{S_{xx}} = \dfrac{\sum_{i=1}^{n} Y_i (x_i - \bar{x})}{S_{xx}}$. Podemos ver a $\hat{\beta}_1$ como una combinación lineal de las variables Y_i, entonces

$$E(\hat{\beta}_1) = E\left(\frac{S_{xy}}{S_{xx}}\right) = E\left(\frac{\sum_{i=1}^{n} Y_i (x_i - \bar{x})}{S_{xx}}\right) = \frac{1}{S_{xx}} E\left(\sum_{i=1}^{n} Y_i (x_i - \bar{x})\right) = \frac{1}{S_{xx}} \sum_{i=1}^{n} E(Y_i)(x_i - \bar{x}) =$$

$$= \frac{1}{S_{xx}} \sum_{i=1}^{n} (\beta_0 + \beta_1 x_i)(x_i - \bar{x}) = \frac{1}{S_{xx}} \left\{ \beta_0 \sum_{i=1}^{n} (x_i - \bar{x}) + \beta_1 \sum_{i=1}^{n} x_i (x_i - \bar{x}) \right\} = \frac{1}{S_{xx}} \beta_1 S_{xx} = \beta_1$$

Notar que $\sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n\bar{x} = \sum_{i=1}^{n} x_i - n\left(\dfrac{\sum_{i=1}^{n} x_i}{n}\right) = 0$

y $\quad \sum_{i=1}^{n} x_i (x_i - \bar{x}) = \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x}) = S_{xx}$

Por lo tanto $\hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}} = \dfrac{\sum_{i=1}^{n} Y_i (x_i - \bar{x})}{S_{xx}}$ es un estimador insesgados de β_1

Veamos ahora la varianza de $\hat{\beta}_1$

$$V(\hat{\beta}_1) = V\left(\frac{S_{xy}}{S_{xx}}\right) = V\left(\frac{\sum_{i=1}^{n} Y_i (x_i - \bar{x})}{S_{xx}}\right) = \frac{1}{S_{xx}^2} V\left(\sum_{i=1}^{n} Y_i (x_i - \bar{x})\right) = \frac{1}{S_{xx}^2} \sum_{i=1}^{n} V(Y_i)(x_i - \bar{x})^2 =$$

$$= \frac{1}{S_{xx}^2} \sum_{i=1}^{n} \sigma^2 (x_i - \bar{x})^2 = \frac{1}{S_{xx}^2} \sigma^2 S_{xx} = \frac{\sigma^2}{S_{xx}}$$

Por lo tanto

$$\boxed{E(\hat{\beta}_1) = \beta_1 \quad \text{y} \quad V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}} \qquad (10.9)$$

Con un enfoque similar calculamos la esperanza y la varianza de $\hat{\beta}_0$

$$E(\hat{\beta}_0) = E(\bar{Y} - \hat{\beta}_1 \bar{x}) = E(\bar{y}) - E(\hat{\beta}_1)\bar{x} = E\left(\frac{\sum_{i=1}^{n} Y_i}{n}\right) - \beta_1 \bar{x} = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) - \beta_1 \bar{x} =$$

$$= \frac{1}{n}\sum_{i=}^{n} (\beta_0 + \beta_1 x_i) - \beta_1 \bar{x} = \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} = \beta_0$$

Calculamos la varianza de $\hat{\beta}_0$, para esto planteamos:

$$V(\hat{\beta}_0) = V(\bar{Y} - \hat{\beta}_1 \bar{x}) = V(\bar{Y}) + V(\hat{\beta}_1)(\bar{x}) - 2Cov(\bar{Y}, \hat{\beta}_1 \bar{x})$$

Tenemos que

$$V(\vec{Y}) = V\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(Y_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}$$

Y $Cov(Y_i, Y_j) = 0$ por indep.

$$Cov(\bar{Y}, \hat{\beta}_1 \bar{x}) = \bar{x} Cov(\bar{Y}, \hat{\beta}_1) = \bar{x} Cov\left(\frac{1}{n}\sum_{i=1}^{n} Y_i, \sum_{i=1}^{n} Y_i \frac{(x_i - \bar{x})}{S_{xx}}\right) = \bar{x}\frac{1}{nS_{xx}}\sum_{i=1}^{n} Cov(Y_i, Y_i(x_i - \bar{x})) =$$

$$= \bar{x}\frac{1}{nS_{xx}}\sum_{i=1}^{n} (x_i - \bar{x}) Cov(Y_i, Y_i) = \bar{x}\frac{1}{nS_{xx}}\sum_{i=1}^{n} (x_i - \bar{x})\sigma^2 = \bar{x}\frac{\sigma^2}{nS_{xx}}\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

Por lo tanto

$$V(\hat{\beta}_0) = V(\bar{Y} - \hat{\beta}_1 \bar{x}) = V(\bar{Y}) + V(\hat{\beta}_1)(\bar{x})^2 - 2Cov(\bar{Y}, \hat{\beta}_1 \bar{x}) = \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{S_{xx}} - 0 = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$$

Entonces

$$\boxed{E(\hat{\beta}_0) = \beta_0 \qquad \text{y} \qquad V(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)} \quad (10.10)$$

Necesitamos estimar la varianza desconocida σ^2 que aparece en las expresiones de $V(\hat{\beta}_0)$ y $V(\hat{\beta}_1)$.

Los residuos $e_i = y_i - \hat{y}_i$ se emplean para estimar σ^2. La suma de los cuadrados de los residuos es

$$SS_R = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (10.11)$$

Puede demostrarse que $E\left(\frac{SS_R}{\sigma^2}\right) = n-2$, en consecuencia $E\left(\frac{SS_R}{n-2}\right) = \sigma^2$.

Entonces se toma como estimador de σ^2 a

$$\hat{\sigma}^2 = \frac{SS_R}{n-2} \qquad (10.12)$$

Puede obtenerse una fórmula más conveniente para el cálculo de SS_R, para esto primero notar que las ecuaciones normales (10.2) se pueden escribir como

$$\begin{cases} \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0 \\ \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)x_i = 0 \end{cases} \Rightarrow \begin{cases} \sum_{i=1}^{n} e_i = 0 \\ \sum_{i=1}^{n} e_i x_i = 0 \end{cases}$$

Entonces

$$SS_R = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)(y_i - \hat{y}_i) = \sum_{i=1}^{n} e_i(y_i - \hat{y}_i) = \sum_{i=1}^{n} e_i\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) =$$
$$= \sum_{i=1}^{n} e_i\left(y_i - \hat{\beta}_0\right) - \sum \hat{\beta}_1 e_i x_i = \sum_{i=1}^{n} e_i\left(y_i - \hat{\beta}_0\right) = \sum_{i=1}^{n} e_i\left(y_i - \bar{y} - \hat{\beta}_1\bar{x}\right) = \sum_{i=1}^{n} e_i(y_i - \bar{y})$$

Por lo tanto

$$SS_R = \sum_{i=1}^{n} e_i(y_i - \bar{y}) = \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)(y_i - \bar{y}) = \sum_{i=1}^{n}\left(y_i - \bar{y} + \hat{\beta}_1\bar{x} - \hat{\beta}_1 x_i\right)(y_i - \bar{y}) =$$
$$= \sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y}) - \sum_{i=1}^{n}\hat{\beta}_1(x_i - \bar{x})(y_i - \bar{y}) = S_{yy} - \hat{\beta}_1 S_{xy}$$

También se puede escribir

$$SS_R = S_{yy} - \hat{\beta}_1 S_{xy} = S_{yy} - \frac{S_{xy}}{S_{xx}} S_{xy} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

En resumen $SS_R = S_{yy} - \hat{\beta}_1 S_{xy}$ ó $SS_R = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$ (10.13)

Por lo tanto $\hat{\sigma}^2 = \frac{S_{yy} - \frac{S_{xy}^2}{S_{xx}}}{n-2}$

Y si anotamos a la ***desviación estándar estimada de*** $\hat{\beta}_0$ ***y*** $\hat{\beta}_1$ ***con*** $s_{\hat{\beta}_0}$ ***y*** $s_{\hat{\beta}1}$ respectivamente entonces

$$s_{\hat{\beta}_1} = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \quad \text{y} \quad s_{\hat{\beta}_0} = \sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)} \qquad (10.14)$$

Ejemplo:
En el ejemplo anterior se calculó, $\bar{x} = 1.9$, $\bar{y} = 5.3885$, $S_{xx} = 26.6$, $S_{xy} = 5.4430$.

Calculamos ahora $S_{yy} = \sum_{i=1}^{20} (y_i - \bar{y})^2 = 1.1733$ y entonces

$$\hat{\sigma}^2 = \frac{S_{yy} - \frac{S_{xy}^2}{S_{xx}}}{n-2} = \frac{1.1733 - \frac{5.4430^2}{26.6}}{18} = 0.003307$$

$$s_{\hat{\beta}_1} = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = \sqrt{\frac{0.003307}{26.6}} = \sqrt{0.000124} = 0.0111$$

$$s_{\hat{\beta}_0} = \sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)} = \sqrt{0.003307\left(\frac{1}{20} + \frac{1.9^2}{26.6}\right)} = 0.02478219$$

Observación:
La varianza de $\hat{\beta}_0$ ***y*** $\hat{\beta}_1$ ***se puede disminuir*** tomando valores x_i muy dispersos con respecto a $\bar{x}$ pues de esta forma aumenta S_{xx}

Para construir intervalos de confianza para los coeficientes de regresión o para construir pruebas de hipótesis con respecto a β_0 o β_1 necesitamos asumir que ***los errores*** ε_i ***tienen distribución normal.*** Entonces $\varepsilon_i \sim N(0, \sigma^2)$

Observación:
Si $\varepsilon_i \sim N(0, \sigma^2)$ entonces, como $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, resulta que $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$. ***Se pueden calcular entonces los EMV de los parámetros y llegaríamos a que son los mismos que los encontrados usando mínimos cuadrados. De modo que la función que cumple la suposición de normalidad de los*** ε_i ***no es otra que la de justificar el uso del método de mínimos cuadrados, que es el más sencillo de calcular.***

Ya vimos que $\hat{\beta}_0$ y $\hat{\beta}_1$ pueden considerarse combinaciones lineales de las Y_i, por lo tanto $\hat{\beta}_0$ ***y*** $\hat{\beta}_1$ son combinación lineal de variables aleatorias independientes con distribución normal y eso implica que

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right) \quad \text{y} \quad \hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right) \qquad (10.15)$$

Y entonces

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}} \sim N(0,1) \qquad \text{y} \qquad \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0,1) \qquad (10.16)$$

Bajo la suposición que los errores tienen distribución normal, se puede probar que

$$\frac{SS_R}{\sigma^2} \sim \chi^2_{\text{n-2}} \qquad (10.17)$$

Y también se puede probar que

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}} \sim t_{\text{n-2}} \qquad \text{y} \qquad \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}} \sim t_{\text{n-2}} \qquad (10.18)$$

5.4 – Inferencias estadísticas sobre los parámetros de regresión

Suponemos que los errores tiene distribución normal, con media cero, varianza σ^2 y son independientes.

Inferencias sobre β_1

Tests de hipótesis sobre β_1

Se desea probar la hipótesis de que la pendiente β_1 es igual a una constante, por ejemplo β_{10}. Supongamos las hipótesis

$$H_0 : \beta_1 = \beta_{10} \qquad \text{contra} \qquad H_0 : \beta_1 \neq \beta_{10}$$

El estadístico de prueba es $T = \dfrac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\dfrac{\hat{\sigma}^2}{S_{xx}}}}$ que bajo H_0 tiene distribución Student con *n-2* grados de libertad.

Por lo tanto la regla de decisión es $\begin{cases} rechazar\ \ H_0\ \ si\ \ |T| > t_{\frac{\alpha}{2},n-2} \\ aceptar\ \ H_0\ \ si\ \ |T| \leq t_{\frac{\alpha}{2},n-2} \end{cases}$

Si $H_1 : \beta_1 > \beta_{10}$ se rechaza $H_0 : \beta_1 = \beta_{10}$ si $T > t_{\alpha,n-2}$

Si $H_1 : \beta_1 < \beta_{10}$ se rechaza $H_0 : \beta_1 = \beta_{10}$ si $T < -t_{\alpha,n-2}$

Un caso especial importante es cuando $H_0 : \beta_1 = 0$ contra $H_0 : \beta_1 \neq 0$
Estas hipótesis están relacionadas con la ***significancia de la regresión***.
Aceptar $H_0 : \beta_1 = 0$ es equivalente a concluir que no hay ninguna relación lineal entre *x* e *Y*.
Si $H_0 : \beta_1 = 0$ se rechaza implica que *x* tiene importancia al explicar la variabilidad en *Y*. También puede significar que el modelo lineal es adecuado, o que aunque existe efecto lineal pueden obtenerse mejores resultados agregando términos polinomiales de mayor grado en *x*.

Ejemplos:
1- El fabricante del resorte de los datos de la ley de Hooke afirma que la constante del resorte β_1 es al menos 0.23 pulg/lb. Se ha calculado que la constante del resorte es $\hat{\beta}_1 = 0.2046$ pulg/lb. ¿Se puede concluir que la afirmación del fabricante es falsa?

Solución:
Se requiere una prueba de hipótesis para contestar la pregunta. Las hipótesis serían

$$H_0 : \beta_1 = 0.23 \qquad \text{contra} \qquad H_0 : \beta_1 < 0.23$$

El estadístico de prueba es $T = \dfrac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\dfrac{\hat{\sigma}^2}{S_{xx}}}} = \dfrac{\hat{\beta}_1 - 0.23}{\sqrt{\dfrac{\hat{\sigma}^2}{S_{xx}}}}$

Se calculó anteriormente $\sqrt{\dfrac{\hat{\sigma}^2}{S_{xx}}} = 0.0111$, entonces el valor t_0 que toma el estadístico es

$$t_0 = \frac{0.2046 - 0.23}{0.0111} = -2.28$$

Calculamos el p-valor recordando que bajo $H_0 : \beta_1 = 0.23$, $T \sim t_{n-2}$:

$$p - valor = P(T < -2.28)$$

Vemos en la tabla de la distribución Student que en la fila $\nu = 18$ grados de libertad

$$\begin{cases} P(T > 2.101) = 0.025 \\ P(T > 2.552) = 0.01 \end{cases} \quad \Rightarrow \qquad 0.01 < p - valor < 0.025$$

Por lo tanto se rechaza $H_0 : \beta_1 = 0.23$

2- La capacidad de una unión soldada de elongarse bajo tensión está afectada por el compuesto químico del metal de soldadura. En un experimento para determinar el efecto del contenido de carbono (*x*) sobre la elongación (*y*) se alongaron 39 soldaduras hasta la fractura, y se midió tanto el contenido de carbono (en partes por mil) como la elongación (en %). Se calcularon los siguientes resúmenes estadísticos:

$$S_{xx} = 0.6561 \;\; ; \;\; S_{xy} = -3.9097 \;\; ; \;\; \hat{\sigma} = 4.3319$$

Suponiendo que *x* e *y* siguen un modelo lineal, calcular el cambio estimado en la elongación debido a un aumento de una parte por mil en el contenido de carbono. ¿Se debe utilizar el modelo lineal para pronosticar la elongación del contenido de carbono?

Solución:
El modelo lineal es $y = \beta_0 + \beta_1 x + \varepsilon$, y el cambio de elongación debido a un aumento de una parte por mil en el contenido de carbono es β_1.
Las hipótesis serían $H_0 : \beta_1 = 0$ contra $H_0 : \beta_1 \neq 0$

La hipótesis nula establece que incrementar el contenido de carbono no afecta la elongación, mientras que la hipótesis alternativa establece que sí afecta la elongación.

El estadístico de prueba $|T| = \left| \dfrac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\dfrac{\hat{\sigma}^2}{S_{xx}}}} \right| = \left| \dfrac{\hat{\beta}_1}{\sqrt{\dfrac{\hat{\sigma}^2}{S_{xx}}}} \right|$ si $H_0 : \beta_1 = 0$ es verdadera tiene distribución Student con $n-2$ gados de libertad.

Calculamos $\hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}} = \dfrac{\sum_{i=1}^{n} y_i (x_i - \bar{x})}{S_{xx}} = \dfrac{-3.9097}{0.6561} = -5.959$

$$\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = \frac{\hat{\sigma}}{\sqrt{S_{xx}}} = \frac{4.3319}{\sqrt{0.6561}} = 5.348$$

El valor que toma el estadístico de prueba es $t_0 = \left| \dfrac{-5.959}{5.348} \right| = 1.114$

Y $p - valor = P(|T| > 1.114) > 2 \times 0.10 = 0.20$

Por lo tanto no hay evidencia en contra de la hipótesis nula. No se puede concluir que el modelo lineal sea útil para pronosticar la elongación a partir del contenido de carbono.

Intervalos de confianza para β_1

Podemos construir intervalos de confianza para β_1 de nivel $1-\alpha$ utilizando el hecho que el estadístico $\dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{\dfrac{\hat{\sigma}^2}{S_{xx}}}} \sim t_{n-2}$. El intervalo sería

$$\left[\hat{\beta}_1 - t_{\frac{\alpha}{2}, n-2} \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}};\ \hat{\beta}_1 + t_{\frac{\alpha}{2}, n-2} \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \right] \qquad (10.19)$$

Ejemplo:
Determinar un intervalo de confianza de nivel 0.95 para la constante del resorte de los datos de la ley de Hooke.

Solución:

Se calculó antes $\hat{\beta}_1 = 0.2046$ y $\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = 0.0111$

El número de grados de libertad es $20 - 2 = 18$, y $\alpha = 0.05$ por lo tanto
$t_{\frac{\alpha}{2},n-2} = t_{0.025,18} = 2.101$

Por lo tanto el intervalo es

$$\left[0.2046 - 2.101(0.0111);\ 0.2046 + 2.101(0.0111)\right] = \left[0.181;\ 0.228\right]$$

Inferencias sobre β_0

De manera similar a lo visto sobre β_1, se pueden deducir intervalos de confianza y tests de hipótesis para β_0
Específicamente, si tenemos las hipótesis

$$H_0 : \beta_0 = \beta_{00} \qquad \text{contra} \qquad H_0 : \beta_0 \neq \beta_{00}$$

El estadístico de prueba es $T \dfrac{\hat{\beta}_0 - \beta_{00}}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}}$ y bajo $H_0 : \beta_0 = \beta_{00}$ tenemos que $T \sim t_{n-2}$

Por lo tanto la regla de decisión es $\begin{cases} rechazar\ \ H_0\ \ si\ \ |T| > t_{\frac{\alpha}{2},n-2} \\ aceptar\ \ H_0\ \ si\ \ |T| \leq t_{\frac{\alpha}{2},n-2} \end{cases}$

Si $H_1 : \beta_0 > \beta_{00}$ se rechaza $H_0 : \beta_0 = \beta_{00}$ si $T > t_{\alpha,n-2}$
Si $H_1 : \beta_0 < \beta_{00}$ se rechaza $H_0 : \beta_0 = \beta_{00}$ si $T < -t_{\alpha,n-2}$

Intervalos de confianza de nivel $1-\alpha$ se deducen de manera análoga a lo visto anteriormente, donde usamos el hecho que el estadístico $T \dfrac{\hat{\beta}_0 - \beta_0}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \sim t_{n-2}$

El intervalo es $\left[\hat{\beta}_0 - t_{\frac{\alpha}{2},n-2}\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)};\ \hat{\beta}_0 + t_{\frac{\alpha}{2},n-2}\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}\right]$ (10.20)

Ejemplo:
En los datos de la ley de Hooke determine un intervalo de confianza de nivel 0.99 para la longitud del resorte no cargado.
Solución:
La longitud del resorte no cargado es β_0. Se ha calculado anteriormente $\hat{\beta}_0 = 4.9997$ y

$$s_{\hat{\beta}_0} = \sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)} = 0.02478219$$

El número de gados de libertad es $20 - 2 = 18$ y como $\alpha = 0.01$ entonces
$t_{\frac{\alpha}{2},n-2} = t_{0.005,18} = 2.878$

Por lo tanto el intervalo es
$$\left[4.9997 - 2.878(0.024782193);\ 4.9997 + 2.878(0.02478219)\right] = \left[4.9283;\ 5.071023\right]$$

5.5 – Intervalo de confianza para la respuesta media

A menudo es de interés ***estimar*** mediante un intervalo de confianza $\beta_0 + \beta_1 x_0$, es decir estimar la media $E(Y/x_0)$ para un valor específico x_0.

Un estimador puntual razonable para $\beta_0 + \beta_1 x_0$ es $\hat{\beta}_0 + \hat{\beta}_1 x_0$.

Sabemos que $E\left(\hat{\beta}_0 + \hat{\beta}_1 x_0\right) = \beta_0 + \beta_1 x_0$.

Como de costumbre necesitamos construir un estadístico a partir de $\hat{\beta}_0 + \hat{\beta}_1 x_0$ que contenga al parámetro de interés, (en este caso $\beta_0 + \beta_1 x_0$) y del cual conozcamos la distribución de probabilidad.

Pensamos en el estadístico $\dfrac{\hat{\beta}_0 + \hat{\beta}_1 x_0 - E\left(\hat{\beta}_0 + \hat{\beta}_1 x_0\right)}{\sqrt{V\left(\hat{\beta}_0 + \hat{\beta}_1 x_0\right)}}$

Nos falta calcular $V\left(\hat{\beta}_0 + \hat{\beta}_1 x_0\right)$. Para esto nuevamente observamos que $\hat{\beta}_0 + \hat{\beta}_1 x_0$ es una combinación lineal de las variables Y_i

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 = \bar{Y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_0 = \frac{1}{n}\sum_{i=1}^{n} Y_i + \hat{\beta}_1 (x_0 - \bar{x}) = \frac{1}{n}\sum_{i=1}^{n} Y_i + \frac{S_{xY}}{S_{xx}}(x_0 - \bar{x}) =$$

$$= \frac{1}{n}\sum_{i=1}^{n} Y_i + \frac{\sum_{i=1}^{n} Y_i (x_i - \bar{x})}{S_{xx}}(x_0 - \bar{x}) = \sum_{i=1}^{n} Y_i \left[\frac{1}{n} + \frac{(x_i - \bar{x})}{S_{xx}}(x_0 - \bar{x})\right]$$

Por lo tanto:

$$V\left(\hat{\beta}_0+\hat{\beta}_1 x_0\right)=V\left(\sum_{i=1}^{n} Y_i\left[\frac{1}{n}+\frac{(x_i-\bar{x})}{S_{xx}}(x_0-\bar{x})\right]\right)=\sum_{i=1}^{n} V(Y_i)\left[\frac{1}{n}+\frac{(x_i-\bar{x})}{S_{xx}}(x_0-\bar{x})\right]^2=$$

$$=\sum_{i=1}^{n}\sigma^2\left[\frac{1}{n}+\frac{(x_i-\bar{x})}{S_{xx}}(x_0-\bar{x})\right]^2=\sum_{i=1}^{n}\sigma^2\left[\frac{1}{n^2}+\frac{(x_i-\bar{x})^2}{S_{xx}^2}(x_0-\bar{x})^2+2\frac{(x_i-\bar{x})}{nS_{xx}}(x_0-\bar{x})\right]=$$

$$=\sigma^2\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}^2}\sum_{i=1}^{n}(x_i-\bar{x})^2+2\frac{(x_0-\bar{x})}{nS_{xx}}\sum_{i=1}^{n}(x_i-\bar{x})\right]=$$

Notar que $\sum_{i=1}^{n}(x_i-\bar{x})=0$ y $\sum_{i=1}^{n}(x_i-\bar{x})^2=S_{xx}$ entonces

$$=\sigma^2\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]$$

Por lo tanto

$$\hat{\beta}_0+\hat{\beta}_1 x_0 \sim N\left(\beta_0+\beta_1 x_0;\ \sigma^2\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]\right) \qquad (10.21)$$

Como σ^2 es desconocido lo reemplazamos por $\hat{\sigma}^2=\frac{SS_R}{n-2}$, y puede probarse que

$\frac{\hat{\beta}_0+\hat{\beta}_1 x_0-(\beta_0+\beta_1 x_0)}{\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]}}$ tiene distribución Student con $n-2$ grados de libertad

Razonando como en casos anteriores, el intervalo de confianza para $\beta_0+\beta_1 x_0$ de nivel $1-\alpha$ es

$$\left[\hat{\beta}_0+\hat{\beta}_1 x_0-t_{\frac{\alpha}{2},n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]};\ \hat{\beta}_0+\hat{\beta}_1 x_0+t_{\frac{\alpha}{2},n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]}\right] \qquad (10.22)$$

Ejemplo:
Mediante los datos de la ley de Hooke calcular un intervalo de confianza de nivel 0.95 para la longitud media de un resorte bajo una carga de 1.4 lb

Solución:
Para aplicar (10.22) necesitamos calcular $\hat{\beta}_0+\hat{\beta}_1 x_0$; $\hat{\sigma}^2$; $\bar{x}$; S_{xx}.
En este caso $x_0=1.4$ y $\alpha=0.05$, por lo tanto $t_{\frac{\alpha}{2},n-2}=t_{0.025,18}=2.101$

Ya tenemos calculado de ejemplos anteriores:
$\hat{\sigma}=0.0575$
$\bar{x}=1.9$

$S_{xx} = 26.6$

$\hat{\beta}_0 = 4.9997$ y $\hat{\beta}_1 = 0.2046$

De aquí ya calculamos $\hat{\beta}_0 + \hat{\beta}_1 x_0 = 4.9997 + 0.2046 \times 1.4 = 5.286$

Entonces el intervalo es:

$$\left[5.286 - 2.101\sqrt{0.0575^2\left[\frac{1}{20} + \frac{(1.4-1.9)^2}{26.6}\right]};\ 5.286 + 2.101\sqrt{0.0575^2\left[\frac{1}{20} + \frac{(1.4-1.9)^2}{26.6}\right]} \right] =$$

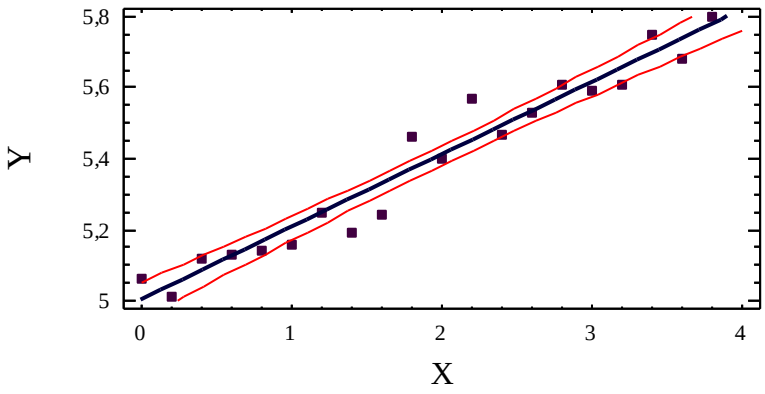
$$= [5.26;\ \ 5.32]$$

Observaciones:

1- Notar que el ancho del intervalo de confianza para $E(Y/x_0)$ depende del valor de x_0. El ancho del intervalo es mínimo cuando $x_0 = \bar{x}$ y crece a medida que $|x_0 - \bar{x}|$ aumenta.

2- Al repetir los cálculos anteriores para varios valores diferentes de x_0 pueden obtenerse intervalos de confianza para cada valor correspondiente de $E(Y/x_0)$.

En la figura siguiente se presenta el diagrama de dispersión con la recta estimada y los correspondientes intervalos de confianza de nivel 0.95 graficados con las líneas inferior y superior referidos al ejemplo anterior. Se origina entonces una ***banda de confianza*** que envuelve a la recta estimada.



5.6 – Intervalos de predicción para futuras observaciones

Una aplicación importante de un modelo de regresión es la predicción de observaciones nuevas o futuras de *Y*, correspondientes a un nivel especificado de la variable *x*.

Si x_0 es el valor de *x* de interés, entonces una ***estimación puntual*** de la observación $Y_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0$ es $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

Para hallar un intervalo de predicción para $Y_0 = \beta_0 + \beta_1 x_0$ de nivel $1-\alpha$ debemos construir un estadístico a partir de $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

Primero notamos que si Y_0 es una nueva observación, entonces Y_0 ***es independiente de las observaciones utilizadas para desarrollar el modelo de regresión***.

Consideramos $Y_0 - \hat{Y}_0$. Calculamos su esperanza y varianza:

$$E\left(Y_0 - \hat{Y}_0\right) = E\left(\beta_0 + \beta_1 x_0 + \varepsilon_0 - \left(\hat{\beta}_0 + \hat{\beta}_1 x_0\right)\right) = \beta_0 + \beta_1 x_0 + E(\varepsilon_0) - \left(\beta_0 + \beta_1 x_0\right) = 0$$

$$V\left(Y_0 - \hat{Y}_0\right) = V\left(Y_0\right) + V\left(\hat{Y}_0\right) = V\left(\beta_0 + \beta_1 x_0 + \varepsilon_0\right) + V\left(\hat{\beta}_0 + \hat{\beta}_1 x_0\right) = \sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right] =$$

$$= \sigma^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]$$

Por lo tanto

$$Y_0 - \hat{Y}_0 \sim N\left(0;\ \sigma^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]\right) \qquad (10.23)$$

En consecuencia

$$\frac{Y_0 - \hat{Y}_0}{\sqrt{\sigma^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}} \sim N(0;\ 1) \qquad (10.24)$$

Si reemplazamos σ^2 por su estimación $\hat{\sigma}^2$ se puede probar que

$$\frac{Y_0 - \hat{Y}_0}{\sqrt{\hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}} \sim t_{n-2} \qquad (10.25)$$

Por el argumento usual llegamos al siguiente ***intervalo de predicción de nivel*** $1-\alpha$ ***para*** Y_0***:***

$$\left[\hat{Y}_0 - t_{\frac{\alpha}{2},n-2}\,\hat{\sigma}\sqrt{\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]};\ \hat{Y}_0 + t_{\frac{\alpha}{2},n-2}\,\hat{\sigma}\sqrt{\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}\right] \qquad (10.26)$$

Ejemplo:

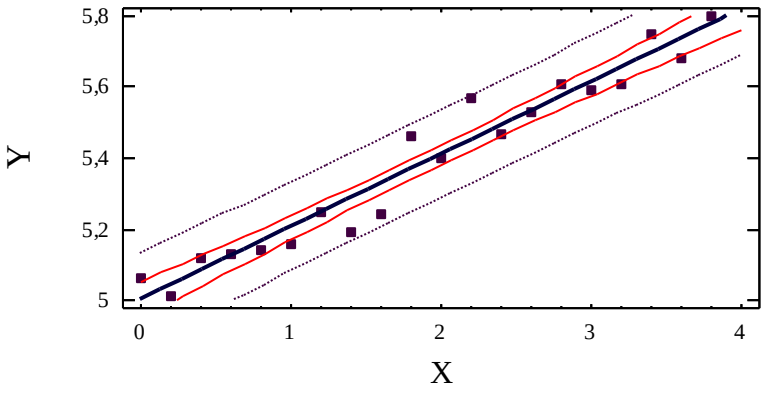
Calcular el intervalo de predicción con nivel 0.95 para la elongación de un resorte bajo una carga de 1.4 lb.

Solución:
El intervalo es

$$\left[5.286-2.101\sqrt{0.0575^2\left[1+\frac{1}{20}+\frac{(1.4-1.9)^2}{26.6}\right]};\ 5.286+2.101\sqrt{0.0575^2\left[1+\frac{1}{20}+\frac{(1.4-1.9)^2}{26.6}\right]}\right]=$$

$$=[5.16165;\ \ 5.41034]$$

Observaciones:
1- Un ***intervalo de confianza*** es un intervalo que contiene, con un nivel de confianza fijado, un parámetro determinado de interés. Un ***intervalo de predicción*** es un intervalo que contiene, con un nivel de confianza fijado, una variable aleatoria de interés.
2- El ancho del intervalo de predicción es mínimo cuando $x_0=\bar{x}$ y crece a medida que $|x_0-\bar{x}|$ aumenta.
Al comparar (10.26) con (10.22) se observa que el intervalo de predicción en el punto x_0 siempre es más grande que el intervalo de confianza en x_0. Esto se debe a que el intervalo de predicción depende tanto del error del modelo ajustado como del error asociado con las observaciones futuras.
3- Al repetir los cálculos anteriores para varios valores diferentes de x_0 pueden obtenerse los intervalos de predicción. En la figura siguiente se presenta el diagrama de dispersión con la recta estimada y los correspondientes intervalos de confianza y de predicción de nivel 0.95 graficados con las líneas inferior y superior referidos al ejemplo anterior. Se originan entonces una ***banda de confianza*** (línea continua) y otra ***banda de predicción*** (línea entrecortada) que envuelven a la recta estimada. Esto ilustra que los intervalos de confianza son menos amplios que los intervalos de predicción.



5.7 – Índice de ajuste

Si consideramos el ajuste por mínimos cuadrados de los pares de datos (x_i, Y_i) al modelo

$$Y = \beta_0 + \varepsilon$$

Entonces es fácil verificar que el estimador de mínimos cuadrados de β_0 es $\bar{Y}$, y la suma de residuos al cuadrado es $S_{YY} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$. Por otro lado si consideramos el modelo lineal

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

Entonces tenemos un valor de $SS_R = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ que será menor o igual a $S_{YY} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$.

La cantidad R^2 se define como

$$R^2 = 1 - \frac{SS_R}{S_{YY}} \qquad (10.27)$$

y es llamado ***coeficiente de determinación***. Vemos que R^2 será cero si $\beta_1 = 0$ y será uno si $SS_R = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 0$, lo que significa ajuste lineal perfecto.

En general $0 \le R^2 \le 1$. El valor de R^2 se interpreta como la proporción de variación de la respuesta Y que es explicada por el modelo. La cantidad $\sqrt{R^2}$ es llamada ***índice de ajuste***, y es a menudo usada como un indicador de qué tan bien el modelo de regresión ajusta los datos. Pero un valor alto de R no significa necesariamente que el modelo de regresión sea correcto.

Ejemplo:

En el ejemplo de la ley de Hooke, tenemos $S_{yy} = \sum_{i=1}^{20}(y_i - \bar{y})^2 = 1.1733$, y

$$SS_R = 1.1733 - \frac{5.4430^2}{26.6} = 0.059526$$

Por lo tanto $R^2 = 1 - \frac{SS_R}{S_{YY}} = 1 - \frac{0.059526}{1.1733} = 0.949266$

El índice de ajuste *R* es a menudo llamado ***coeficiente de correlación muestral***. Si la variable fijada *x* es una ***variable aleatoria,*** entonces tendríamos una v.a. bidimensional (X, Y) con una distribución de probabilidad conjunta, y tenemos una muestra de pares (X_i, Y_i) $i = 1,...,n$. Supongamos que estamos interesados en estimar ρ el coeficiente de correlación entre *X* e *Y*. Es decir

$$\rho = \frac{E[(X - E(X))(Y - E(Y))]}{\sqrt{V(X)V(Y)}}$$

Es razonable estimar

$E[(X - E(X))(Y - E(Y))]$ con $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$

$V(X)$ con $\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$ y $V(Y)$ con $\frac{1}{n}\sum_{i=1}^{n}(Y_i-\overline{Y})$

Por lo tanto un estimador natural de ρ es

$$\hat{\rho}=\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2\sum_{i=1}^{n}(Y_i-\overline{Y})^2}}=\frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}=R \qquad (10.28)$$

Es decir ***el índice de ajuste estima la correlación entre X e Y***

Si X es una variable aleatoria, entonces se observan pares independientes (X_i,Y_i) con $i=1,...,n$ que cumplen el modelo

$$Y_i=\beta_0+\beta_1X_i+\varepsilon_i$$

Si asumimos que X_i y ε_i ***son independientes y que las*** ε_i ***tienen todas la misma distribución con*** $E(\varepsilon_i)=0$***, entonces*** $E(Y_i/X_i)=\beta_0+\beta_1X_i$
Si además suponemos que $\varepsilon_i\sim N(0,\sigma^2)$ ***entonces se puede probar que los estimadores de máxima verosimilitud para los parámetros*** β_0 ***y*** β_1 son

$$\hat{\beta}_0=\overline{Y}-\hat{\beta}_1\overline{X}$$

y

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}Y_i(X_i-\overline{X})}{\sum_{i=1}^{n}(X_i-\overline{X})^2}=\frac{S_{XY}}{S_{XX}}$$

Es decir son los mismos estimadores a los dados por el método de mínimos cuadrados en el caso de suponer que *X* es una variable matemática.

También se puede probar que bajo las suposiciones hechas (10.17) y (10.18) ***siguen siendo válidas***.
Las distribuciones de los estimadores dependen ahora de las distribuciones de las X_i. Puede probarse que siguen siendo insesgados, y que su distribución condicional en las X_i es normal, pero en general su distribución no será normal.

5.8 – <u>Análisis de residuos</u>

El ajuste de un modelo de regresión requiere varias suposiciones. La estimación de los parámetros del modelo requiere la suposición de que los errores son variables aleatorias independientes

con media cero y varianza constante. Las pruebas de hipótesis y la estimación de intervalos requieren que los errores estén distribuidos de manera normal. Además se supone que el grado del modelo es correcto, es decir, si se ajusta un modelo de regresión lineal simple, entonces se supone que el fenómeno en realidad se comporta de una manera lineal.
Se debe considerar la validez de estas suposiciones como dudosas y examinar cuán adecuado es el modelo que se propone. A continuación se estudian métodos que son útiles para este propósito.

Los residuos de un modelo de regresión son $e_i = y_i - \hat{y}_i \quad i = 1,2,\ldots,n$. A menudo el análisis de los residuos es útil para verificar la hipótesis de que los errores tienen una distribución que es aproximadamente normal con varianza constante, y también para determinar la utilidad que tiene la adición de más términos al modelo.

Es posible ***estandarizar*** los residuos mediante el cálculo de $\frac{e_i}{\sqrt{\hat{\sigma}^2}} \quad i = 1,2,\ldots,n$.

También se puede probar que la varianza del *i*-ésimo residuo e_i es igual a

$$V(e_i) = \sigma^2\left[1 - \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right)\right]$$

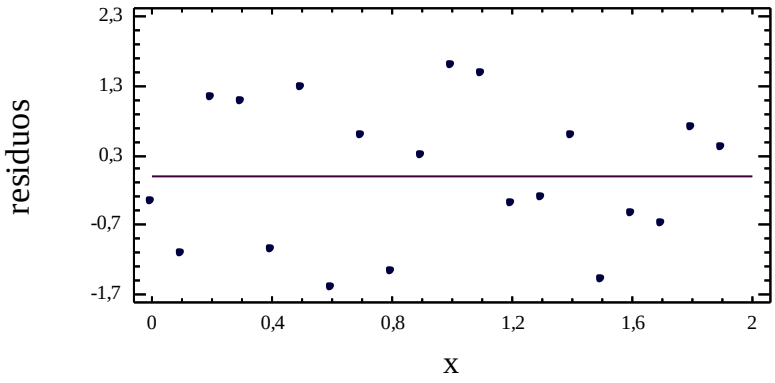
Y entonces podemos considerar al *i*-ésimo ***residuo estudentizado*** que se define como

$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2\left[1 - \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right)\right]}}$$

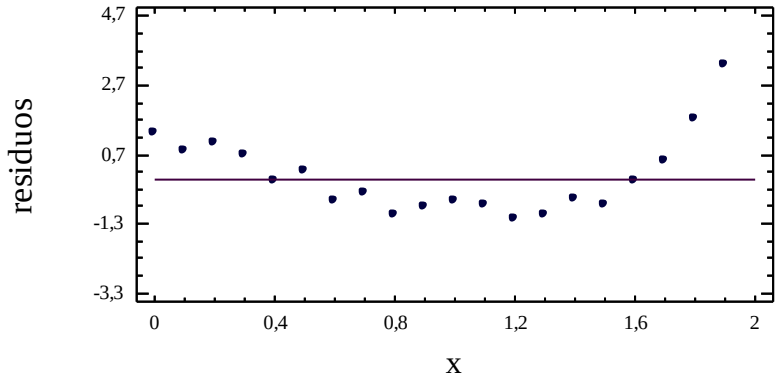
y tiene desviación estándar unitaria.

Si los errores tienen una distribución normal, entonces aproximadamente el 95% de los residuos estandarizados deben caer en el intervalo $(-2;\ 2)$. Los residuos que se alejan mucho de este intervalo pueden indicar la presencia de un ***valor atípico***, es decir, una observación que no es común con respecto a los demás datos.
A menudo es útil hacer una gráfica de residuos contra la variable independiente *x*. En este caso la gráfica tendría que ser una nube de puntos sin ningún patrón en el intervalo $(-2;\ 2)$; pues $e_i = y_i - \hat{y}_i$ sería lo que queda de y_i al quitarle la influencia de x_i. Si en la gráfica aparece algún patrón quiere decir que no estamos quitando de las *y* toda la influencia de las *x*.
Patrones usuales para las gráficas de residuos suelen ser los de las siguientes figuras: en la figura a) se representa la situación ideal, una nube de puntos sin ningún patrón en el intervalo $(-2;\ 2)$. Las figuras b) , c) y d) representan anomalías. Si los residuos aparecen como en b) o c) indican que el modelo es inadecuado. La figura d) muestra residuos que indican que la varianza de las observaciones varía con la magnitud de *x*. Comúnmente se utiliza una transformación de datos sobre la respuesta *y* para eliminar este problema. Las transformaciones más utilizadas para estabilizar la varianza son $\sqrt{y}$, $\ln(y)$ o $1/y$.
En la figura d) la varianza de las observaciones disminuye con el aumento de *x*

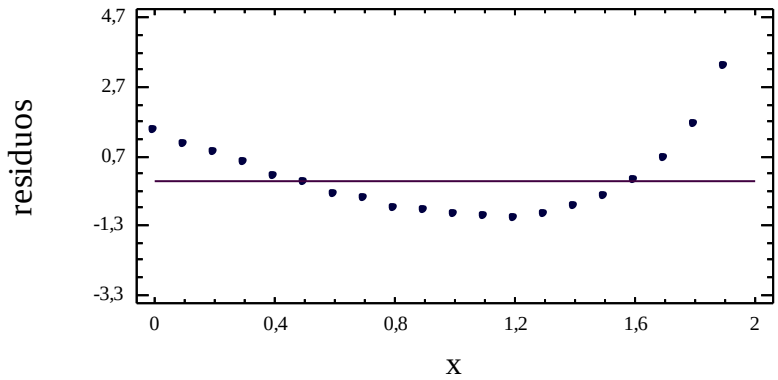
a)


residuos
2,3
1,3
0,3
-0,7
-1,7
0
0,4
0,8
1,2
1,6
2
x

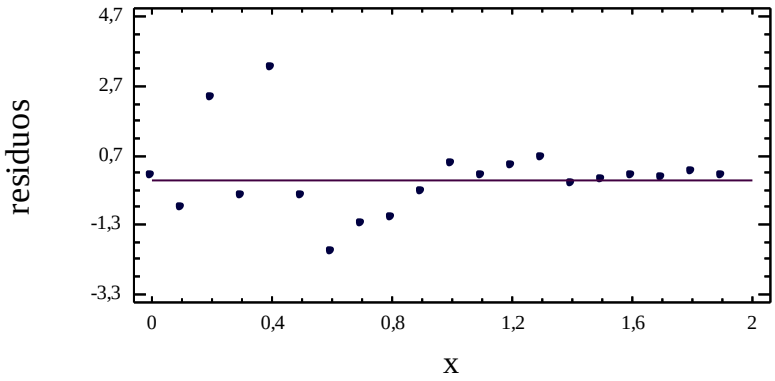
b)


residuos
4,7
2,7
0,7
-1,3
-3,3
0
0,4
0,8
1,2
1,6
2
x

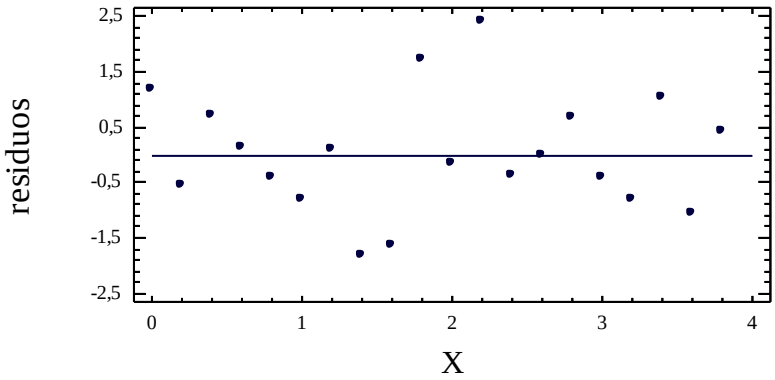
c)


residuos
4,7
2,7
0,7
-1,3
-3,3
0
0,4
0,8
1,2
1,6
2
x

d)



Ejemplo:
Para los datos sobre la ley de Hooke la gráfica de residuos es



Para el caso en que (X,Y) es una v.a. bidimensional, no siempre se está interesado en la relación lineal que defina $E(Y/X)$. Si no, únicamente saber si *X* e *Y* son variables aleatorias independientes. Si asumimos que la distribución conjunta de (X,Y) es una distribución llamada ***normal bivariada***, entonces ***probar que*** $\rho = 0$ ***es equivalente a probar que X e Y son independientes***.
Se puede probar que si la distribución conjunta de (X,Y) es ***normal bivariada***, entonces *R* es el estimador de máxima verosimilitud de ρ. Pero es difícil obtener la distribución de probabilidad para *R*. Se puede superar esta dificultad en muestras bastante grandes al utilizar el hecho que el

estadístico $\left(1/2\right)\ln\left(\frac{1+R}{1-R}\right)$ tiene aproximadamente una distribución normal con media $\mu=\left(1/2\right)\ln\left(\frac{1+\rho}{1-\rho}\right)$ y varianza $\sigma^2=\frac{1}{n-3}$.

Por lo tanto para probar la hipótesis $H_0:\rho=\rho_0$ podemos utilizar el estadístico de prueba

$$Z=\frac{\left(1/2\right)\ln\left(\frac{1+R}{1-R}\right)-\left(1/2\right)\ln\left(\frac{1+\rho_0}{1-\rho_0}\right)}{\frac{1}{\sqrt{n-3}}} \qquad (10.29)$$

Para construir intervalos de confianza de nivel $1-\alpha$ para ρ , se despeja en $\mu=\left(1/2\right)\ln\left(\frac{1+\rho}{1-\rho}\right)$ el coeficiente ρ y se llega a

$$\rho=\frac{e^{2\mu}-1}{e^{2\mu}+1} \qquad (10.30)$$

Ejemplo:
En un estudio de los tiempos de reacción, el tiempo de respuesta a un estímulo visual (*x*) y el tiempo de respuesta a un estímulo auditivo (*y*) se registraron para cada una de 10 personas.
Los tiempos se midieron en minutos. Se presentan en la siguiente tabla.

x	161	203	235	176	201	188	228	211	191	178
y	159	206	241	163	197	193	209	189	169	201

a) Determinar un intervalo de confianza de nivel 0.95 para la correlación entre los tiempos de reacción.
b) Determinar el p-valor para $H_0:\rho=0.3$ contra $H_1:\rho>0.3$

Solución:

a) Se calcula $R=\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2}}=\frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}=0.8159$

Luego calcula $\left(1/2\right)\ln\left(\frac{1+R}{1-R}\right)=\left(1/2\right)\ln\left(\frac{1+0.8159}{1-0.8159}\right)=1.1444$

Como $\left(1/2\right)\ln\left(\frac{1+R}{1-R}\right)$ está distribuido normalmente con varianza $\sigma^2=\frac{1}{n-3}$, el intervalo para $\mu=\left(1/2\right)\ln\left(\frac{1+\rho}{1-\rho}\right)$ es

$$\left[1.1444-1.96\left(\frac{1}{\sqrt{10-3}}\right);\quad 1.1444+1.96\left(\frac{1}{\sqrt{10-3}}\right)\right]=\left[0.4036;\ \ 1.8852\right]$$

Para hallar el intervalo para ρ aplicamos la transformación (10.30) y se obtiene

$$\frac{e^{2(0.4036)}-1}{e^{2(0.4036)}+1}<\rho<\frac{e^{2(1.8852)}-1}{e^{2(1.8852)}+1}\quad\Rightarrow\quad 0.383<\rho<0.955$$

b) Si $H_0:\rho=0.3$ es verdadera entonces el estadístico

$$Z=\frac{\left(1/2\right)\ln\left(\frac{1+R}{1-R}\right)-\left(1/2\right)\ln\left(\frac{1+0.3}{1-0.3}\right)}{\frac{1}{\sqrt{10-3}}}$$ tiene aproximadamente distribución $N(0,1)$

El valor observado de $\left(1/2\right)\ln\left(\frac{1+R}{1-R}\right)$ es 1.1444, por lo tanto el estadístico toma el valor $z_0=2.2088$

Entonces $p-valor=P\left(Z>2.2088\right)\approx 0.0136$. Entonces se rechaza $H_0:\rho=0.3$ y se concluye que $\rho>0.3$

Práctica

Regresión lineal simple

1) Se utiliza regresión lineal para analizar los datos de un estudio donde se investigó la relación que existe entre la temperatura de la superficie de una carretera (x) y la deformación del pavimento (y). El resumen de cantidades es el siguiente:

$n = 20,\ \sum y_i = 12.75,\ \sum y_i^{\,2} = 8.86,\ \sum x_i = 1478,\ \sum x_i^{\,2} = 143215.8,\ \sum x_i y_i = 1083.67$

a) Calcular las estimaciones de mínimos cuadrados de la pendiente y la ordenada al origen. Hacer un gráfico de la recta de regresión, y estimar σ^2.

b) Utilice la ecuación de la recta ajustada para predecir la deformación del pavimento observada cuando la temperatura de la superficie sea 85ºF .

2) Se tiene la siguiente información sobre la relación entre una medida de la corrosión del hierro (Y) y la concentración de $NaPO_4$ (X, en ppm)

X	2.50	5.03	7.60	11.60	13.00	19.60	26.20	33.00	40.00	50.00	55.00
Y	7.68	6.95	6.30	5.75	5.01	1.43	0.93	0.72	0.68	0.65	0.56

a) Construya un gráfico de dispersión de los datos. ¿Parece ser razonable el modelo de regresión lineal?

b) Calcule la ecuación de la recta de regresión estimada, utilícela para pronosticar el valor de la rapidez de corrosión que se observaría para una concentración de 33 ppm, y calcule el residuo correspondiente.

c) Estime la desviación estándar de observaciones alrededor de la recta de regresión verdadera.

d) ¿Se puede concluir que el modelo de regresión lineal simple especifica una relación útil entre las dos variables?. Establezca y pruebe las hipótesis adecuadas al nivel de significación de 0.05.

3) En pruebas diseñadas para medir el efecto de cierto aditivo en el tiempo de secado de pintura se obtuvieron los siguientes datos

Concentración de aditivo (%)	4.0	4.2	4.4	4.6	4.8	5.0	5.2	5.4	5.6	5.8
Tiempo de secado (horas)	8.7	8.8	8.3	8.7	8.1	8.0	8.1	7.7	7.5	7.2

Se tiene el siguiente resumen estadístico:

$\bar{x} = 4.9 \quad \bar{y} = 8.11 \quad S_{xx} = 3.29967 \quad S_{xy} = -2.75 \quad S_{yy} = 2.58894$

a) Realizar un gráfico de dispersión y estimar la recta de regresión lineal

b) Estimar la varianza

c) Pronostique el tiempo de secado para una concentración de 4.4%

d) ¿Puede utilizarse la recta de mínimos cuadrados para pronosticar el tiempo de secado respecto a una concentración de 7%?

e) ¿Para qué concentración pronosticaría un tiempo de secado de 8.2 horas?

4) Un químico está calibrando un espectrómetro que se utilizará para medir la concentración de monóxido de carbono en muestras atmosféricas. Para comprobar la calibración, se miden muestras de concentración conocida.
Las concentraciones verdaderas (x) y las medidas (y) están dadas en la tabla siguiente:

x (ppm)	0	10	20	30	40	50	60	70	80	90	100
Y (ppm)	1	11	21	28	37	48	56	68	75	86	96

Para comprobar la calibración se ajusta un modelo lineal $y = \beta_0 + \beta_1 x + \varepsilon$. Idealmente, el valor de β_0 debe ser 0 y el valor de β_1 debe ser 1.

a) Calcule los estimadores de mínimos cuadrados $\hat{\beta}_0$ y $\hat{\beta}_1$

b) ¿Se puede rechazar la hipótesis nula $H_0 : \beta_0 = 0$?. Utilice $\alpha = 0.05$

c) ¿Se puede rechazar la hipótesis nula $H_0 : \beta_1 = 1$? .Utilice $\alpha = 0.05$

d) ¿Los datos proporcionan suficiente evidencia para concluir que la máquina está fuera de calibración?

5) En un experimento para investigar la relación entre el diámetro de un clavo (x) y su fuerza retirada final (y), se colocaron clavos de forma anular enhebrados en madera de abeto de Douglas, y después se midieron sus fuerzas de retirada en N/mm. Se obtuvieron los resultados siguientes para 10 diámetros diferentes (en mm):

x	2.52	2.87	3.05	3.43	3.68	3.76	3.76	4.50	4.50	5.26
y	54.74	59.01	72.92	50.85	54.99	60.56	69.08	77.03	69.97	90.70

a) Calcule la recta de mínimos cuadrados para predecir la fuerza a partir del diámetro
b) Determine el intervalo de confianza de nivel 0.95 para la media de la fuerza de retirada de clavos de 4 mm de diámetro
c) Determine el intervalo de predicción de nivel 0.95 para la fuerza de retirada de clavos de 4 mm de diámetro
d) ¿Puede concluir que la media de la fuerza de retirada de clavos de 4 mm de diámetro es 60 N/mm con un nivel de significancia de 0.05?

6) Un comerciante realizó un estudio para determinar la relación que hay entre los gastos de la publicidad semanal y las ventas. Registró los datos siguientes:
Costos de publicidad (en $): 40, 20, 25, 20, 30, 50, 40, 20, 50, 40, 25, 50.
Ventas (en $): 385, 400, 395, 365, 475, 440, 490, 420, 560, 525, 480, 510.
a) Haga un gráfico de dispersión
b) Encuentre la recta de regresión estimada para pronosticar las ventas semanales, a partir de los gastos de publicidad.
c) Estime las ventas semanales cuando los costos de la publicidad sean 35$. ¿Es válido estimar las ventas semanales cuando los costos de la publicidad sean 75$?.
d) Pruebe la hipótesis de que $\beta_1 = 6$ contra la alternativa de que $\beta_1 < 6$, utilice $\alpha = 0.025$.
e) Construya un intervalo de confianza de 95% para la media de las ventas semanales cuando se gastan 45$ en publicidad.
f) Construya un intervalo de predicción de 95% para la media de las ventas semanales cuando se gastan 45$ en publicidad.
g) ¿Qué proporción de la variabilidad total en las ventas está explicada por el costo en publicidad?

7) Considere los siguientes datos sobre el número de millas para ciertos automóviles, en millas por galón (mpg) y su peso en libras (wt)

Modelo	GMC	Geo	Honda	Hyundai	Infiniti	Isuzu	Jeep	Land	Lexus	Linclon
Wt (*x*)	4520	2065	2440	2290	3195	3480	4090	4535	3390	3930
Mpg (*y*)	15	29	31	28	23	21	15	13	22	18

a) Estime la recta de regresión lineal.
b) Estime las millas para un vehículo que pesa 4000 libras.
c) Suponga que los ingenieros de Honda afirman que, en promedio, el Civic (o cualquier otro modelo de vehículo que pese 2440 libras) recorre mas de 30 mpg. Con base en los resultados del análisis de regresión, ¿es esta afirmación creíble?, ¿por qué?.
Sugerencia: calcule el intervalo de confianza para la media mpg cuando el peso es de 2440 libras, con $\alpha = 0.05$
d) Los ingenieros de diseño para el Lexus ES300 tienen por objetivo lograr 18 mpg como ideal para dicho modelo (o cualquier otro que pese 3390 libras), aunque se espera que haya cierta variación. ¿Es probable que sea realista ese objetivo?. Comente al respecto.
e) ¿Qué proporción de la variabilidad total en el millaje está explicada por el peso del motor?

8) Los valores siguientes son 26 lecturas sobre la congestión del transito y la concentración de monóxido de carbono efectuadas en un sitio de muestreo para determinar la calidad del aire de cierta ciudad. Enunciando las suposiciones necesarias , resolver los siguientes incisos :
a) Prepare un diagrama de dispersión.
b) Calcule el coeficiente de correlación de la muestra.
c) Pruebe que H_0 : $\rho=0$ al nivel de significación de 0.05 y saque sus conclusiones.
d) Construya el intervalo de confianza del 95% para ρ.
e) ¿Hay suficiente evidencia de una correlación mayor que 0.95 ?.Utilice $\alpha=0.05$.

congestion de transito		congestion de transito	
(automoviles por hora), X	CO (ppm), Y	(automoviles por hora), X	CO (ppm), Y
100	8.8	375	13.2
110	9.0	400	14.5
125	9.5	425	14.7
150	10.0	450	14.9
175	10.5	460	15.1
190	10.5	475	15.5
200	10.5	500	16.0
225	10.6	525	16.3
250	11.0	550	16.8
275	12.1	575	17.3
300	12.1	595	18.0
325	12.5	600	18.4