Teorema N4: Sea $\mathbf{X}' = (X_1, \ldots, X_k)$ un vector $k-$dimensional con distribución normal multivariada, luego la distribución marginal de cualquier subconjunto de componentes tiene distribución normal multivariada. En particular cada componente es normal.

Demostración: Sea $\mathbf{X}^* = (X_{k_1}, \ldots, X_{k_h})$, $k_1 < k_2 < \ldots < k_h$, luego se tiene que $\mathbf{X}^* = A\mathbf{X}$, donde A es la matriz de $h \times k$ dada por:

$$a_{ij} = \begin{cases} 1 & \text{si } j = k_i \\ 0 & \text{si } j \neq k_i \end{cases}$$

$1 \leq i \leq h$, $1 \leq j \leq k$.

Es fácil ver que A es una matriz de rango h.

Teorema N5: Si $\mathbf{X}$ es un vector $k-$dimensional con distribución $N_k(\boldsymbol{\mu}, \mathbf{Q})$, luego

$$(\mathbf{X}-\boldsymbol{\mu})'\mathbf{Q}^{-1}(\mathbf{X}-\boldsymbol{\mu}) \sim \chi^2_k$$

.

Demostración: Por lo ya visto, resulta que $\mathbf{X} = \mathbf{B}\mathbf{Y} + \boldsymbol{\mu}$ donde $\mathbf{Y}$ es $N(\mathbf{0_k}, \mathbf{I_k})$

$$\mathbf{Y} = \mathbf{B}^{-1}(\mathbf{X}-\boldsymbol{\mu})$$

y además

$$\mathbf{B}\mathbf{B}' = \mathbf{Q}$$

Luego

$$\mathbf{Y}\mathbf{Y}' = (\mathbf{X}-\boldsymbol{\mu})'\mathbf{B}'^{-1}\mathbf{B}^{-1}(\mathbf{X}-\boldsymbol{\mu}) = (\mathbf{X}-\boldsymbol{\mu})'\mathbf{Q}^{-1}(\mathbf{X}-\boldsymbol{\mu})$$

El teorema resulta del hecho que

$$\mathbf{Y}'\mathbf{Y} = \sum_{i=1}^{k} Y_i^2$$

tiene distribución χ^2_k, ya que las Y_i son variables aleatorias independientes con distribución $N(0,1)$.

Teorema N6: Si $\mathbf{X}$ es un vector $k-$dimensional con distribución $N_k(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_k)$ y $\mathbf{P}$ una matriz simétrica e idempotente de rango r, entonces

$$\frac{(\mathbf{X}-\boldsymbol{\mu})'\mathbf{P}(\mathbf{X}-\boldsymbol{\mu})}{\sigma^2} \sim \chi^2_r$$

.

Tests y Regiones de Confianza

Hasta ahora hemos trabajado sólo con las hipótesis Ω. Sin embargo para deducir tests y regiones de confianza con nivel exacto será necesario que hagamos un supuesto adicional: **normalidad conjunta de los errores**

Supondremos que las y_i's se distribuyen conjuntamente según una normal multivariada.

Podremos deducir:

- intervalos de confianza de nivel exacto para funciones paramétricas estimables
- tests de nivel exacto para hipótesis que involucran a los parámetros
- conjuntos o regiones de confianza para la estimación simultánea de más de una función paramétrica estimable.

Nuestro nuevo modelo será:

$$\boxed{\Omega: \quad \mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}) \quad rg(\mathbf{X}) = r \quad \boldsymbol{\beta} \in \Re^p}$$

Observemos que en este caso suponer que $\Sigma_{\mathbf{Y}} = \sigma^2 \mathbf{I}$ es equivalente a asumir que las $y_i, 1 \leq i \leq n$, son independientes.

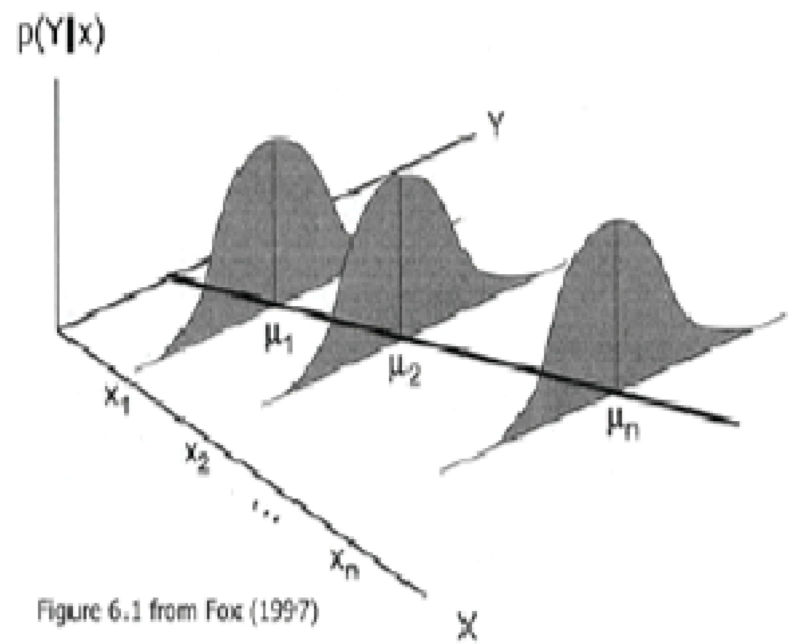


Figure 6.1 from Fox (1997)

Bajo estas condiciones se obtiene el siguiente resultado:

Teorema: Supongamos que se tiene el modelo

$$\Omega : \quad \mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}) \quad \mathrm{rg}(\mathbf{X}) = p \quad \boldsymbol{\beta} \in \Re^p \,.$$

Luego, $\widehat{\boldsymbol{\beta}}$ y s^2 son funciones de estadísticos suficientes y completos y por lo tanto, $\widehat{\boldsymbol{\beta}}$ y s^2 son estimadores IMVU de $\boldsymbol{\beta}$ y σ^2, respectivamente.

Si nuestro modelo es

$$E(\mathbf{Y}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

podríamos tener interés en testear hipótesis como las que siguen:

$$H_o : \beta_j = 0 \text{ vs. } H_1 : \beta_j \neq 0$$
$$H_o : \beta_1 - \beta_2 = 0 \text{ vs. } H_1 : \beta_1 - \beta_2 \neq 0$$
$$H_o : \beta_1 = \beta_2 = \ldots = \beta_p = 0 \text{ vs. } H_1 : \text{existe } j : \beta_j \neq 0$$

Todas estas hipótesis son de la forma $\mathbf{c}'\boldsymbol{\beta} = 0$ o $\mathbf{C}\boldsymbol{\beta} = \mathbf{0}$.

Supongamos que tenemos q funciones estimables $\psi_1, \psi_2, \ldots, \psi_q$ donde:

$$\psi_i = \sum_{j=1}^{p} c_{ij}\beta_j \quad 1 \leq i \leq q$$

Por ser estimables, por el Teorema de Gauss–Markov tenemos que

$$\widehat{\psi}_i = \sum_{j=1}^{n} a^*_{ij} y_j \quad 1 \leq i \leq q\,,$$

donde $\mathbf{a}^*_i \in \mathcal{V}_r \subset \Re^n$; de manera que

$$\begin{aligned} \Psi &= \mathbf{C}\boldsymbol{\beta} \qquad \mathbf{C} \in \Re^{q\times p} \\ \widehat{\Psi} &= \mathbf{A}^*\mathbf{Y} \qquad \mathbf{A}^* \in \Re^{q\times n} \end{aligned}$$

Más aún, sabemos que

$$\begin{aligned} \widehat{\Psi} &= \mathbf{C}\widehat{\boldsymbol{\beta}} \\ \Sigma_{\widehat{\Psi}} &= \sigma^2 \mathbf{A}^*\mathbf{A}^{*'} \end{aligned}$$

Estimamos a σ^2 por

$$s^2 = \frac{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2}{n - r}$$

Bajo estas nuevas hipótesis obtenemos el siguiente resultado:

Teorema: Supongamos que se cumple Ω, es decir $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, $\mathrm{rg}(\mathbf{X}) = r$, $\boldsymbol{\beta} \in \Re^p$ y que además que $\psi_1, \psi_2, \ldots, \psi_q$ son q funciones estimables l.i., de manera que $\mathrm{rg}(\mathbf{C}) = q$. Entonces,

i) $\widehat{\Psi} \sim N_q(\Psi, \Sigma_{\widehat{\Psi}})$ (o lo que es igual $N_q(\Psi, \sigma^2\mathbf{A}^*\mathbf{A}^{*'})$)

ii) $\widehat{\Psi}$ y $\dfrac{s^2(n-r)}{\sigma^2}$ son independientes

iii) $\dfrac{(n-r)s^2}{\sigma^2} \sim \chi^2_{n-r}$

En el caso de rango completo, es decir cuando $r = p$, obtenemos el siguiente resultado:

Teorema: Supongamos que se cumple Ω, es decir $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, $\text{rg}(\mathbf{X}) = p$, $\boldsymbol{\beta} \in \Re^p$. Entonces,

i) $\widehat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$

ii) $\dfrac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\mathbf{X}'\mathbf{X})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\sigma^2} \sim \chi^2_p$

iii) $\widehat{\boldsymbol{\beta}}$ y $\dfrac{(n-p)s^2}{\sigma^2}$ son independientes

iv) $\dfrac{(n-p)s^2}{\sigma^2} \sim \chi^2_{n-p}$

Estos resultados nos permiten deducir intervalos de confianza o tests para cada uno de los coeficientes del modelo lineal:

Como $\widehat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$, entonces $\widehat{\beta}_i = \mathbf{e}_i'\widehat{\boldsymbol{\beta}} \sim N(\beta_i, \sigma^2\mathbf{e}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{e}_i)$.
Si denotamos $\Sigma_{\widehat{\boldsymbol{\beta}}} = \sigma^2\mathbf{D}$

$$\boxed{\widehat{\beta}_i \sim N(\beta_i, \sigma^2 d_{ii})}$$

siendo d_{ii} el i-ésimo elemento diagonal de $\mathbf{D}$.

Si para un i fijo queremos testear

$$\boxed{H_o : \beta_i = 0 \quad \text{vs.} \quad H_1 : \beta_i \neq 0}$$

tenemos que bajo H_o

$$\frac{\widehat{\beta}_i}{s\sqrt{d_{ii}}} \sim t_{n-p}$$

Por lo tanto, rechazaremos H_o con nivel α si

$$\left|\frac{\widehat{\beta}_i}{s\sqrt{d_{ii}}}\right| > t_{n-p,\frac{\alpha}{2}}$$

En el caso de regresión simple tendríamos

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad 1 \le i \le n, \quad \epsilon_i \sim N(0, \sigma^2)$$

Entonces:

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix}$$

y la inversa resulta

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{n\sum_{i=1}^{n} x_i^2 - n^2\overline{x}^2} \begin{pmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{pmatrix}$$

$$\widehat{\beta}_0 = -\overline{x}\,\widehat{\beta}_1 + \overline{y}$$

y

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

Luego, si queremos testear

$$H_o : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0$$

el estadístico será

$$T = \left| \frac{\widehat{\beta}_1}{s\sqrt{d_{11}}} \right| = \left| \frac{\widehat{\beta}_1}{s/\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}} \right|$$

y rechazaremos H_o si

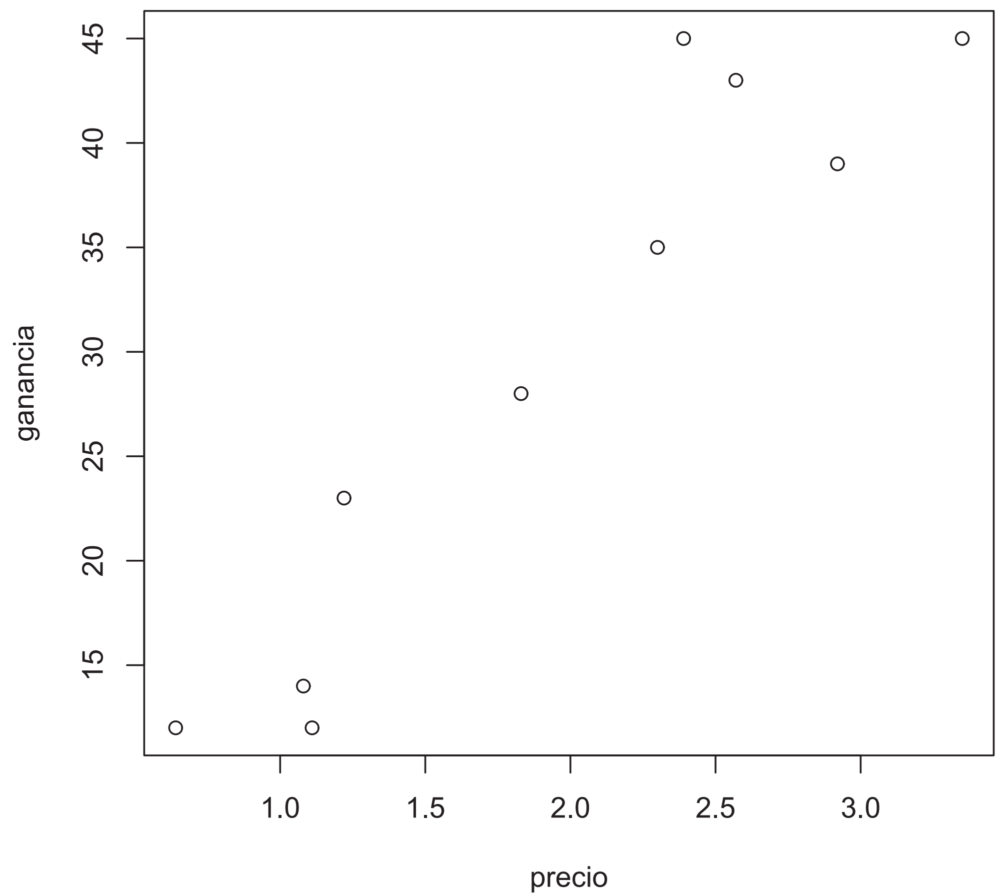
$$\boxed{|T| > t_{n-2,\frac{\alpha}{2}}}$$

Veamos un ejemplo: Precio del papel.

Y: ganancia en 1972

x: precio de papel en 1973

```
> Ejemplo Precio del Papel
   precio  ganancia
       x       y
 1   1.83   28
 2   3.35   45
 3   0.64   12
 4   2.30   35
 5   2.39   45
 6   1.08   14
 7   2.92   39
 8   1.11   12
 9   2.57   43
10   1.22   23
```



45
40
35
30
25
20
15
ganancia
1.0
1.5
2.0
2.5
3.0
precio

```
> sal.lm
Coefficients:
 (Intercept)         x
    2.027775      14.20517

Degrees of freedom: 10 total; 8 residual
Residual standard error: 5.025083

> summary(sal.lm)

Call: lm(formula = y ~ x, x = T)
Residuals:
    Min        1Q     Median      3Q        Max
 -5.796   -4.222    0.1386    2.952     9.022
```

```
Coefficients:
                 Value    Std. Error    t value    Pr(>|t|)
(Intercept)   2.0278     3.9383          0.5149      0.6206
          x  14.2052     1.8565          7.6516      0.0001

Residual standard error: 5.025 on 8 degrees of freedom
Multiple R-Squared: 0.8798
F-statistic: 58.55 on 1 and 8 degrees of freedom, the p-value
is 0.00006008

Correlation of Coefficients:
  (Intercept)
            x  -0.915
 X'X=
                (Intercept)       x
(Intercept)       10.00         19.4100
          x       19.41         45.0013
```

```
(X'X)^(-1)=
              (Intercept)         x
(Intercept)    0.6142273  -0.264929
          x   -0.2649290   0.136491

> matriz de covarianza de coeficientes

             (Intercept)         x
(Intercept)    15.510133 -6.689844
          x    -6.689844  3.446597
```

También podríamos interesarnos realizar in I. de C. para la esperanza de una nueva observación independiente de las demás que cumpla el modelo

$$y_i = \beta_1 x_1 + \beta_1 x_2 + \ldots + \beta_p x_p + \epsilon_i$$

en $\mathbf{x}_o = (\mathbf{x}_{o1}, \mathbf{x}_{o2}, \ldots, \mathbf{x}_{op})'$ donde $\epsilon_i \sim N(0, \sigma^2)$ independientes.

Como $E(y_o) = \mathbf{x}_o'\boldsymbol{\beta}$, podemos estimarlo por $\widehat{E(y_o)} = \mathbf{x}_o'\widehat{\boldsymbol{\beta}} = \widehat{y}_o$

Por lo tanto, de acuerdo con lo que hemos visto

$$\widehat{y}_o = \mathbf{x}_o'\widehat{\boldsymbol{\beta}} \sim N(\mathbf{x}_o'\boldsymbol{\beta}, \sigma^2 \mathbf{x}_o'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_o)$$

y es independiente de

$$\frac{(n-p)s^2}{\sigma^2} \sim \chi^2_{n-p}$$

por lo tanto

$$T = \frac{\widehat{y}_o - \mathbf{x}_o'\boldsymbol{\beta}}{s\sqrt{\mathbf{x}_o'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_o}} \sim t_{n-p}$$

En consecuencia,

$$\boxed{\widehat{y}_o \pm t_{n-2,\frac{\alpha}{2}}\ s\sqrt{\mathbf{x}_o'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_o}}$$

es un intervalo de nivel exacto $1-\alpha$.

Asimismo, podríamos estar interesados en la predicción de y_o, una nueva observación que cumpla el modelo, y en un intervalo para ella, que llamaremos intervalo de predicción.

Observemos que el predictor de y_o es $\widehat{y}_o = \mathbf{x}_o'\widehat{\boldsymbol{\beta}}$. En efecto, $E(\widehat{y}_o - y_o) = 0$. ¿Qué distribución tiene $\widehat{y}_o - y_o$?

Tenemos que

$$\begin{aligned} \widehat{y}_o &\sim N(\mathbf{x}_o'\boldsymbol{\beta}, \sigma^2\mathbf{x}_o'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_o) \\ y_o &\sim N(\mathbf{x}_o'\boldsymbol{\beta}, \sigma^2) \end{aligned}$$

y dado que y_o es independiente de las restantes y_i's con las que estimamos, entonces por la independencia entre estas dos normales queda que

$$\widehat{y}_o - y_o \sim N(0, \sigma^2(1 + \mathbf{x}_o'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_o))$$

Por lo tanto, el intervalo de predicción de nivel $1 - \alpha$ estará dado por

$$\boxed{\widehat{y}_o \pm t_{n-2,\frac{\alpha}{2}}\; s\sqrt{1 + \mathbf{x}_o'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_o}}$$

Ejemplo. Los siguientes son datos que corresponden a 10 porcentajes y_i de una sustancia que fueron medidos en experiencias de laboratorio y que se desean relacionar con la temperatura x_i a la que fueron realizados dichas experiencias.

i	x	y
1	100	45
2	110	52
3	120	54
4	130	63
5	140	62
6	150	68
7	160	75
8	170	76
9	180	92
10	190	88

La tabla con los estadísticos calculados es:

Coeficiente	Estimación	Error Standard	Valor de t
β_0	-4.47273	5.63433	-0.79
β_1	0.49636	0.03812	13.02
s	3.46213	g.l.=8	

Intervalos de Estimación y de Predicción

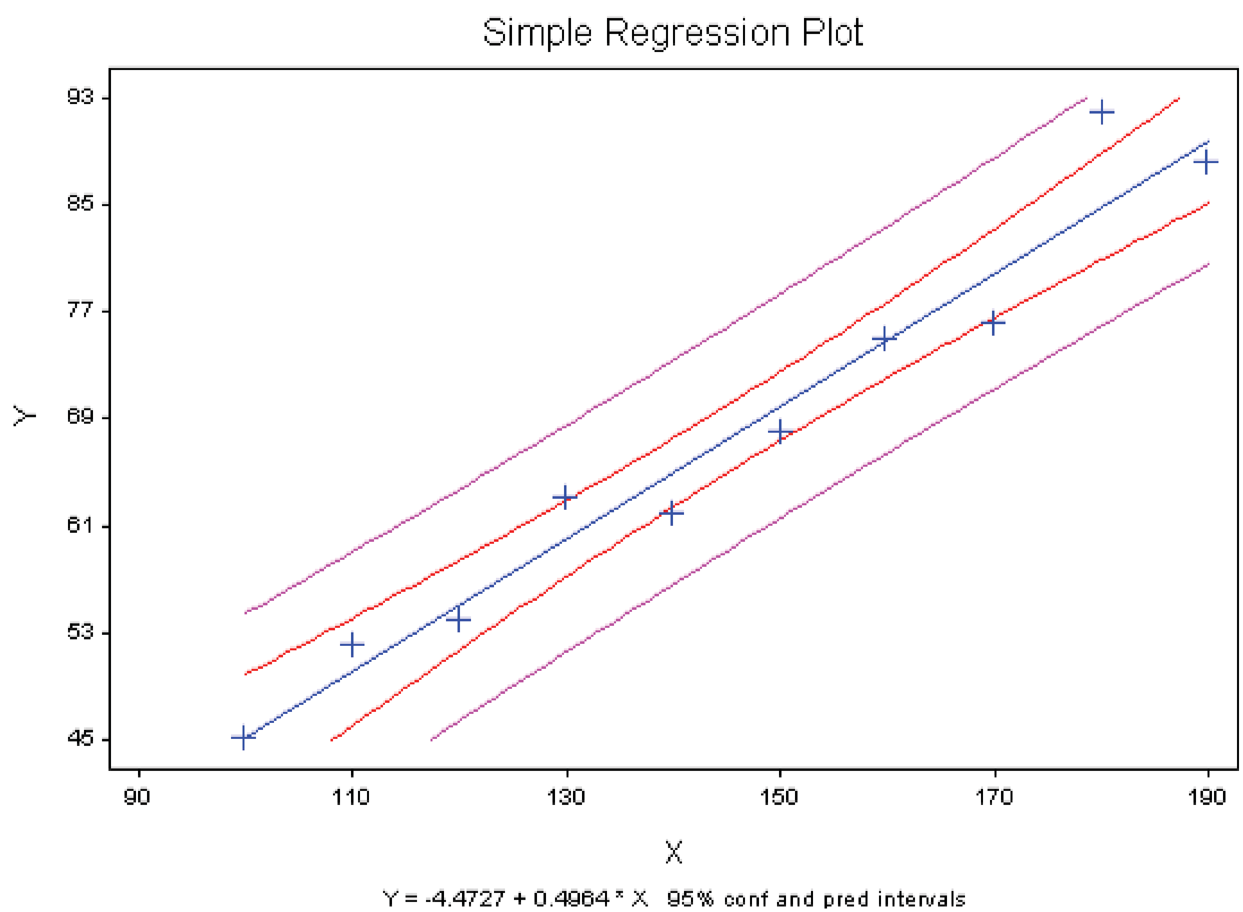


Tabla de Resultados

```
UNWEIGHTED LEAST SQUARES LINEAR REGRESSION OF Y

PREDICTOR
VARIABLES     COEFFICIENT    STD ERROR     STUDENT'S T      P
---------     -----------    ---------     -----------    ------
CONSTANT        -4.47273      5.63433         -0.79       0.4502
X                0.49636      0.03812         13.02       0.0000

R-SQUARED            0.9549      RESID. MEAN SQUARE (MSE)     11.9864
ADJUSTED R-SQUARED   0.9493      STANDARD DEVIATION            3.46213

SOURCE        DF       SS          MS          F         P
----------    ---   ----------  ----------   -----    ------
REGRESSION     1      2032.61     2032.61    169.58   0.0000
RESIDUAL       8      95.8909     11.9864
TOTAL          9      2128.50

CASES INCLUDED 10   MISSING CASES 0
```

- El valor estimado de $\widehat{\beta}_1 \simeq 0{,}5$, $\Rightarrow$ esperamos que el porcentaje aumente 0.5 por cada incremento de un grado en la temperatura.

- $s_{\widehat{\beta}_1} = 0{,}03812$

- Si testeamos $H_0 : \beta_1 = 0$ $t = \frac{0{,}49636}{0{,}038112} = 13{,}02$ y $t_{8,0,025} = 2{,}306004$

$\Rightarrow$ los datos nos dan evidencia suficiente al nivel 5 % como para concluir que la pendiente es no nula.