Lección 8.1: Análisis de Regresión

Variación Experimental

Alfaomega

Alfaomega-UAQro CIMAT

2016

Índice

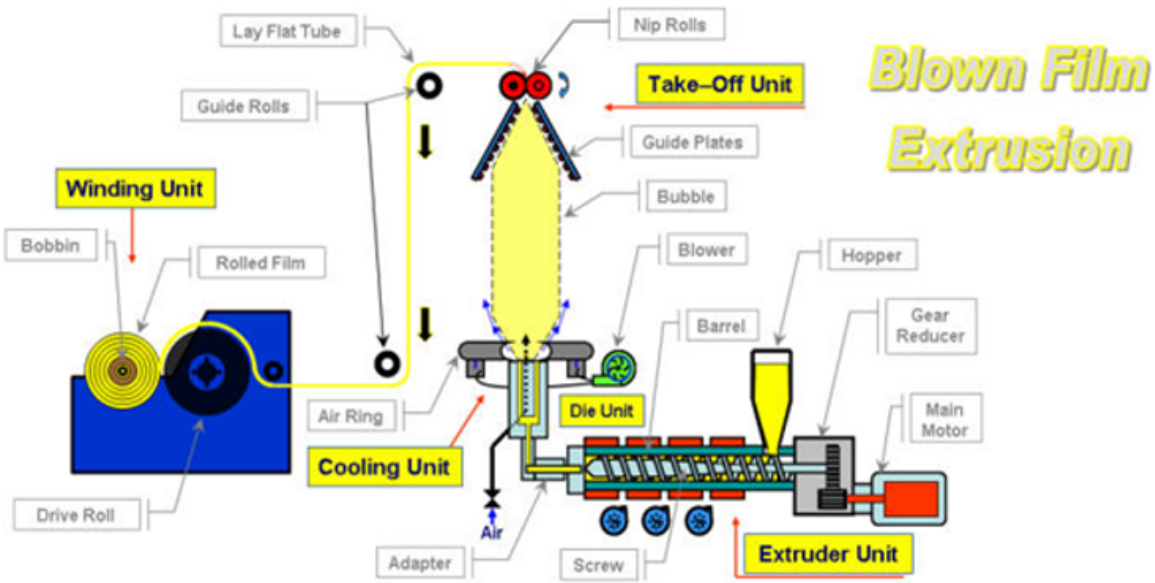
1. Presentación
2. Variación experimental
   - Modelo de regresión líneal simple
   - Descripción de las discrepancias
3. Inferencia estadística
   - Prueba de hipótesis del modelo de regresión
4. Estimación con R

- En esta lección se describirá el procedimiento para estudiar el modelo de regresión lineal simple.
- Primero se motivará la relación entre dos variables. Se plantea una propuesta didáctica para la construcción del modelo de regresión, para ello se retoma el enfoque de la variación experimental, ver lección 1.2.
- En la lección 8.2 se extiende a más variables independientes, y se aplica la programación en R.
- En el capítulo 9.1 se presenta la extensión de dos variables.

Ejemplo de un proceso

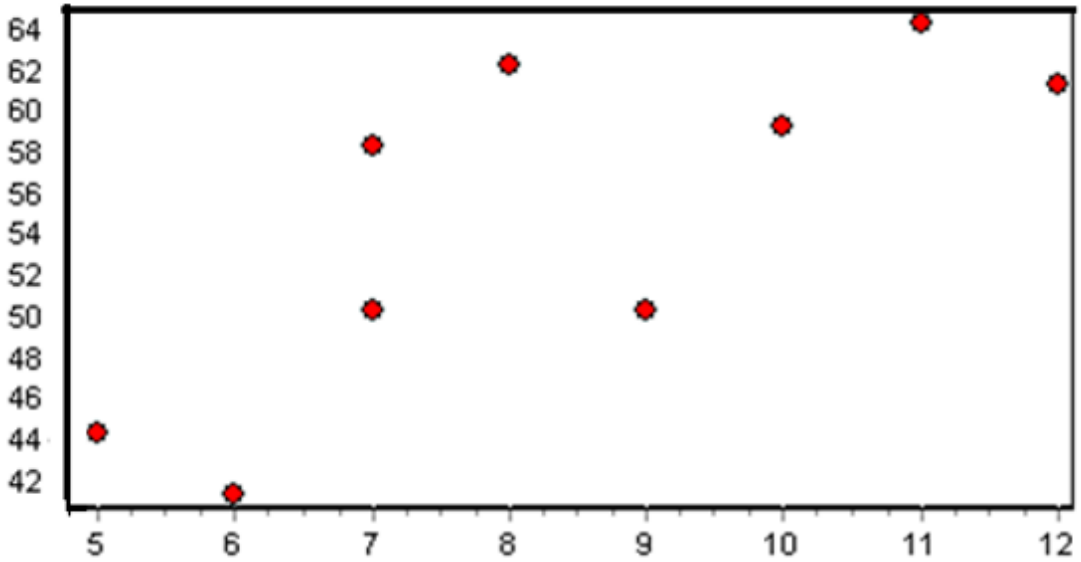
Proceso de plásticos

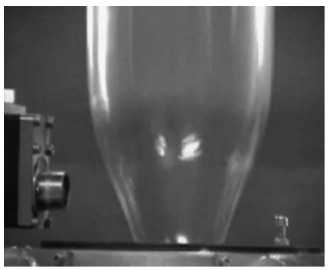
Entre las variables de interés en un proceso de extrusión en la elaboración de bolsas de plástico están, la variable X : temperatura de la masa del dado y, la variable Y : resistencia a máxima de tensión a la ruptura de la película.



Relación entre dos variables

Descripción de los resultados del proceso de plásticos



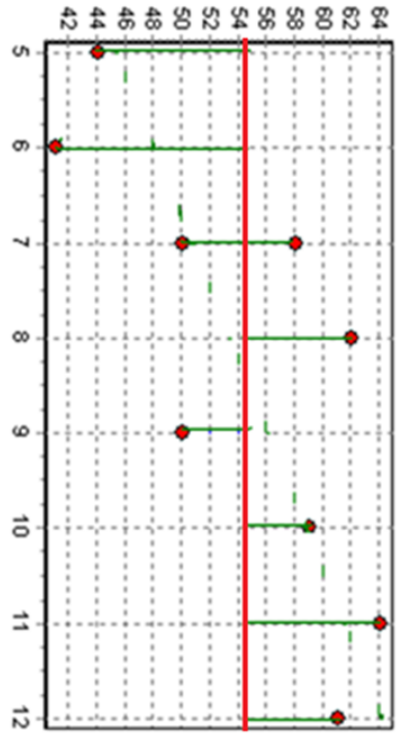


X	Y
5	44
6	41
7	58
7	50
8	62
9	50
10	59
11	64
12	61

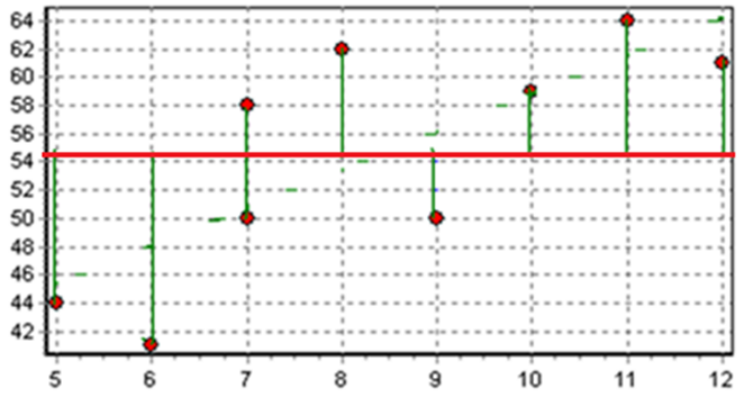
Varianza experimental

Evaluación de la discrepancia, variable Y

Estimación del modelo: $Y = \mu + \epsilon$



$$d_i = y_i - \overline{y}$$



$$d_i = y_i - \overline{y}$$

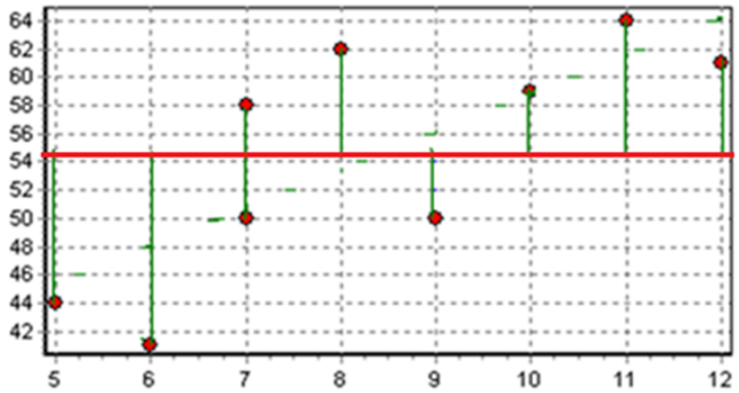
La línea roja $\overline{y}$ es una estimación del parámetro μ.

Las líneas verdes d_i miden la discrepancia de los puntos a la línea roja. La varianza evalúa esta discrepancia

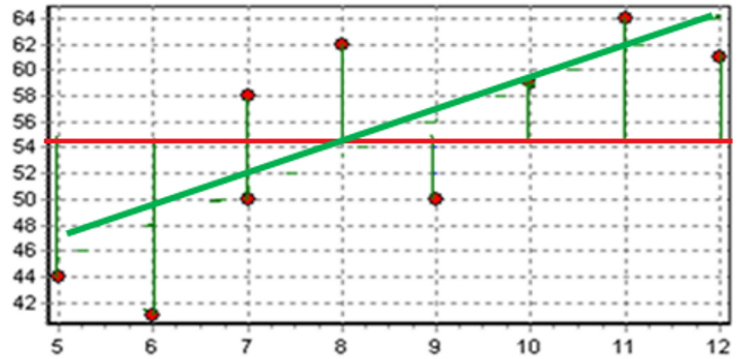
Varianza experimental

Descripción de la estimación del modelo $Y = \mu(X) + \epsilon$

Gráfica que describe la discrepancia de los datos experimentales con respecto a la media $\overline{y}$



El modelo verde y estima a $\mu(X) = \beta_0 + \beta_1 X$



Comentarios

- Al estimar el modelo $Y = \mu(X) + \epsilon = \beta_0 + \beta_1 X + \epsilon$, surge la necesidad de evaluar las discrepancias de los datos experimentales y el modelo ajustado -línea verde-.
- Se puede ver que la situación es similar al modelo

$$Y = \mu + \epsilon$$

  Donde se han evaluado las discrepancias $d_i = y - \overline{y}$.
- Comente y escriba algunas ideas sobre este planteamiento.

Modelo y Discrepancia

Modelo:

$$Y = \mu(X) + \epsilon = \beta_0 + \beta_1 X + \epsilon,$$

Discrepancia

$Y - \mu(X) = \epsilon,$

$(\beta_0 + \beta_1 X) - \mu(X) = \epsilon.$

ϵ es variable aleatoria con distribución normal $(0, \sigma^2)$ para cada valor de X

Minimizar la discrepancia

¿Cuál es la idea intuitiva de la discrepancia?

En principio que sea lo más pequeña posible. Matemáticamente es minimizar:

$$\sum_{i=1}^{n} \epsilon_i \epsilon_i = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \mu)^2$$

Modelo y Discrepancia

Modelo:

$Y = \mu(X) + \epsilon = \beta_0 + \beta_1 X + \epsilon,$

Discrepancia

$Y - \mu(X) = \epsilon,$

$(\beta_0 + \beta_1 X) - \mu(X) = \epsilon.$

ϵ es variable aleatoria con distribución normal $(0, \sigma^2)$ para cada valor de X

Minimizar la discrepancia

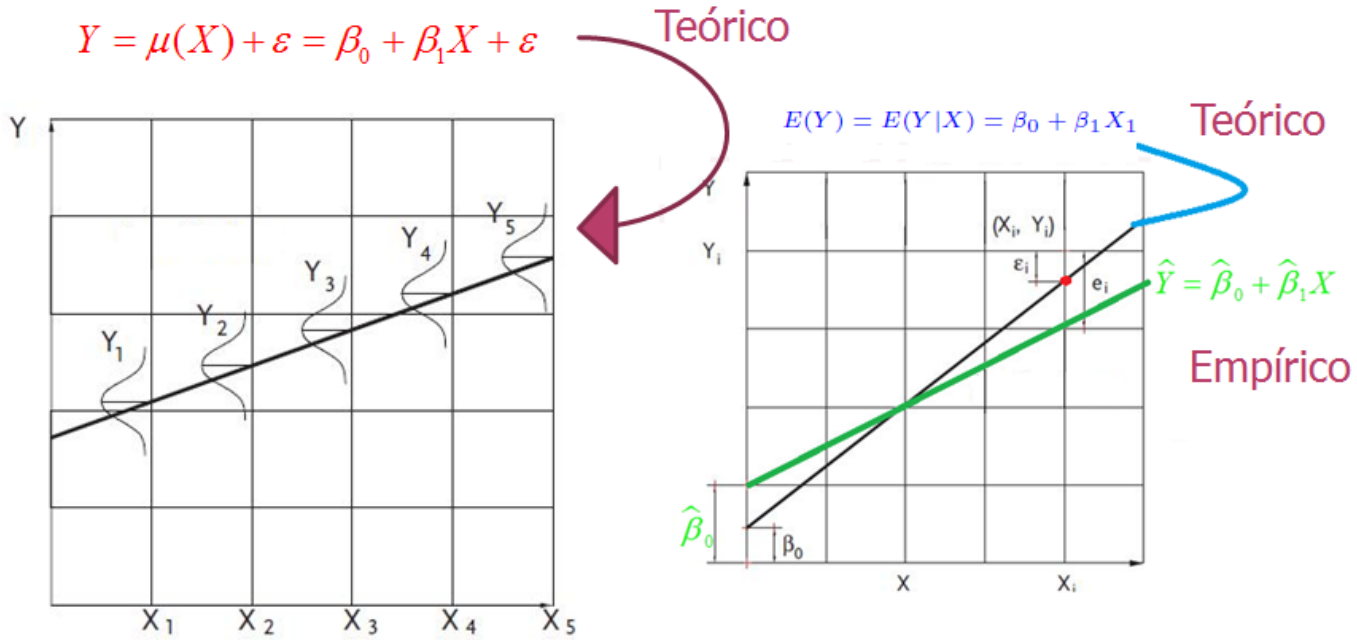
¿Cuál es la idea intuitiva de la discrepancia?

En principio que sea lo más pequeña posible. Matemáticamente es minimizar:

$$\sum_{i=1}^{n} \epsilon_i \epsilon_i = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \mu)^2$$

Modelo

Descripción de la relación modelo teórico y empírico.



Presentación de los supuestos para estimar los parámetros del modelo de regresión

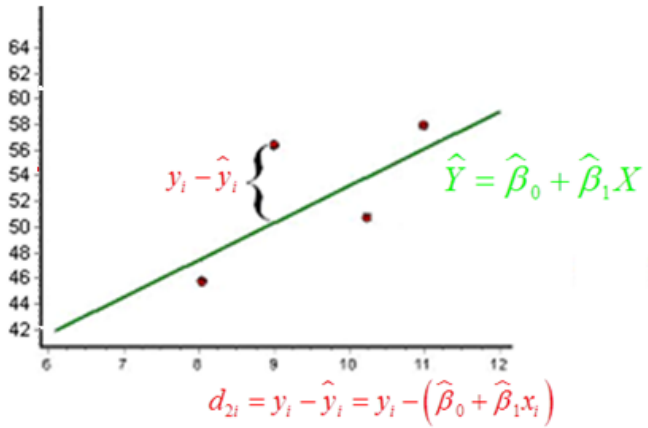
Modelo

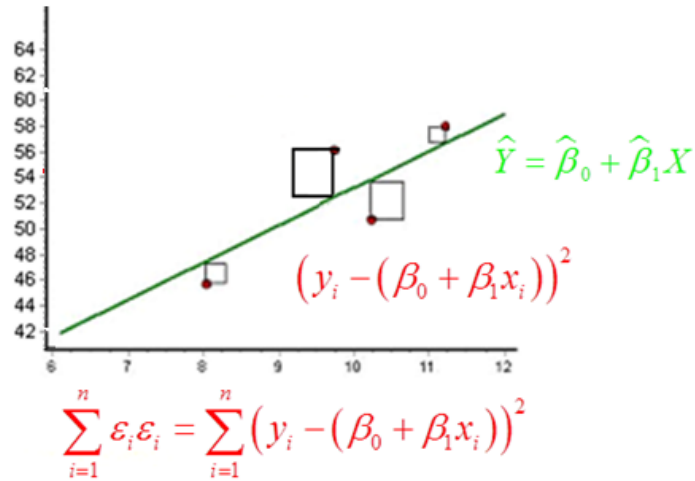
Evaluación de la discrepancia d_{2i}

$$d_{2i} = y_i - \widehat{y_i} = y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i)$$

Modelo:

$$Y = \beta_0 + \beta_1 X + \epsilon$$





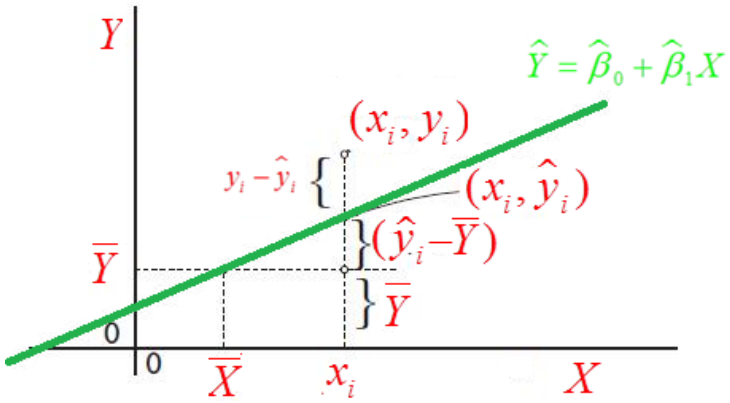
Suma de mínimos cuadrados

$$\sum_{i=1}^{n} \epsilon_i \epsilon_i = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

Objetivo minimizar la discrepancia.

Modelo

Estimación de los parámetros del modelo: Solución del proceso de optimización.



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

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

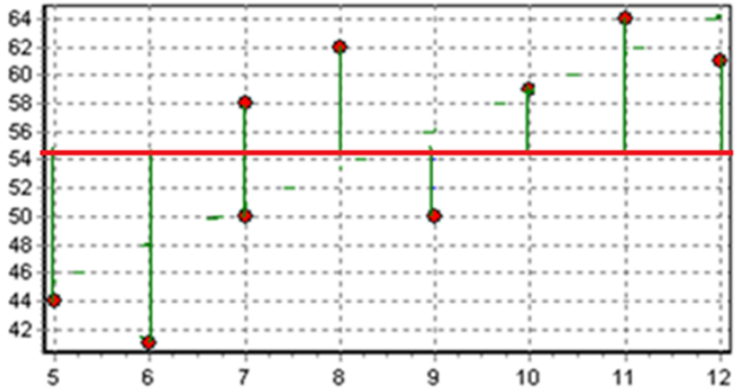
$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X} = 54.333 - 2.708(8.333) = 31.795$$

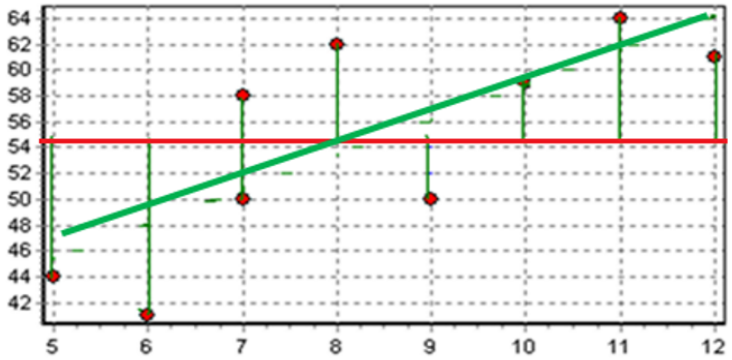
$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \overline{X})^2} = \frac{119.163}{43.992} = 2.705$$

Así, el modelo estimado es: $\widehat{y} = 31.795 + 2.705x$

Modelo

Descripción de la discrepancia en cada parte

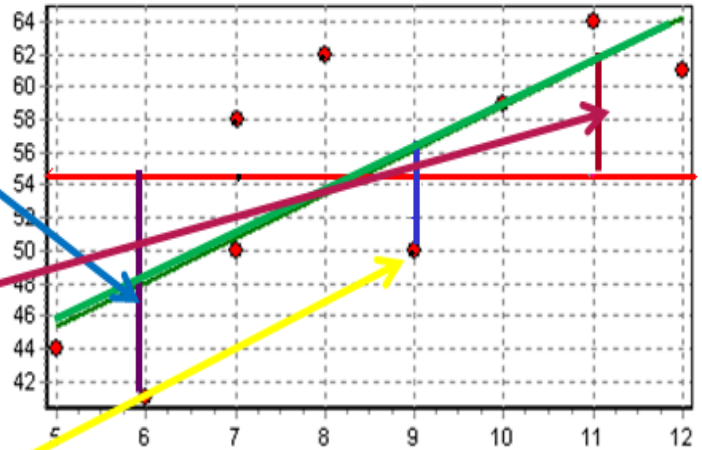




$d_i = y_i - \bar{Y}$

$d_{1i} = \hat{y}_i - \bar{Y}$

$d_{2i} = y_i - \hat{y}_i$



Modelo

Procedimiento para calcular de la discrepancia en cada parte

$d_i = d_{1i} + d_{2i}$

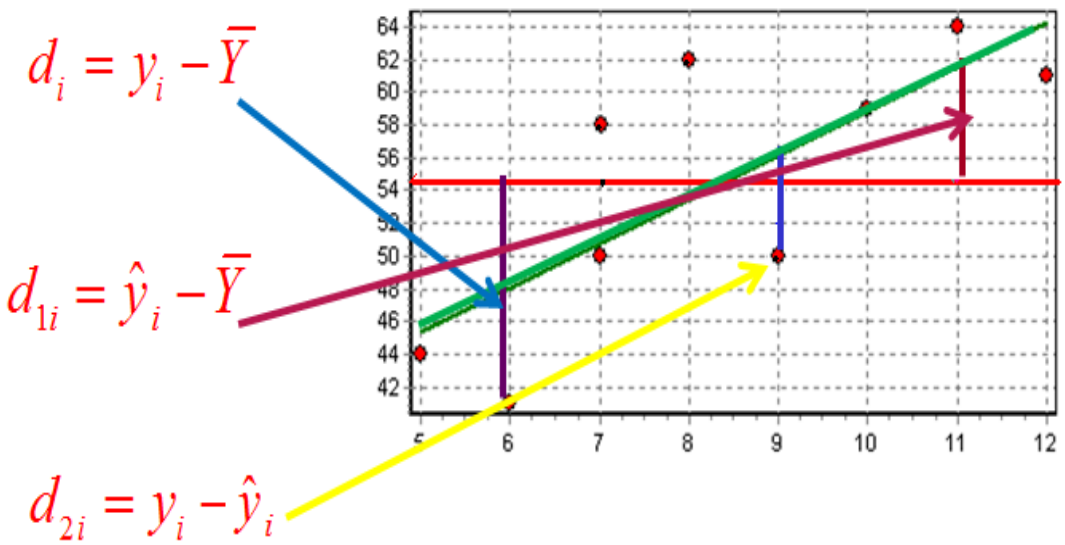
$y_i - \overline{Y} = (\widehat{y}_i - \overline{Y}) + (y_i - \widehat{y}_i)$

$(y_i - \overline{Y})^2 = ((\widehat{y}_i - \overline{Y}) + (y_i - \widehat{y}_i))^2$

La acción es sumar:

$sumar[(y_i - \overline{Y})^2] = sumar[((\widehat{y}_i - \overline{Y}) + (y_i - \widehat{y}_i))^2]$

$sumar[(y_i - \overline{Y})^2] = sumar[(\widehat{y}_i - \overline{Y})^2 + (y_i - \widehat{y}_i)^2 - 2(\widehat{y}_i - \overline{Y})(y_i - \widehat{y}_i)]$



Modelo

Procedimiento para calcular de la discrepancia en cada parte

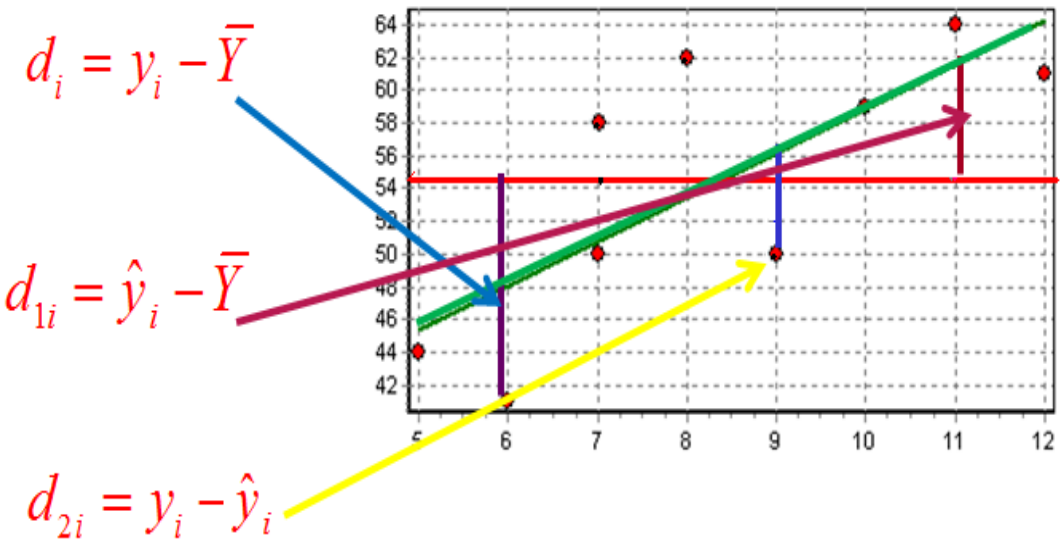
$d_i = d_{1i} + d_{2i}$

$y_i - \overline{Y} = (\widehat{y}_i - \overline{Y}) + (y_i - \widehat{y}_i)$

$(y_i - \overline{Y})^2 = ((\widehat{y}_i - \overline{Y}) + (y_i - \widehat{y}_i))^2$

$\sum_{i=1}^{n}[(y_i - \overline{Y})^2] = \sum_{i=1}^{n}[((\widehat{y}_i - \overline{Y}) + (y_i - \widehat{y}_i))^2]$

$\sum_{i=1}^{n}[(y_i - \overline{Y})^2] = \sum_{i=1}^{n}[(\widehat{y}_i - \overline{Y})^2 + (y_i - \widehat{y}_i)^2 - 2(\widehat{y}_i - \overline{Y})(y_i - \widehat{y}_i)]$



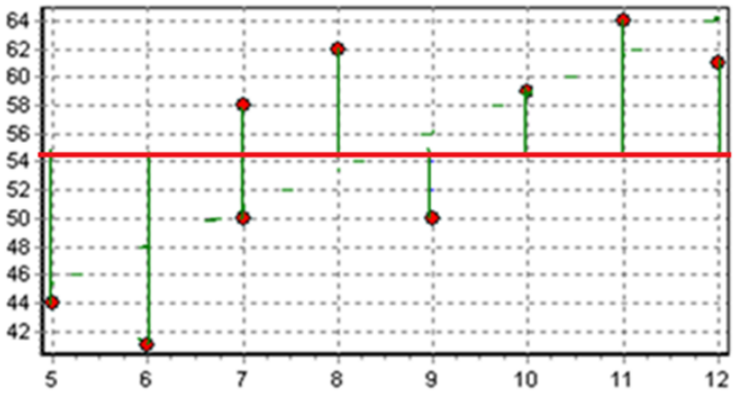
Cálculo de las discrepancias

Referencia de las tres discrepancias

x_i	y_i	$\widehat{y}_i$	$y_i - \overline{Y}$	$\widehat{y}_i - \overline{Y}$	$y_i - \widehat{y}_i$
5	44	45.31	-10.33	-9.02	-1.31
6	41	48.02	-13.33	-6.31	-7.02
7	58	50.72	3.67	-3.61	7.28
7	50	50.72	-4.33	-3.61	-0.73
8	62	53.43	7.67	-0.90	8.57
9	50	56.13	-4.33	1.80	-6.13
10	59	58.84	4.67	4.51	0.16
11	64	61.54	9.67	7.21	2.46
12	61	64.25	6.67	9.92	-3.25

Evaluación de la discrepancia del:

Modelo rojo $\overline{Y}$ con los valores y_i así: $(y_i - \overline{Y})$



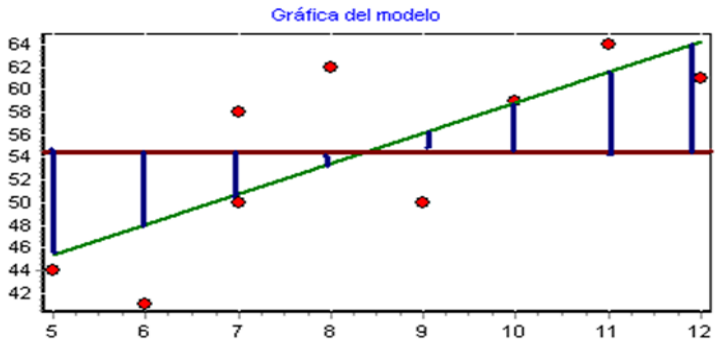
$$d_i = y_i - \overline{Y}$$

obs.	$(y_i - \overline{Y})$ Discrepancia: Obs y media	$(y_i - \overline{Y})^2$ Cuadrado de la discrepancia
1	-10.33	106.70
2	-13.33	177.68
3	3.67	13.46
4	-4.33	18.74
5	7.67	58.82
6	-4.33	18.74
7	4.67	21.80
8	9.67	93.50
9	6.67	44.48
		554.00

Terminología: Discrepancia del total

Evaluación de la discrepancia del:

Modelo rojo $\overline{Y}$ con los valores estimados $\widehat{y}_i$ así: $(\widehat{y}_i - \overline{Y})$



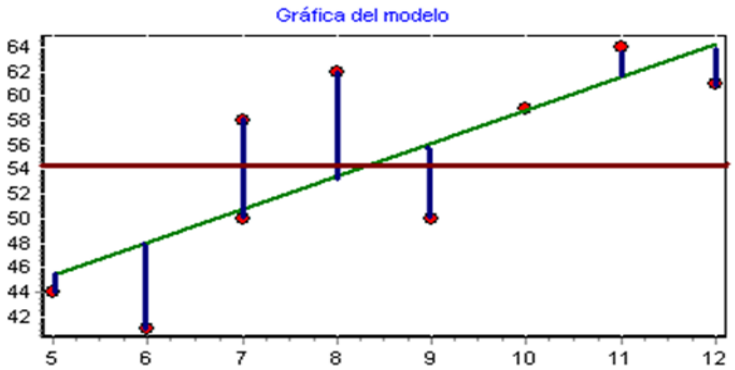
$$d_i = \widehat{y}_i - \overline{Y}$$

	$(\widehat{y}_i - \overline{Y})$	$(\widehat{y}_i - \overline{Y})^2$
obs.	Valor estimado: menos la media	Cuadrado de la discrepancia
1	-9.02	81.36
2	-6.31	39.82
3	-3.61	13.03
4	-3.61	13.03
5	-0.90	0.81
6	1.80	3.24
7	4.51	20.34
8	7.21	51.98
9	9.92	98.41
		321.841

Terminología: Discrepancia de lo que explica el modelo

Evaluación de la discrepancia del:

Modelo verde $\widehat{y} = 31.795 + 2.705$ a los valores y_i de la variable Y



$d_i = y_i - \widehat{y}_i$

	$(y_i - \widehat{y}_i)$	$(y_i - \widehat{y}_i)^2$
1	-1.31	1.71
2	-7.02	49.28
3	7.28	52.99
4	-0.73	0.53
5	8.57	73.44
6	-6.13	37.57
7	0.16	0.02
8	2.46	6.05
9	-3.25	10.56
		232.159

Terminología: Discrepancia de lo que no explica el modelo:

Residuales: $y_i - \widehat{y}_i$

Construcción del análisis de la varianza del modelo

	$(y_i - \overline{Y})$	$(y_i - \overline{Y})^2$	$(\widehat{y}_i - \overline{Y})$	$(\widehat{y}_i - \overline{Y})^2$	$(y_i - \widehat{y}_i)$	$(y_i - \widehat{y}_i)^2$
1	-10.33	106.70	-9.02	81.36	-1.31	1.71
2	-13.33	177.68	-6.31	39.82	-7.02	49.28
3	3.67	13.46	-3.61	13.03	7.28	52.99
4	-4.33	18.74	-3.61	13.03	-0.73	0.53
5	7.67	58.82	-0.90	0.81	8.57	73.44
6	-4.33	18.74	1.80	3.24	-6.13	37.57
7	4.67	21.80	4.51	20.34	0.16	0.02
8	9.67	93.50	7.21	51.98	2.46	6.05
9	6.67	44.48	9.92	98.41	-3.25	10.56
		554.00		321.841		232.159

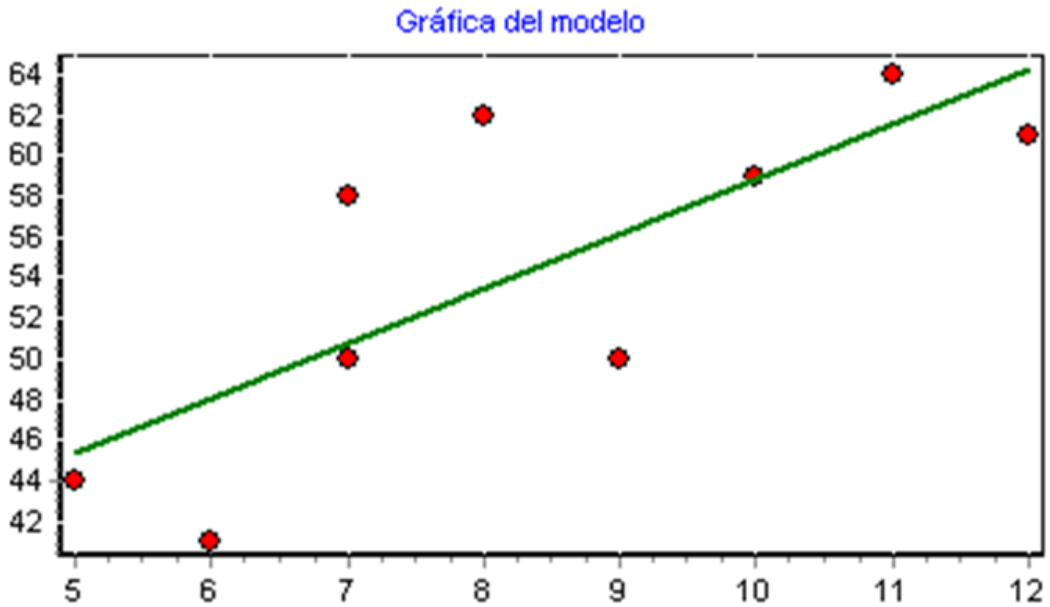
$$\sum_{i=1}^{n}(y_i - \overline{Y})^2 = \sum_{i=1}^{n}(\widehat{y}_i - \overline{Y})^2 + \sum_{i=1}^{n}(y_i - \widehat{y})^2$$

Modelo de regresión
Planteamiento inicial

Modelo de regresión lineal simple

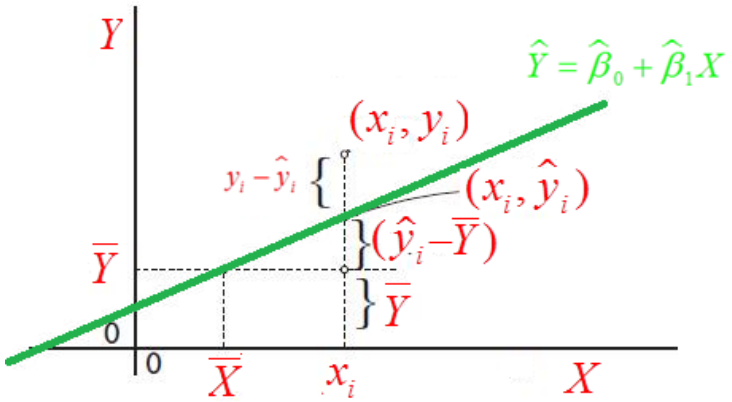
$$Y = \mu(X) + \epsilon = \beta_0 + \beta_1 X + \epsilon$$

ϵ v.a. con distribución normal $(0, \sigma^2)$



X	Y
5	44
6	41
7	58
7	50
8	62
9	50
10	59
11	64
12	61

Hipótesis
Análisis de la varianza



$$\widehat{y} = 31.705 + 2.70X$$

Hipótesis sobre el parámetro β_1

$H_0 : \beta_1 = 0$

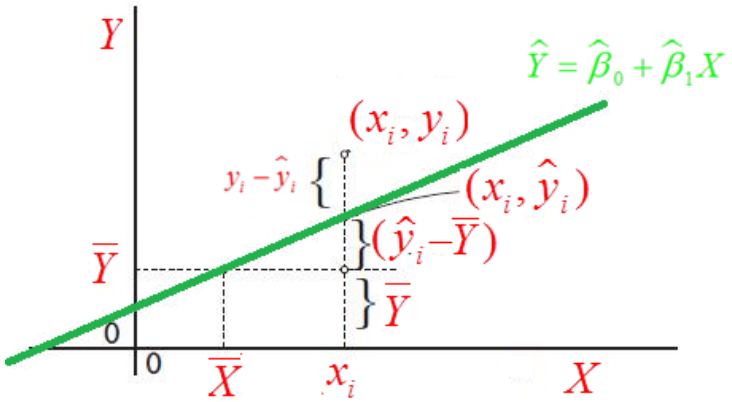
$H_1 : \beta_1 \neq 0$

F de V	SC	gl	CM	RV
Modelo	321.841	1	321.841	9.704
Error	232.159	7	33.166	
Total	554.000	8		

Tabla del Análisis de la Varianza
SC: Suma de cuadrados, CM: Cuadrado medio

RV es la variable razón de varianzas y su distribución es F con $gl(num = 1)$ y $gl(den = 7)$

Hipótesis
Análisis de la varianza



$$\widehat{y} = 31.705 + 2.70X$$

Hipótesis sobre el parámetro β_1

$H_0 : \beta_1 = 0$

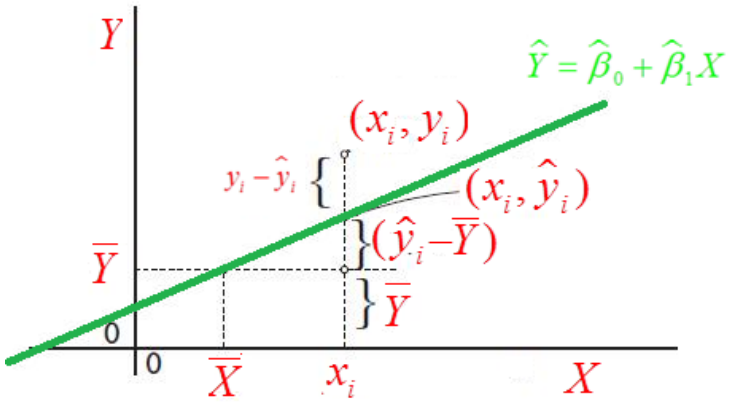
$H_1 : \beta_1 \neq 0$

F de V	SC	gl	CM	RV
Modelo	321.841	1	321.841	9.704
Error	232.159	7	33.166	
Total	554.000	8		

Tabla del Análisis de la Varianza
SC: Suma de cuadrados, CM: Cuadrado medio

RV es la variable razón de varianzas y
su distribución es F con $gl(num = 1)$ y $gl(den = 7)$

Hipótesis
Análisis de la varianza



$$\widehat{y} = 31.705 + 2.70X$$

Hipótesis sobre el parámetro β_1

$H_0 : \beta_1 = 0$

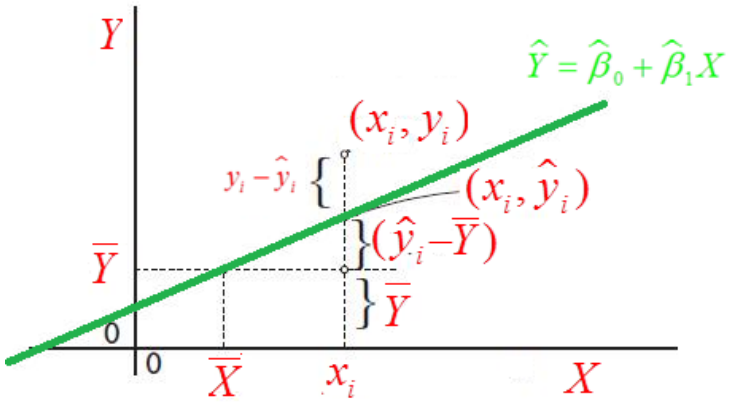
$H_1 : \beta_1 \neq 0$

F de V	SC	gl	CM	RV
Modelo	321.841	1	321.841	9.704
Error	232.159	7	33.166	
Total	554.000	8		

Tabla del Análisis de la Varianza
SC: Suma de cuadrados, CM: Cuadrado medio

RV es la variable razón de varianzas y
su distribución es F con $gl(num = 1)$ y $gl(den = 7)$

Hipótesis
Análisis de la varianza



$$\widehat{y} = 31.705 + 2.70X$$

Hipótesis sobre el parámetro β_1

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

F de V	SC	gl	CM	RV
Modelo	321.841	1	321.841	9.704
Error	232.159	7	33.166	
Total	554.000	8		

Tabla del Análisis de la Varianza
SC: Suma de cuadrados, CM: Cuadrado medio

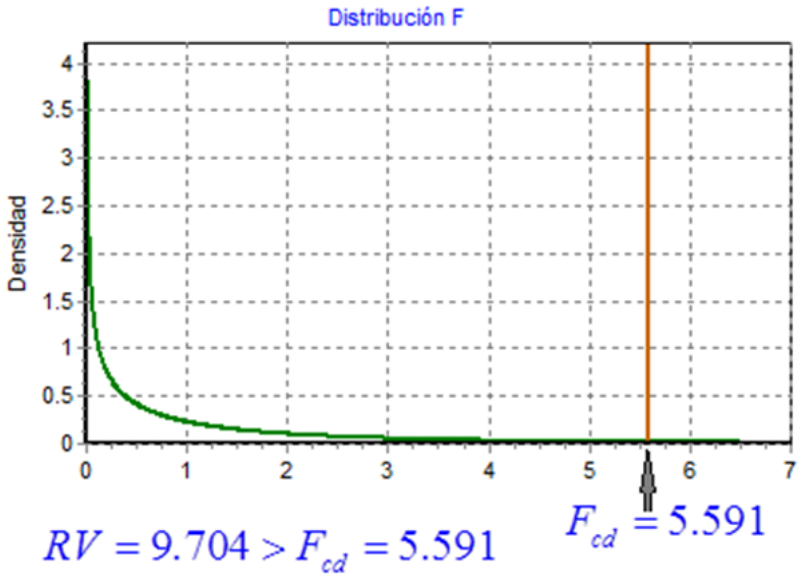
RV es la variable razón de varianzas y
su distribución es F con $gl(num = 1)$ y $gl(den = 7)$

Prueba de hipótesis del modelo de regresión

Estadístico RV

Estadístico RV

$$RV = \frac{\frac{SC_{Modelo}}{1}}{\frac{SC_{Error}}{7}} = \frac{CM_{Modelo}}{CM_{Error}}$$



Regla de decisión

Decisión con el valor crítico: $F_{cd} = 5.591$
Se rechaza la hipótesis nula puesto que $RV = 9.704 > 5.591$
Decisión con la probabilidad: $\alpha = 0.05$
Se rechaza la hipótesis nula ya que $p = 0.0170 < \alpha = 0.05$
Donde $p = P(RV \geq 9.704) = 0.017$

Nota. Puede consultar valores de las distribuciones en www.calest.com/CalculadoraDist.aspx

Prueba de hipótesis del modelo de regresión

Alternativa con el estadístico $t - Student$

Hipótesis sobre el parámetro β_1

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

Estadístico $\widehat{\beta}_1$

$\widehat{\beta}_1 = \beta_1 + t(1-\alpha, gl)ES(\widehat{\beta}_1)$

Valores para $\frac{\alpha}{2}$ y $1 - \frac{\alpha}{2}$

$\widehat{\beta}_{1ci} = \beta_1 + t(\frac{\alpha}{2}, gl)ES(\widehat{\beta}_1)$

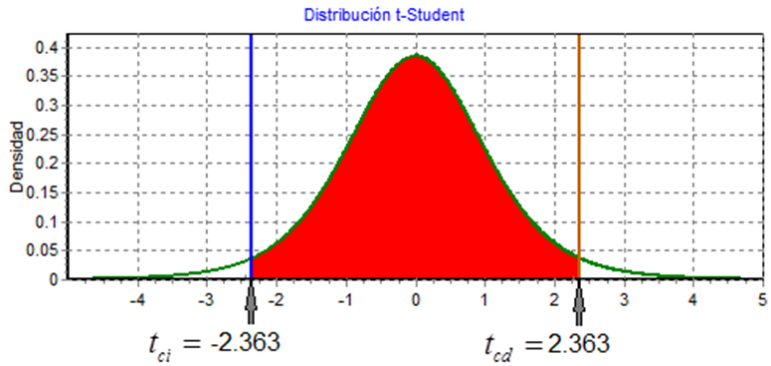
$\widehat{\beta}_{1cd} = \beta_1 + t(1-\frac{\alpha}{2}, gl)ES(\widehat{\beta}_1)$

Estadístico t-student

$t_{ci} = \frac{\widehat{\beta}_{1ci} - \beta_1}{ES(\widehat{\beta}_1)}$

$t_{cd} = \frac{\widehat{\beta}_{1cd} - \beta_1}{ES(\widehat{\beta}_1)}$

$ES(\widehat{\beta}_1) = \sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{X})^2}}$



$t = \frac{\widehat{\beta}_{1ci} - \beta_1}{ES(\widehat{\beta}_1)} = \frac{2.708}{0.868}$

Se rechaza la hipótesis nula puesto que $3.1 > 2.363$

Prueba de hipótesis del modelo de regresión

Alternativa con el estadístico $t - Student$

Hipótesis sobre el parámetro β_1

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

Estadístico $\widehat{\beta}_1$

$\widehat{\beta}_1 = \beta_1 + t(1-\alpha, gl)ES(\widehat{\beta}_1)$

Valores para $\frac{\alpha}{2}$ y $1 - \frac{\alpha}{2}$

$\widehat{\beta}_{1ci} = \beta_1 + t(\frac{\alpha}{2}, gl)ES(\widehat{\beta}_1)$

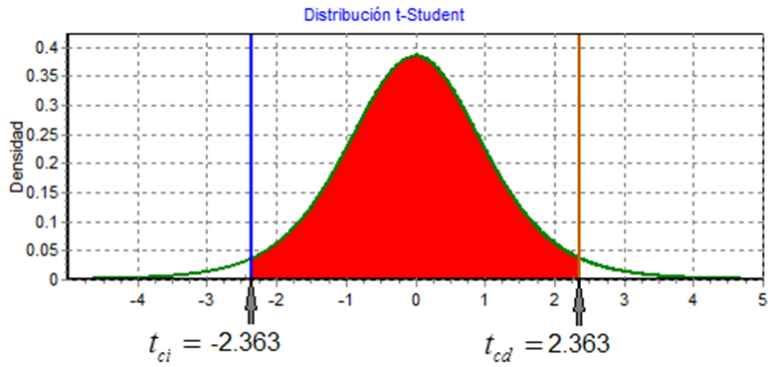
$\widehat{\beta}_{1cd} = \beta_1 + t(1-\frac{\alpha}{2}, gl)ES(\widehat{\beta}_1)$

Estadístico t-student

$t_{ci} = \frac{\widehat{\beta}_{1ci} - \beta_1}{ES(\widehat{\beta}_1)}$

$t_{cd} = \frac{\widehat{\beta}_{1cd} - \beta_1}{ES(\widehat{\beta}_1)}$

$ES(\widehat{\beta}_1) = \sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{X})^2}}$



$t = \frac{\widehat{\beta}_{1ci} - \beta_1}{ES(\widehat{\beta}_1)} = \frac{2.708}{0.868}$

Se rechaza la hipótesis nula puesto que $3.1 > 2.363$

Prueba de hipótesis del modelo de regresión

Alternativa con el estadístico $t - Student$

Hipótesis sobre el parámetro β_1

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

Estadístico $\widehat{\beta}_1$

$\widehat{\beta}_1 = \beta_1 + t(1-\alpha, gl)ES(\widehat{\beta}_1)$

Valores para $\frac{\alpha}{2}$ y $1 - \frac{\alpha}{2}$

$\widehat{\beta}_{1ci} = \beta_1 + t(\frac{\alpha}{2}, gl)ES(\widehat{\beta}_1)$

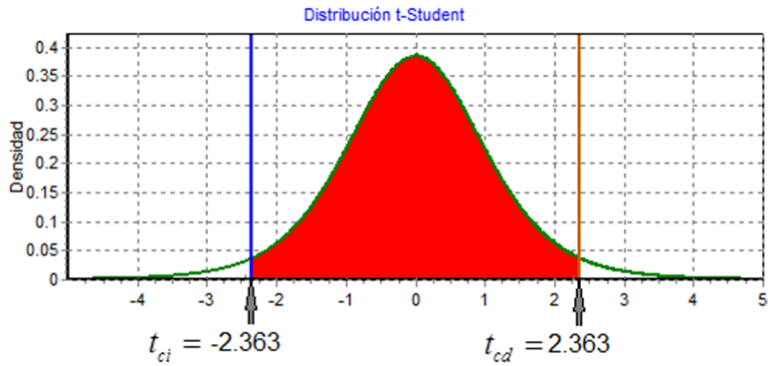
$\widehat{\beta}_{1cd} = \beta_1 + t(1-\frac{\alpha}{2}, gl)ES(\widehat{\beta}_1)$

Estadístico t-student

$t_{ci} = \frac{\widehat{\beta}_{1ci} - \beta_1}{ES(\widehat{\beta}_1)}$

$t_{cd} = \frac{\widehat{\beta}_{1cd} - \beta_1}{ES(\widehat{\beta}_1)}$

$ES(\widehat{\beta}_1) = \sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{X})^2}}$



$$t = \frac{\widehat{\beta}_{1ci} - \beta_1}{ES(\widehat{\beta}_1)} = \frac{2.708}{0.868}$$

Se rechaza la hipótesis nula puesto que $3.1 > 2.363$

Modelo de regresión
Estimación de los parámetros con R

En esta parte se dan las ideas principales usando R para la estimación de los parámetros de regresión del ejemplo discutido aquí. Varias de los elementos de programación mediante R se ha descrito desde el capítulo 1. En lo siguiente ubiquese en consola de R

Programación en R

```
> TEM<-scan()
1: 5 6 7 7 8 9 10 11 12
10:
Read 8 items
> RES<-scan()
1: 44 41 58 50 62 50 59 64 61
10:
Read 9 items
> var<-cbind(TEM,RES) Junta las variables TEM y RES en dos columnas
> dump("var",file=".....") Guarda las varibles en un archivo
> source(mco) Llama a la función que estima los parámetros
> mco(TEM,RES) Se aplica la función mco
Solución [1] 31.795 2.705
```

script: Mínimos Cuadrados: mco

$$
\begin{aligned}
&mco < -function(x, y)\\
&\{\\
&scx = sum((x - mean(x))^2)\\
&scxy = sum(((x - mean(x)) * (y - mean(y)))\\
&beta1 = \frac{scxy}{scx}\\
&beta0 = mean(y) - beta1 * mean(x)\\
&c(beta0, beta1)\\
&\}
\end{aligned}
$$

Nota. El análisis de la varianza y la solución de otros ejemplos se presentan en la lección 8.2