4.2.3 Fuentes de variación en la regresión lineal

Vamos a estudiar ahora las fuentes de variabilidad que intervienen en **el análisis de regresión.**

Si solamente consideramos la variable y como ya hemos visto, su variabilidad se mide por la dispersión de los valores de y alrededor de su media y . Sin embargo, al relacionar la variable y con la variable x estamos tratando de explicar la variabilidad de y a partir de una cierta relación lineal que la liga con una variable x.

La relación estimada de y con x está dada por $y = a + b\,x$.

Entonces, podemos pensar que la variabilidad total de y puede ser analizada a través de dos partes componentes: la variabilidad de y alrededor de la línea de regresión estimada y y la variabilidad representada por las diferencias entre la recta de regresión estimada y y la media y.

0 sea:

$$y_i - \bar{y} = \hat{y}_i - \hat{y}_i + y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

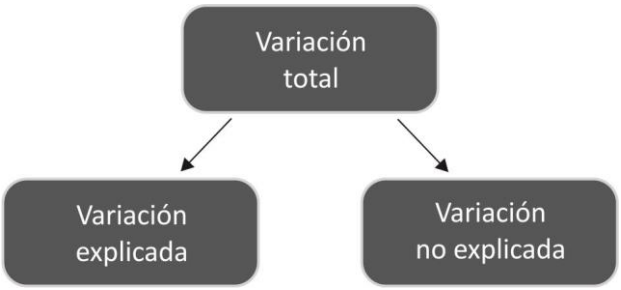
Desvío de y alrededor de su medida	Desvío de y_i alrededor de la recta de regresión estimada	Desvío de la recta alrededor de su media

Ahora bien, las diferencias $y - y_i$ es lo que hemos denominado elresiduo o error aleatorio y este permanece aún después de estimar la recta de regresión.

No olvidemos que $y_i - y_i$ las diferencias que se observan entre los valores de y_i y la recta de regresión estimada. Debido a esta situación, esta Variabilidad se conoce con el nombre de **variación no explicada**.

Al encontrarse una relación lineal entre la variable aleatorio y y la variable x ya no consideramos la variabilidad de y con respecto a su media y sino que la estudiamos con respecto a la recta de regresión $y = a + b\,x$.

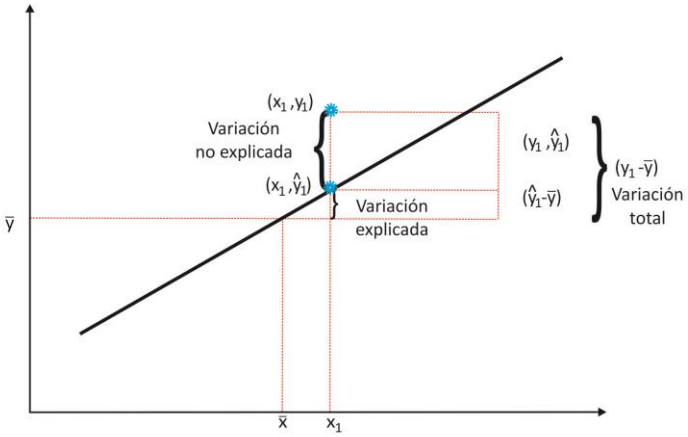
Entonces decimos que la variabilidad expresada por las diferencias $y_i - y$ ya ha sido explicada por la recta de regresión. Debido a ello, esta variabilidad se denomina **variación explicada**.



De acuerdo a lo expresado, podemos rescribir el diagrama anterior de la siguiente manera:
Variación total = Variación no explicada + Variación explicada

$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

Gráficamente tenemos:



Puede demostrarse que las dos fuentes de variación en que se descompone la variabilidad total son independientes por lo cual si elevamos las diferencias al cuadrado y las sumamos, se mantiene la relación:

variación explicada + variación no explicada = variación total.

En símbolos:

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + (\hat{y}_i - \bar{y})^2$$

Hemos obtenido, de esta manera, tres sumas de cuadrados.

La suma de cuadrados ubicada en el lado izquierdo de la igualdad se denomina **suma de cuadrados total (SCT)**.

El primer término del lado derecho de la igualdad no es más que la suma de los residuos o errores aleatorios e y por ello se la denomina **suma de cuadrados de error (SCe)**.

El segundo término que representa la parte de la variabilidad total explicada por la regresión se denomina **suma de cuadrados debida a la regresión de y sobre x (SCR)**.

En síntesis:

SCT = SCR + SCe

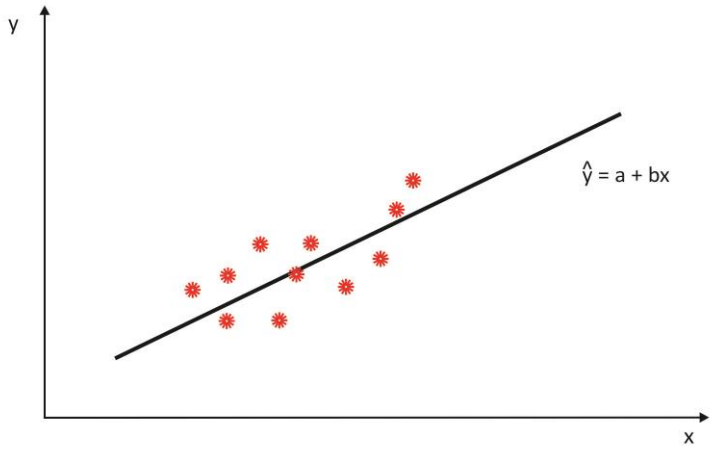
Luego volveremos a usar estos conceptos.

Una vez establecidas las diferentes **fuentes de variabilidad** que intervienen en un **análisis de regresión**, estudiaremos el **error estándar de la estimación de una recta de regresión**.

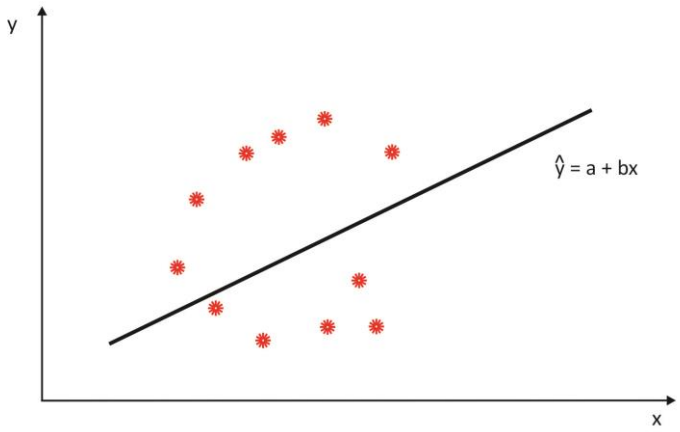
4.2.4 Error estándar de la regresión

Es evidente que una medida de la variabilidad de los puntos observados alrededor de la recta de regresión nos permitirá una determinación respecto a si la recta es o no un buen estimador de la relación que liga a las variables **y** y **x**.

Si la relación entre x e y se da en el siguiente gráfico.



es evidente que la estimación o predicción de y será más precisa (menos variable) que si la relación se diera el siguiente gráfico.



También dijimos que las diferencias entre los valores observados de y y la recta de regresión estimada están representados por los valores e .

Entonces, la variabilidad alrededor de la recta de regresión puede ser evaluada por la suma de cuadrados de error. Aunque sin demostrarlo, diremos que esta suma de cuadrados tiene n - 2 grados de libertad.

Si dividimos esta suma de cuadrados por sus correspondientes gradosde libertad, obtenemos el cuadrado medio del error residual, el cual, por otra parte, es también una estimación de la varianza poblacional σ .

La varianza de los errores es:

$$S_e^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n - 2}$$

donde los y_i se calculan mediante la ecuación $y_i = a + b\, x_i$ y los y_i son los valores observados de y.

También es posible calcular S_e^2 mediante la siguiente "fórmula de cálculo":

$$S_e^2 = \frac{\sum y_i^2 - a \sum y_i - b \sum x_i y_i}{n - 2}$$

aunque a veces resultan pequeñas diferencias por redondeo.

Calculamos ahora la varianza de error con los datos del ejemplo.

$$S_e^2 = \frac{0{,}31431}{18} = 0{,}01746$$

$$S_e = 0{,}1321$$

Hasta aquí hemos estimado la recta de regresión y presentado una medida de la variabilidad de los puntos alrededor de la recta estimada.

4.3 Verificación de supuestos

A continuación veremos cómo podemos comprobar el cumplimiento de los supuestos del modelo de regresión y para ello utilizaremos el programa de computación SPSS.

El análisis de los supuestos del modelo de regresión se estudia a través del comportamiento de los residuales.

Ya hemos dicho que un residual se define como:

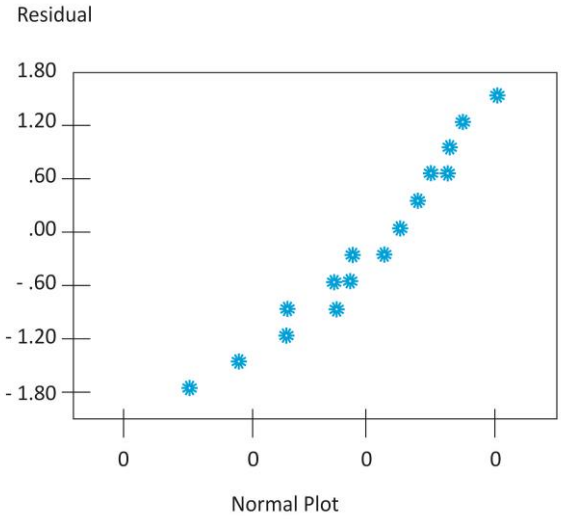
$$e_i = y_i - \hat{y}_i \qquad i = 1, \ldots, n$$

Analizando los residuales se puede llegar a comprobar:

a) si la distribución de e es aproximadamente normal.
b) si su variabilidad es constante.
c) si existe alguna evidencia de una relación no lineal entre las variables.
d) si existen observaciones atípicas (outliers).
e) si los residuales o errores aleatorios se distribuyen independientemente.

El análisis de los supuestos referidos a los residuales se efectúa gráficamente.

a) El supuesto de normalidad de los residuales se puede verificar utilizando un plot normal. Utilizando la sentencia "EXAMINE" para el ejemplo de las universidades, tenemos:



Observando el gráfico vemos que los residuales se distribuyen alrededor de la línea de 45°.

b) El supuesto de varianza constante se comprueba utilizando un gráfico en el que el eje de las abscisas representa a los valores estimados de y y el eje de las ordenadas a los residuales. Tanto los valores estimados de y, como los residuales aparecen estandarizados. En regresión se trabaja con las variables estandarizadas para evitar el sesgo que introducen las distintas unidades de medida.

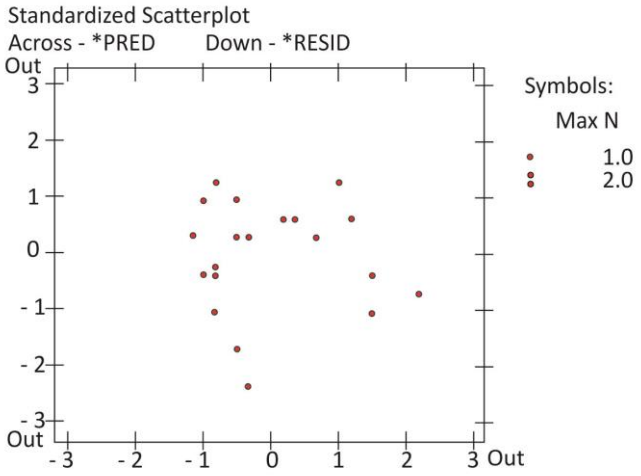
Un residual estandarizado es el residual dividido por su error estándar.

$$\text{Residual estandarizado} = \frac{h_i}{S_{y/x}\sqrt{1 - h_i}} = 0{,}954$$

siendo:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

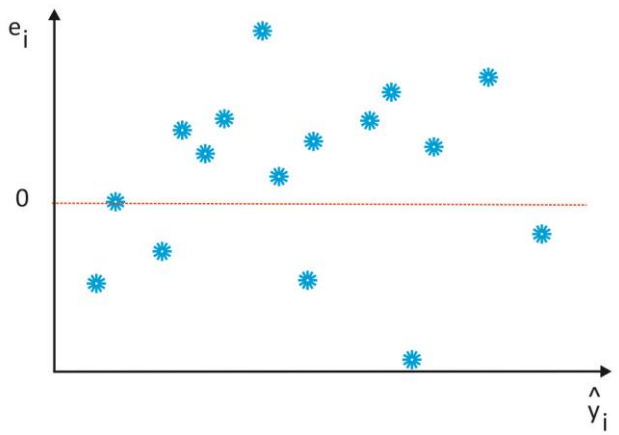
Utilizando la sentencia "REGRESSION" en el ejemplo obtuvimos el siguiente gráfico:



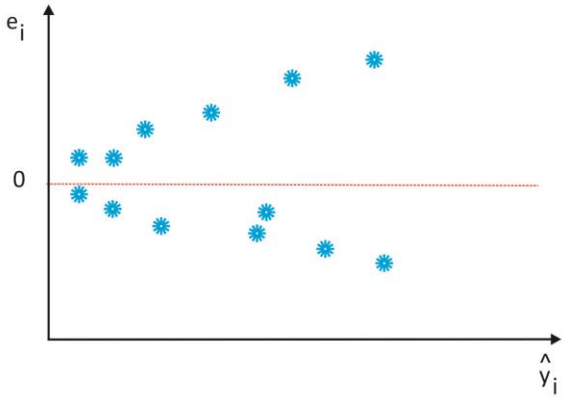
Si la varianza de los e es constante se debería esperar que la variabilidad de los residuales se distribuyera aleatoriamente.

Vemos que, en nuestro ejemplo, los residuales se distribuyen aleatoriamente alrededor de 0, sin mostrar evidencia de una mayor amplitud a medida que los valores de y estimados se alejan del 0. En general, si la varianza permanece constante, el gráfico debe presentar una banda de amplitud uniforme.

El gráfico es:



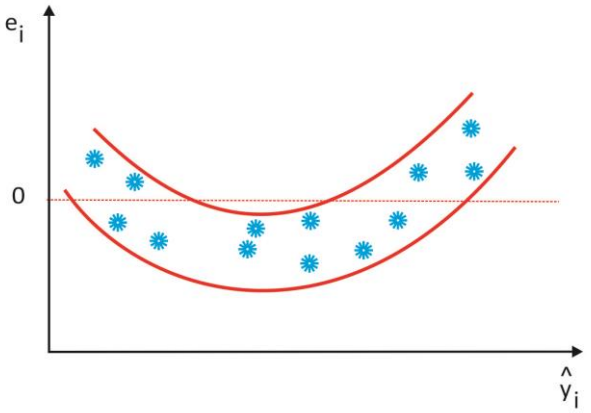
En cambio, si se hubiera obtenido el siguiente gráfico:



habría una indicación de que la varianza de los residuales se incrementa a medida que se incrementa x. A medida que la variable toma valores más grandes se observa una mayor variabilidad en la variable y.

Un **ejemplo** de esta situación podría ser el caso en que se estuvieran analizando observaciones dadas en un cierto orden cronológico y la varianza se fuera incrementando a medida que pasa el tiempo.

c) **Independencia de los errores con la variable en estudio**: otra situación que se puede presentar al analizar el comportamiento de los residuales es la referente a la presentación de una tendencia lineal. Graficando en un diagrama de dispersión los residuales y los valores estimados de y, puede aparecer el siguiente comportamiento:



Se advierte la violación del supuesto que establece que el error aleatorio es independiente de la observación y_i y en consecuencia, se debe dudar de la aleatoriedad de la variable y.

d) **Detección de outliers**: algunos residuales pueden ser mucho más grandes que los demás, considerando su magnitud en valor absoluto. Existen procedimientos de test de hipótesis para determinar si una observación es un outlier, pero una regla práctica, aplicable cuando n > 20, aconseja rechazar un residual cuando el residual estandarizado es mayor que 3.

La siguiente tabla presenta los 10 valores de residuales estandarizados más grandes para el ejemplo dado. Los mismos fluctúan dentro de los límites - 3 y 3.

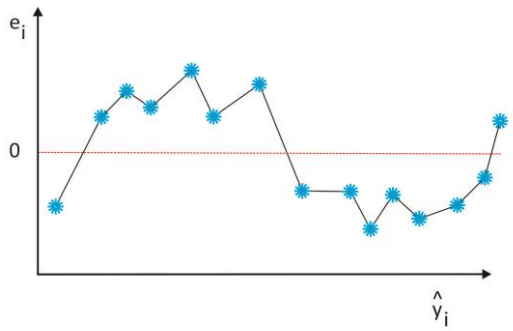
Outliers - Standardized Residual	
Case #	*ZRESID
16	- 2.20915
15	- 1.68793
3	1.47412
6	1.23607
7	- 1.07655
18	- 1.01903
17	.93221
2	.89265
8	.67044
10	.62145

e) **Test de Durbin Watson**:

También gráficamente se puede estudiar el comportamiento de los residuales a través del tiempo.

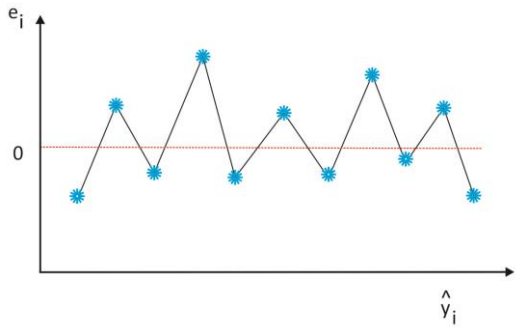
Cuando se trabaja con series cronológicas es muy importante efectuar un gráfico contrastando los residuales e_i versus y y_i para detectar la presencia de correlaciones positivas y negativas entre errores calculados en tiempos consecutivos.

Veamos el siguiente gráfico:



En él se presenta una correlación positiva entre los e_i .

En cambio, en el siguiente vemos una correlación negativa entre los errores:



Existe un test muy conocido para probar la existencia de correlación serial (autocorrelación) en una serie de tiempo: el **test de Durbin-Watson**.

Si se piensa que los errores están consecutivamente relacionados, se calcula el estadístico de Durbin Watson que simbolizaremos con D como:

$$D = \frac{\sum (e_i - e_{i-1})^2}{\sum e_i^2}$$

Aunque existen tablas que computan la distribución de probabilidad de D, solamente diremos que $0 \leq D \leq 4$ y que un valor de D cercano a 2 está indicando ausencia de autocorrelación en los errores.

A medida que D se acerca al valor 0, se concluye que existe una autocorrelación positiva entre los e_i. El acercamiento de D al valor 4, indica la presencia de una autocorrelación negativa.

El programa SPSS calcula el estadístico de Durbin-Watson que en el ejemplo que venimos estudiando ha dado D = 1,86. Evidentemente es un valor muy cercano a 2 por lo que podemos concluir que no existe autocorrelación entre los residuos. Este resultado es lógico pues no se está analizando una serie cronológica.

A continuación veremos la resolución del problema utilizando el programa SPSS.

En la siguiente tabla se puede ver el resultado del R^2 , el error típico de la estimación $S_{Y/X}$ y el valor del test de Durbin Watson.

Resumen del modelo				
Modelo	R	R Cuadrado	Error típ. de la estimación	Durbin-Watson
1	,976	,952	,13214	1,857

A continuación se observa la tabla de análisis de la varianza. Como la significación es 0,000 < a 0,05, se decide rechazar la hipótesis nula que establece que la pendiente de la recta de regresión es 0.

ANOVA						
Modelo		Suma de cuadrados	g	Media cuadrática	F	Sig.
1	Regresión	6,287	1	6,287	360,047	,000
	Residual	314	18	,017		
	Total	6,601	19			

En el siguiente cuadro se observan las estimaciones de los párametros (a y b) cada una con su correspondiente error estándar o típico. Es de notar que el coeficiente de regresión b nos había dado 0,0001216 y el programa informa con hasta 3 decimales por eso aparece como 0. El valor de a esta redondeado a - 0,019.

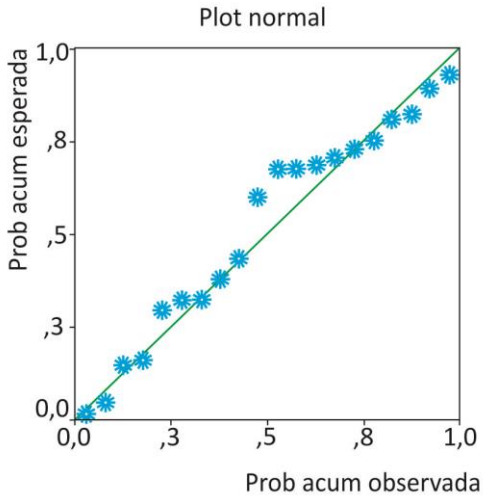
A continuación aparecen los valores de t observado tanto para la ordenada al origen como para la pendiente. La siguiente columna muestra la significación del test de hipótesis.

No se rechaza la hipótesis nula que establece que $\alpha = 0$ (sig. = 0,718) pero si se rechaza la hipótesis que establece que $\beta = 0$ (sig. = 0.0001).

A continuación aparecen los valores extremos de los intervalos de confianza para la estimación de α y β.

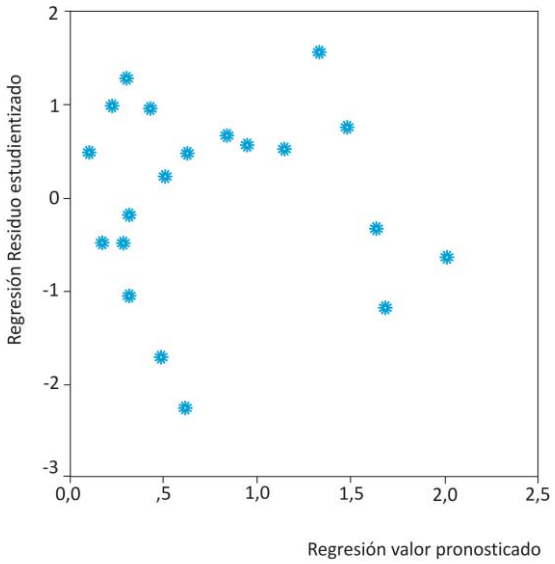
Coeficientes							
		Coeficientes no estandarizados				Intervalo de confianza para B a 95%	
Modelo		B	Error típ.	t	Sig.	Límite inferior	Límite inferior
1	(Constante: a)	-,019	,051	-,366	,718	-,126	,089
	Alumnos (b)	,000	,000	18,975	,000	,000	,000

Este grafico ya visto sirve para considerar el supuesto de normalidad de los residuales así como la detección de valores outliers. Podemos pensar que la distribución es normal.



En el siguiente grafico se prueba la independencia de los residuales así como la homogeneidad de varianzas. No se observa una distribución aleatoria de los residuales alrededor del 0 (no independencia de errores).

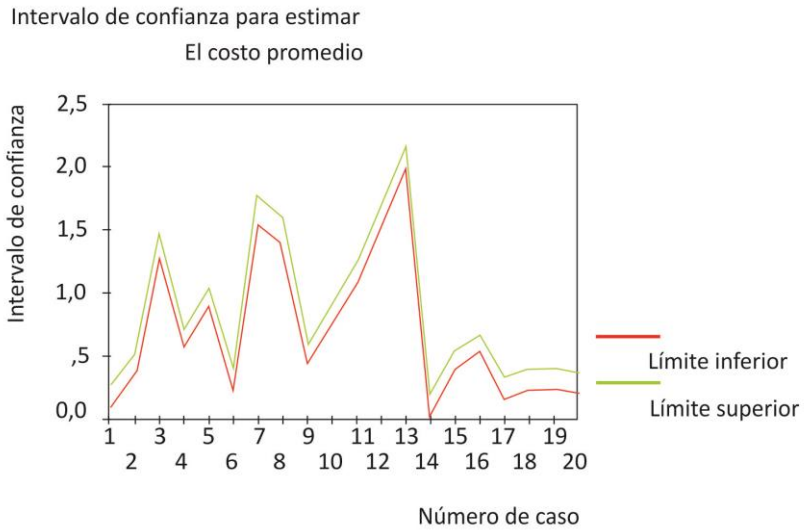
Además se observa una mayor dispersión de los datos para valores chicos y para valores grandes de los valores pronosticados de y.

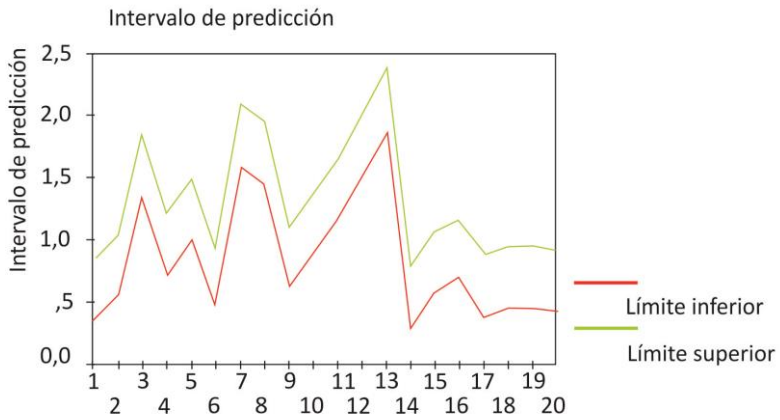


El programa también calcula los valores ajustados, los residuales y los límites superior e inferior de estimación de un promedio y de in valor individual (intervalo de confianza de estimación e intervalo de predicción).

Valores y ajustados	Residuales estandarizados	Limite inferior estim. prom.	Límite superior estim. prom.	Límite inferior estim. indiv.	Límite superior estim. indiv.
-1,04843	- ,49198	,08023	,26258	- ,12080	,46361
-,57795	92443	36987	,51421	,15520	,72889
1,02690	1,55863	1,27503	1,45538	1,07331	1,65710
-,25034	,46282	,56640	,69459	,34557	,91542
,31779	,56602	,89201	1,02260	,67211	1,24250
-,77853	1,29002	,24721	,40611	,03790	,61543
1,52591	- 1,18349	1,53694	1,76758	1,35164	1,95288
1,24629	,71951	1,39064	1,59218	1,19606	1,78675
-,46001	,23520	,44124	,57853	,22390	,79587
,12757	,63788	,78527	,91049	,56329	1,13247
,64539	,50469	1,07130	1,22021	,85832	1,43318
1,47202	- ,34146	1,50881	1,73370	1,32173	1,92078
2,11793	-,65208	1,84431	2,14130	1,67797	2,30764
-1,13615	,47581	,02561	,21629	- ,17259	,41448
-,52405	- 1,74511	,40256	,54353	,18662	,75947
-,30001	- 2,27221	,53697	,66687	,31680	,88704
-,93472	,98045	,15080	,32283	- ,05382	,52745
-,81679	- 1,06538	,22366	,38565	,01546	,593850
-,79692	- ,20620	,23590	,39627	,02712	,60505
-,85589	- ,49276	,19955	,36478	- ,00749	,57182

Con estos datos fueron construidas las fajas de confianza de estimación.





Para ello debemos establecer el error estándar de la estimación que se calcula por medio de la siguiente fórmula:

$$S_y = S_e \sqrt{[1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}]}$$

Haremos la misma estimación pero ahora utilizando un intervalo de confianza. Para ello debemos calcular S_y.

$$S_y = 0{,}1321 \sqrt{[1 + \frac{1}{20} + \frac{(15.000 - 6524{,}45)^2}{425317485}]} = 0{,}1321 \sqrt{1 + 0{,}05 + 0{,}168} = 0{,}1321 \bullet 1{,}10 = 0{,}145$$

El intervalo es el siguiente:

$$P(\hat{y}_i - t_{n-2;\, 1-\alpha/2} \,.\, S_y \leq \hat{Y}_i \leq \hat{y}_i + t_{n-2;\, 1-\alpha/2} \,.\, S_y) = 1 - \alpha$$

donde y y es la estimación puntual que acabamos de calcular. Reemplazando convenientemente, tenemos:

$$P(1{,}81 - 2{,}101 \,.\, 0{,}145 \leq \hat{Y}_i \leq 1{,}81 + 2{,}101 \,.\, 0{,}145) = 0{,}95$$

$$P(1{,}81 - 0{,}305 \leq \hat{Y}_i \leq 1{,}81 + 0{,}305) = 0{,}95$$

$$P(1{,}50 \leq \hat{Y}_i \leq 2{,}11) = 0{,}95$$

El costo mensual es un valor comprendido en el intervalo $ 1,50, $ 2,11.