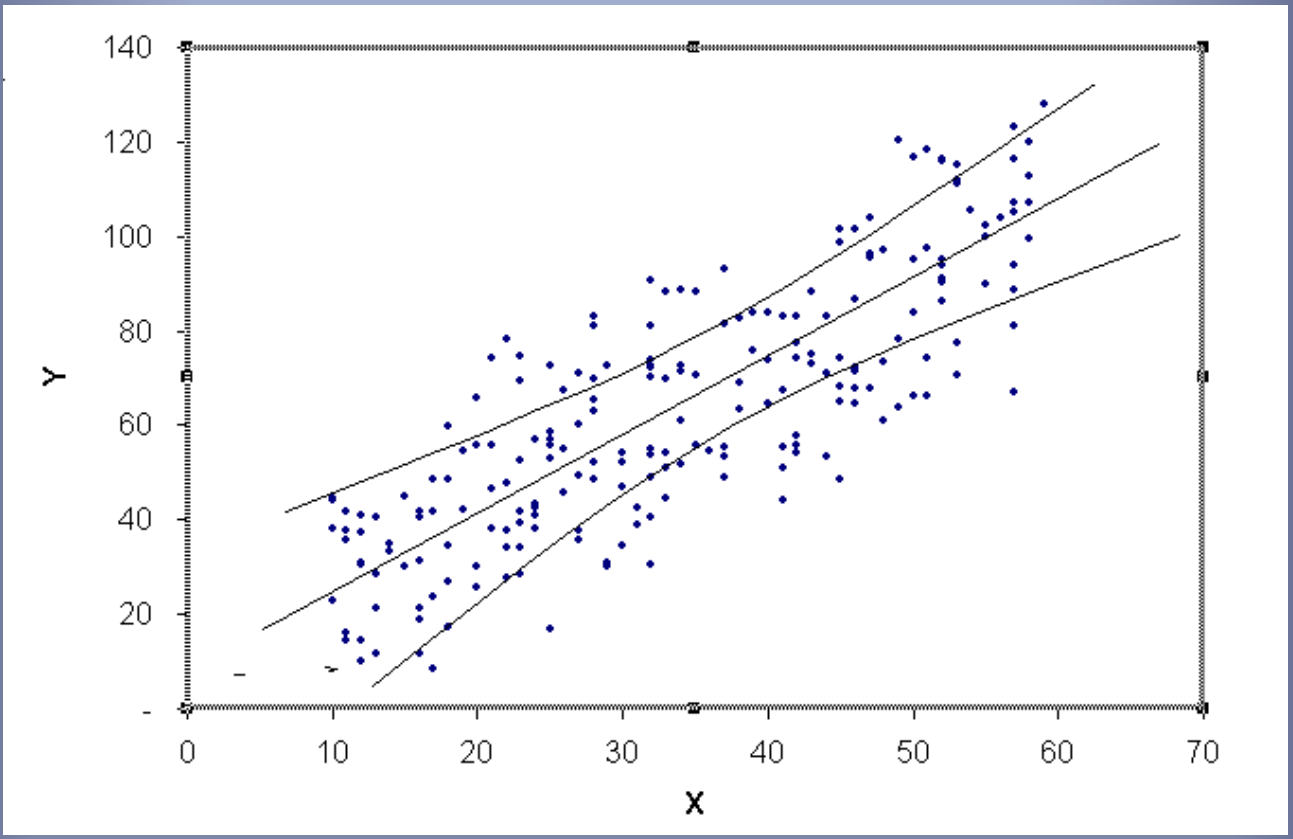


Análisis de Regresión Simple

Notas de Clase

Profr. Francisco Sánchez Villarreal 2021

ANALISIS DE REGRESIÓN SIMPLE

1. INTRODUCCION

El término "regresión" fue usado por primera vez en 1886 cuando Francis Galton publicó un artículo en el que mencionaba que aunque los padres altos tienen en general hijos altos, la estatura de éstos tiende a ser menor que la de sus padres y padres bajos tienen en general hijos bajos, pero la estatura de éstos tiende a ser mayor que la de sus padres. Es decir, existe una tendencia a "regresar" a la estatura media de la población. Tal enunciado conocido como la Ley de Regresión Universal de Galton, fue confirmada por Karl Pearson mediante el análisis de más de 1000 casos.

Actualmente el análisis de regresión es una técnica estadística que trata de la relación de una variable dependiente en función de una o más variables independientes.

2. MODELO LINEAL DE DOS VARIABLES

Supóngase, por ejemplo, que se desea investigar la relación que existe entre el gasto en consumo familiar semanal y el ingreso disponible o la dependencia del producto de una cosecha de la temperatura media, las lluvias y la fertilización. En ambos casos la relación es de naturaleza estadística, ya que las variables explicativas no permiten hacer una predicción precisa del gasto en el primer caso o del producto cosechado en el segundo, debido primordialmente a la existencia de muchos factores que no se incorporan en la relación funcional.

Considérese, en una forma un tanto simplista, que se identifica por Y el gasto y por X el ingreso de una población de familias. Ambas variables se suponen, pueden asociarse de acuerdo a un modelo lineal como el siguiente:

$$Y = \alpha + \beta X + u \qquad (2.0.1)$$

Donde α es la ordenada al origen, β es la pendiente y u es un error aleatorio.

En forma esquemática cada familia identificada por una pareja ordenada (Xi,Yi) se representa por un punto en el plano y el conjunto forma una "nube" tal como se ilustra en la fig. 1.

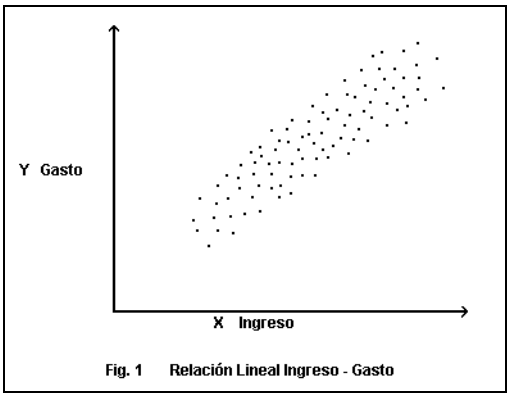


Fig. 1 Relación Lineal Ingreso - Gasto

Podemos suponer que la relación entre el ingreso y el gasto es aceptablemente lineal, pues en el modelo propuesto se han omitido multitud de factores que afectan individualmente la relación. Esos factores actuarán unos en favor y otros en contra de modo que si en conjunto se representa su influencia por u, es razonable suponer un efecto de cancelación y que en consecuencia

$$\mathbf{E(u) = 0} \qquad (2.0.2)$$

Como u es la suma de múltiples variables y tiende a concentrarse en torno a cero, también es razonable suponer que se distribuye normalmente con media 0 y varianza σ_u^2.

$$U \approx N\left(0, \sigma_u^2\right) \qquad (2.0.3)$$

Como un supuesto adicional se considerará que la varianza σ_u^2 es constante e independiente de X (Fig.2) y que distintos valores de u son independientes entre sí, es decir que no covarían.

$$E(u_i u_j) = \begin{cases} o & i \neq j \\ \sigma_u^2 & i = j \end{cases} \qquad (2.0.4)$$

2.1 Los Estimadores de Mínimos Cuadrados Ordinarios.

Supóngase que se cuenta con una muestra de n pares de observaciones (X,Y) y que se desea contar con estimadores de α y β a los que denotaremos por $\hat{\alpha}$ y $\hat{\beta}$. El valor de la variable dependiente Y puede expresarse en la siguiente forma

$$Y_i = \hat{\alpha} + \hat{\beta} X_i + e_i \qquad \text{i=1,2,...,n} \qquad (2.1.1)$$

Donde el residual e_i es la forma empírica del error teórico u_i.

De acuerdo con el modelo, un estimador de $\mathbf{Y_i}$ tendría la siguiente expresión

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}\, X_i \qquad (2.1.2)$$

A partir de (2.1.1) y (2.1.2) es posible concluir que el residual es igual a la siguiente diferencia:

$$\boxed{e_i = Y_i - \hat{Y}_i} \qquad (2.1.3)$$

Esta relación se interpreta con mayor facilidad mediante la observación de la figura 2.

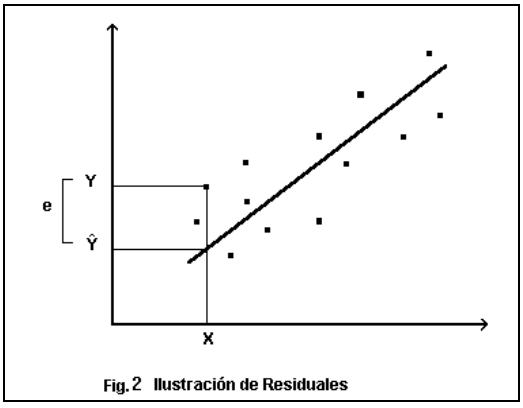


Fig. 2 Ilustración de Residuales

Un modelo ajustado deseable, sería aquel que minimizara los residuales **ei**, en consecuencia, como criterio de optimización se tomará aquel procedimiento de estimación que minimice la suma de cuadrados de los residuales

$$\sum_{i=1}^{n} e_i^2 = min \qquad (2.1.4)$$

Para lograr la minimización se toma una expresión alternativa de (2.1.4) y se procede a minimizar, con el procedimiento de la primera derivada igualada a cero.

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

$$= \sum \left(Y_i - \hat{\alpha} - \hat{\beta} X_i\right)^2$$

La última expresión se deriva parcialmente respecto a $\hat{\alpha}$

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \hat{\alpha}} = -2\sum_{i=1}^{n}\left(Y_i - \hat{\alpha} - \hat{\beta} X_i\right)$$

La derivada se iguala a cero, se distribuye la suma y se despeja $\sum_{i=1}^{n} Y_i$

$$-2\sum_{i=1}^{n}\left(Y_i - \hat{\alpha} - \hat{\beta} X_i\right) = 0$$

$$\sum_{i=1}^{n} Y_i = n\hat{\alpha} + \hat{\beta}\sum_{i=1}^{n} X_i$$

Ahora el procedimiento de derivación se aplica en forma análoga para $\hat{\beta}$

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \hat{\beta}} = -2\sum_{i=1}^{n}\left(Y_i - \hat{\alpha} - \hat{\beta} X_i\right)X_i$$

$$= -2\sum_{i=1}^{n}\left(Y_i - \hat{\alpha} - \hat{\beta} X_i\right)X_i = 0$$

$$\sum_{i=1}^{n} X_i Y_i = \hat{\alpha}\sum_{i=1}^{n} X_i + \hat{\beta}\sum_{i=1}^{n} X_i^2$$

Finalmente se tiene un sistema de dos ecuaciones lineales en los estimadores, conocido como el sistema de ecuaciones normales.

$$\begin{aligned} \sum_{i=1}^{n} Y_i &= n\hat{\alpha} + \hat{\beta}\sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} X_i Y_i &= \hat{\alpha}\sum_{i=1}^{n} X_i + \hat{\beta}\sum_{i=1}^{n} X_i^2 \end{aligned} \qquad (2.1.5)$$

Al resolverlo para $\hat{\alpha}$ y $\hat{\beta}$ se obtienen las expresiones de los estimadores de mínimos cuadrados.

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} Y_i \sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} X_i Y_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} \qquad (2.1.6)$$

$$\hat{\beta} = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}$$

Ejemplo 1.

Considere un experimento en el que se dispone de 12 parcelas cuyo suelo se ha enriquecido con la aplicación de diversas dosis de fertilizante (X). La variable de respuesta es el rendimiento en Bushels por Acre (Y). Se pretende conocer la relación lineal que existe entre ambas variables. Se presenta el diagrama de dispersión y el cuadro con los datos originales y las columnas para los cálculos auxiliares.

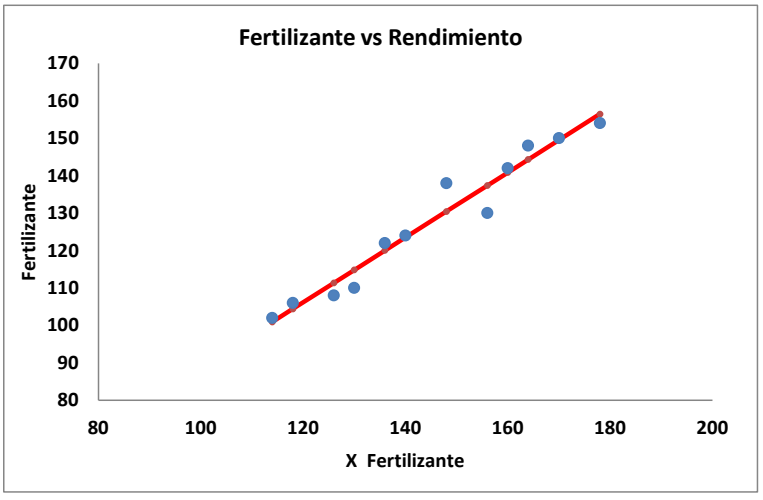
Parcela	Fertilizante X	Rendimiento Y	XY	X^2	Y est	e	e^2
1	114.0	102.0	11,628.0	12,996.0	100.94	1.06	1.12
2	118.0	106.0	12,508.0	13,924.0	104.41	1.59	2.52
3	126.0	108.0	13,608.0	15,876.0	111.35	- 3.35	11.24
4	130.0	110.0	14,300.0	16,900.0	114.82	- 4.82	23.25
5	136.0	122.0	16,592.0	18,496.0	120.03	1.97	3.89
6	140.0	124.0	17,360.0	19,600.0	123.50	0.50	0.25
7	148.0	138.0	20,424.0	21,904.0	130.44	7.56	57.22
8	156.0	130.0	20,280.0	24,336.0	137.37	- 7.37	54.39
9	160.0	142.0	22,720.0	25,600.0	140.84	1.16	1.34
10	164.0	148.0	24,272.0	26,896.0	144.31	3.69	13.59
11	170.0	150.0	25,500.0	28,900.0	149.52	0.48	0.23
12	178.0	154.0	27,412.0	31,684.0	156.46	- 2.46	6.04
Suma	1,740.00	1,534.00	226,604.00	257,112.00	1,534.00	0.00	175.08

Se estiman los parámetros, ordenada al origen y pendiente de la recta ajustada.

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} Y_i \sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} X_i Y_i =}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} = 2.05819$$

$$\hat{\beta} = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} = 0.86741$$

A continuación el diagrama de dispersión y la recta ajustada



Estimaciones Mediante Desviaciones

Se divide la primera ecuación en (2.1.5) entre n se obtiene una ecuación alternativa equivalente.

$$\frac{\sum_{i=1}^{n} Y_i}{n} = \frac{n\hat{\alpha} + \hat{\beta}\sum_{i=1}^{n} X_i}{n}$$

$$\overline{Y} = \hat{\alpha} + \hat{\beta}\,\overline{X} \qquad (2.1.7)$$

Ello significa que la recta estimada pasa por el punto de medias $\left(\overline{X}, \overline{Y}\right)$

A continuación se resta (2.1.7) de (2.1.2)

$$\begin{array}{r} \hat{Y}_i = \hat{\alpha} + \hat{\beta}\, X_i \\ \overline{Y} = \hat{\alpha} + \hat{\beta}\,\overline{X} \\ \hline \hat{Y}_i - \overline{Y} = \hat{\beta}\left(X_i - \overline{X}\right) \end{array} \qquad (2.1.8)$$

Para facilitar las expresiones algebraicas, se definen las siguientes desviaciones:

$$y_i = Y_i - \overline{Y}$$
$$x_i = X_i - \overline{X}$$
$$\hat{y}_i = \hat{Y}_i - \overline{Y}$$

De donde se obtiene otra expresión para los residuales en términos de desviaciones.

$$\begin{aligned} e_i &= Y_i - \hat{Y}_i \\ &= Y_i - \overline{Y} - \hat{Y}_i + \overline{Y} \\ &= \left(Y_i - \overline{Y}\right) - \left(\hat{Y}_i - \overline{Y}\right) \\ &= y_i - \hat{y}_i \end{aligned}$$

A partir de (2.1.8) se obtiene una expresión alternativa que involucra al estimador de la pendiente en función de las desviaciones de las observaciones respecto de su media.

$$\hat{y}_i = \hat{\beta}\ x_i \qquad (2.1.9)$$

Los residuales se pueden expresar alternativamente como diferencia de desviaciones.

$$\begin{aligned} e_i &= y_i - \hat{y}_i \\ &= y_i - \hat{\beta}\ x_i \end{aligned}$$

De donde la suma de cuadrados también tiene una expresión alterna.

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}\left(y_i - \hat{\beta}\, x_i\right)^2 \qquad (2.1.10)$$

Si se minimiza (2.1.10) se obtiene la siguiente ecuación:

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \beta} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}\, x_i\right) x_i = 0$$

De ella, al despejar $\hat{\beta}$ se obtiene su expresión simplificada en términos de desviaciones

$$\boxed{\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}} \qquad (2.1.11)$$

Puesto que $\overline{Y} = \hat{\alpha} + \hat{\beta}\,\overline{\mathrm{X}}$ se estima la ordenada al origen en función de la pendiente y de las medias de las variables involucradas en el modelo.

$$\boxed{\hat{\alpha} = \overline{Y} + \hat{\beta}\,\overline{\mathrm{X}}} \qquad (2.1.12)$$

De la fórmula (2.1.11) se podrá expresar al estimador del coeficiente de regresión como una suma ponderada de los valores observados de Y

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

$$= \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2} - \frac{\sum_{i=1}^{n} x_i \overline{Y}}{\sum_{i=1}^{n} x_i^2}$$

$$= \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2} - \overline{Y}\,\frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2} \qquad \text{La suma } \sum_{i=1}^{n} x_i = 0 \text{ por ser primer momento}$$

$$= \sum_{i=1}^{n} w_i Y_i \qquad (2.1.13)$$

Donde las wi se consideran ponderadores expresados como sigue

$$w_i = \frac{x_i}{\sum_{i=1}^{n} x_i^2}$$

Los ponderadores w_i cumplen las siguientes igualdades:

$$\sum_{i=1}^{n} w_i = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2} = 0$$

$$\sum_{i=1}^{n} w_i^2 = \frac{\sum_{i=1}^{n} x_i^2}{\left(\sum_{i=1}^{n} x_i^2\right)^2} = \frac{1}{\sum_{i=1}^{n} x_k^2} \qquad (2.1.14)$$

$$\sum_{i=1}^{n} w_i x_i = \frac{\sum_{i=1}^{r} x_i^2}{\sum_{i=1}^{n} x_k^2} = 1$$

$$\sum_{i=1}^{n} w_i x_i = \sum_{i=1}^{n} w_i \left(X_i - \bar{X} \right) = \sum_{i=1}^{n} w_i x_i - \bar{X} \sum_{i=1}^{n} w_i = 1$$

En términos de las w_i se proponen nuevas expresiones para $\hat{\alpha}$ y $\hat{\beta}$ a fin de probar algunas propiedades de los estimadores.

$$\begin{aligned} \hat{\alpha} &= \bar{Y} - \hat{\beta}\,\bar{X} \\ &= \bar{Y} - \bar{X} \sum_{i=1}^{n} w_i Y_i \qquad \text{por (2.1.13)} \\ &= \frac{\sum_{i=1}^{n} Y_i}{n} - \bar{X} \sum_{i=1}^{n} w_i Y_i \\ &= \sum_{i=1}^{n} \left(1/n - \bar{X} w_i \right) Y_i \qquad (2.1.15) \end{aligned}$$

Esperanza y Varianza de los Estimadores

A continuación se obtendrán fórmulas para el cálculo de esperanzas y varianzas de $\hat{\alpha}$ y $\hat{\beta}$.

Se parte de la relación (2.1.13)

$$\begin{aligned} \hat{\beta} &= \sum_{i=1}^{n} w_i Y_i \\ &= \sum_{i=1}^{n} w_i \left(\alpha + \beta X_i + u_i \right) \qquad \text{Mediante sustitución de (2.0.1)} \\ &= \alpha \sum_{i=1}^{n} w_i + \beta \sum_{i=1}^{n} w_i X_i + \sum_{i=1}^{n} w_i u_i \\ &= \hat{\beta} + \sum_{i=1}^{n} w_i u_i \qquad (2.1.16) \end{aligned}$$

De donde al tomar esperanza

$$\begin{aligned} E\left(\hat{\beta} \right) &= \beta + \sum_{i=1}^{n} w_i E\left(u_i \right) \qquad \text{Por (2.0.2)} \\ &= \beta \end{aligned}$$

Lo que permite concluir que $\hat{\beta}$ es un estimador insesgado.

De manera análoga, de (2.1.15) y (2.0.1) se verifica que $\hat{\alpha}$ es, también, un estimador insesgado

$$\begin{aligned} \hat{\alpha} &= \sum_{i=1}^{n} \left(1/n - \bar{X} w_i \right) Y_i \\ &= \sum_{i=1}^{n} \left(1/n - \bar{X} w_i \right) \left(\alpha + \beta X_i + u_i \right) \\ &= \alpha + \sum_{i=1}^{n} \left(1/n - \bar{X} w_i \right) u_i \end{aligned}$$

y al tomar esperanza se obtiene

$$\mathrm{E}(\hat{\alpha}) = \alpha + \sum_{i=1}^{n}\left(1/n - \bar{X}w_i\right)E(u_i) = \alpha$$

Para obtener la varianza de $\hat{\beta}$ se parte de (2.1.16)

$$\hat{\beta} = \beta + \sum_{i=1}^{n} w_i u_i$$

$$\hat{\beta} - \beta = \sum_{i=1}^{n} w_i u_i$$

Al utilizar la definición de varianza, se tiene

$$V\left(\hat{\beta}\right) = E\left[\left(\hat{\beta} - \beta\right)^2\right]$$

$$= E\left(\sum_{i=1}^{n} w_i u_i\right)^2$$

$$= E\left(\sum_{i=1}^{n} w_i^2 u_i^2 + 2\sum\sum_{i<j} w_i w_j u_i u_j\right)$$

$$= \sum_{i=1}^{n} w_i^2 E\left(u_i^2 + 2\sum\sum_{i<j} w_i w_j E\left(u_i u_j\right)\right) \qquad \text{por (2.0.4)}$$

$$= \sigma_u^2 \sum_{i=1}^{n} w_i^2$$

$$V\left(\hat{\beta}\right) = \frac{\sigma_u^2}{\sum_{i=1}^{n} x_i^2} \qquad (2.1.17)$$

En forma análoga, es posible obtener las fórmulas para la varianza de la ordenada al origen y de la covarianza de ambos estimadores.

$$V(\hat{\alpha}) = \sigma_u^2 \frac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n} x_i^2} \qquad (2.1.18)$$

$$\mathrm{COV}\left(\hat{\alpha}, \hat{\beta}\right) = -\frac{\bar{X}\sigma_u^2}{\sum_{i=1}^{n} x_i^2} \qquad (2.1.19)$$

2.2 Teorema de Gauss-Markov

"Los estimadores de mínimos cuadrados son los mejores estimadores lineales insesgados (MELI). Es decir, que dentro de la clase de estimadores insesgados son los de mínima varianza."

Para demostrar tal afirmación se partirá de un estimador lineal arbitrario

$$\tilde{\beta} = \sum_{i=1}^{n} C_i Y_i \qquad (2.2.20)$$

Donde $C_i = w_i + d_i$, las w_i son definidas según (2.1.14) y las d_i son constantes arbitrarias.

Para que $\tilde{\beta}$ sea insesgado se deben cumplir ciertas condiciones que se establecerán a continuación.

$$\tilde{\beta} = \sum_{i=1}^{n} C_i\left(\alpha + \beta X_i + u_i\right)$$

$$= \alpha \sum_{i=1}^{n} C_i + \beta \sum_{i=1}^{n} C_i X_i + \sum_{i=1}^{n} C_i u_i$$

La propiedad de insesgamiento implica que $E\left(\tilde{\beta}\right) = \beta$, por lo tanto

$$E\left(\tilde{\beta}\right) = \alpha \sum_{i=1}^{n} C_i + \beta \sum_{i=1}^{n} C_i X_i + \sum_{i=1}^{n} E\left(u_i\right) = \beta$$

Esta propiedad se cumple si a su vez se cumplen las siguientes condiciones:

$$\sum_{i=1}^{n} C_i = 0 \qquad \text{y} \qquad \sum_{i=1}^{n} C_i X_i = 1$$

Como consecuencia de lo anterior se tendrá

$$\sum_{i=1}^{n} C_i = \sum_{i=1}^{n}\left(w_i + d_i\right) = 0 \Rightarrow \sum_{i=1}^{n} d_i = 0$$

$$\sum_{i=1}^{n} C_i X_i = \sum_{i=1}^{n}\left(w_i + d_i\right) X_i = \sum_{i=1}^{n} w_i X_i + \sum_{i=1}^{n} d_i X_i = 1$$

Se cumplen $\sum_{i=1}^{n} d_i X_i = 0$, de donde al cumplirse las condiciones definidas se dispone de una expresión alternativa para el estimador.

$$\tilde{\beta} = \beta + \sum_{i=1}^{n} C_i u_i \qquad (2.2.21)$$

Se procede a expresar la varianza de $\tilde{\beta}$ en términos de las Ci

$$V\left(\tilde{\beta}\right) = E\left(\left(\tilde{\beta} - \beta\right)^2\right)$$

$$= E\left(\left(\sum_{i=1}^{n} C_i u_i\right)^2\right) \qquad \text{al despejar en (2.2.21)}$$

$$= E\left(\sum_{i=1}^{n} C_i^2 u_i^2 + 2 \sum_{i<j} \sum C_i C_j u_i u_j\right)$$

$$= \sigma^2 \sum_{i=1}^{n} C_i^2$$

Ahora se analiza $\sum_{i=1}^{n} C_i^2$ expresada en función de w_i y d_i

$$\sum_{i=1}^{n} C_i^2 = \sum_{i=1}^{n} w_i^2 + \sum_{i=1}^{n} d_i^2 + 2\sum_{i=1}^{n} w_i d_i$$

Pero la suma de los productos de ponderadores por las di es nula.

$$\sum_{i=1}^{n} w_i d_i = \sum_{i=1}^{n} \frac{x_i d_i}{\sum x^2} = \frac{\sum_{i=1}^{n} (X_i - \bar{X}) d_i}{\sum x^2} = \frac{\sum_{i=1}^{n} X_i d_i - \bar{X} \sum_{i=1}^{n} d_i}{\sum x^2} = 0$$

De donde

$$\begin{aligned} V(\tilde{\beta}) &= \sigma \sum_{i=1}^{n} C_i^2 \\ &= \left(\sum_{i=1}^{n} w_i^2 + \sum_{i=1}^{n} d_i^2 \right) \sigma_u^2 \\ &= \frac{\sigma_u^2}{\sum_{i=1}^{n} x_i} + \sigma^2 \sum_{i=1}^{n} d_i^2 \\ &= V(\hat{\beta}) + \sigma^2 \sum_{i=1}^{n} d_i^2 \end{aligned}$$

Es posible concluir que para que $\tilde{\beta}$ sea tan eficiente como $\hat{\beta}$ se requiere que $\sum_{i=1}^{n} d_i^2 = 0$ y ello se cumple solamente que $d_i = 0$ para toda i.

Así $\hat{\beta}$ el estimador de mínimos cuadrados es insesgado y de mínima varianza.

2.3 Estimación de la varianza del error σ^2

La varianza del error σ^2 es un parámetro usualmente desconocido y por ello se justifica el planeamiento de un estimador adecuado.

De (2.1.10) se tiene

$$e_i = y_i - \hat{\beta} x_i$$

por otro lado se obtiene una expresión alterna de y_i

Sea

$$\begin{aligned} Y_i &= \alpha + \beta X_i + u_i \\ \bar{Y} &= \alpha + \beta \bar{X} + \bar{u} \end{aligned}$$

la diferencia

$$Y_i - \bar{Y} = \beta (X_i - \bar{X}) + (u_i - \bar{u})$$

$$y_i = \beta x_i + (u_i - \bar{u})$$

así al sustituir en e_i se logra

$$\begin{aligned} e_i &= \beta x_i + (u_i - \bar{u}) - \hat{\beta} x_i \\ &= -(\hat{\beta} - \beta) x_i + (u_i - \bar{u}) \end{aligned}$$

de donde la suma de los cuadrados de los residuales es

$$\sum_{i=1}^{n} e_i^2 = (\hat{\beta} - \beta)^2 \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} (u_i - \bar{u})^2 - 2(\hat{\beta} - \beta) \sum_{i=1}^{n} x_i (u_i - \bar{u})$$

Si tomamos la esperanza de la suma de cuadrados de residuales

$$E(\sum_{i=1}^{n} e_i^2) = \sum_{i=1}^{n} x_i^2 E\left(\hat{\beta} - \beta\right)^2 + E\left(\sum_{i=1}^{n}\left(u_i - \bar{u}\right)^2\right) - 2E\left(\left(\hat{\beta} - \beta\right)\sum_{i=1}^{n} x_i\left(u_i - \bar{u}\right)\right) \qquad (2.3.22)$$

Para mayor claridad se procede sumando a sumando con (2.3.22)

$$\sum_{i=1}^{n} x_i^2 E\left(\hat{\beta} - \beta\right)^2 = \sum_{i=1}^{n} x_i^2 V\left(\hat{\beta}\right)$$

$$= \sum_{i=1}^{n} x_i^2 \frac{\sigma_u^2}{\sum_{i=1}^{n} x_i^2} \qquad \text{por (2.1.17)}$$

$$= \sigma_u^2 \qquad (2.3.23)$$

$$E\left(\sum_{i=1}^{n}\left(u_i - \bar{u}\right)^2\right) = E\left(\sum_{i=1}^{n}\left(u_i^2 - 2u_i\bar{u} + \bar{u}^2\right)\right)$$

$$= \left(\sum_{i=1}^{n} u_i^2 - \frac{\left(\sum_{i=1}^{n} u_i\right)^2}{n}\right)$$

$$= E\left(\sum_{i=1}^{n} u_i^2 - \frac{\left(\sum_{i=1}^{n} u_i^2 + 2\sum_{i<j}\sum u_i u_j\right)}{\mathrm{n}}\right)$$

$$= n\sigma_u^2 - \sigma_u^2$$

$$= (n-1)\sigma_u^2 \qquad (2.3.24)$$

Ahora al utilizar la relación ya familiar

$$\hat{\beta} - \beta = \sum_{i=1}^{n} w_i u_i$$

$$= \frac{\sum_{i=1}^{n} x_i u_i}{\sum_{i=1}^{n} x_k^2}$$

Se aplica el cociente anterior al sumando restante

$$E\left(\left(\hat{\beta} - \beta\right)\sum_{i=1}^{n} x_i\left(u_i - \bar{u}\right)\right) = E\,\frac{\sum_{i=1}^{n} x_i u_i}{\sum_{k=1}^{n} x_k^2}\left(\sum_{i=1}^{n} x_i u_i - \bar{u}\sum_{i=1}^{n} x_i\right)$$

$$= E\frac{\left(\sum_{i=1}^{n} x_i u_i\right)^2}{\sum_{k=1}^{n} x_k^2}$$

$$= E\frac{\sum_{i=1}^{n} x_i^2 u_i^2 + \sum\sum_{i<j} x_i x_j u_i u_j}{\sum_{k=1}^{n} x_k^2}$$

$$= \sigma_u^2 \qquad (2.3.25)$$

Finalmente al sustituir (2.3.23), (2.3.24) y (2.3.25) en (2.3.22)

$$E\left(\sum_{i=1}^{n} e_i^2\right) = \sigma_u^2 + (n-1)\sigma_u^2 - 2\sigma_u^2$$

$$= (n-2)\sigma_u^2$$

De donde se concluye que un estimador insesgado de σ_u^2 es el siguiente

$$\hat{\sigma}_u^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} \qquad \hat{\sigma}_u = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}} \qquad (2.3.26)$$

Ejemplo:

Se tienen 15 parejas de observaciones (X,Y) y se procederá a calcular las columnas de desviaciones respecto a la media y columnas auxiliares para obtener estimaciones puntuales de los parámetros.

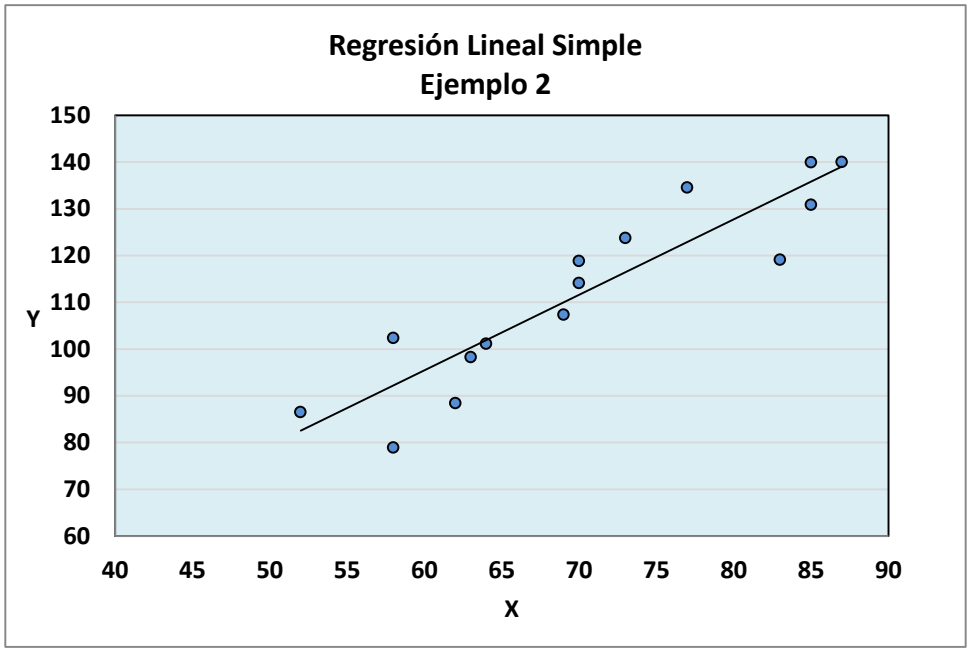
Obs.	X_i	Y_i	x_i	y_i	$x_i y_i$	x_i^2	y_i^2	$\hat{Y}_i$	e_i	e_i^2
1	52.0	86.5	-18.4	-25.74	473.616	338.560	662.548	82.511	3.989	15.912
2	58.0	78.9	-12.4	-33.34	413.416	153.760	1111.556	92.205	-13.305	177.030
3	58.0	102.3	-12.4	-9.94	123.256	153.760	98.804	92.205	10.095	101.904
4	62.0	88.4	-8.4	-23.84	200.256	70.560	568.346	98.668	-10.268	105.433
5	63.0	98.2	-7.4	-14.04	103.896	54.760	197.122	100.284	-2.084	4.342
6	64.0	101.1	-6.4	-11.14	71.296	40.960	124.100	101.899	-0.799	0.639
7	69.0	107.3	-1.4	-4.94	6.916	1.960	24.404	109.978	-2.678	7.172
8	70.0	118.8	-0.4	6.56	-2.624	0.160	43.034	111.594	7.206	51.930
9	70.0	114.1	-0.4	1.86	-0.744	0.160	3.460	111.594	2.506	6.281
10	73.0	123.7	2.6	11.46	29.796	6.760	131.332	116.441	7.259	52.696
11	77.0	134.5	6.6	22.26	146.916	43.560	495.508	122.904	11.596	134.475
12	83.0	119.1	12.6	6.86	86.436	158.760	47.060	132.598	-13.498	182.193
13	85.0	130.8	14.6	18.56	270.976	213.160	344.474	135.829	-5.029	25.294
14	85.0	139.9	14.6	27.66	403.836	213.160	765.076	135.829	4.071	16.571
15	87.0	140.0	16.6	27.76	460.816	275.560	770.618	139.061	0.939	0.882
Suma	**1,056.00**	**1,683.60**	**0.00**	**0.00**	**2,788.06**	**1,725.60**	**5,387.44**	**1,683.60**	**0.00**	**882.75**
Media	**70.40**	**112.24**								

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^1} = \frac{2{,}788.06}{1{,}725.80} = 1.615705$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = -1.50561$$

$$\hat{\sigma}^2 = \frac{\sum e^2}{n-2} = \frac{882.75}{15-2} = 67.9042$$

La siguiente gráfica de dispersión muestra las parejas de datos representados por puntos, junto con la recta ajustada.



Si se utilizan las fórmulas derivadas de las ecuaciones normales, se llega al mismo resultados, la ventaja de utilizar desviaciones es la simplificación de las fórmulas para los aspectos inferencales.

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} Y_i \sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} X_i Y_i =}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} = 2.05819$$

$$\hat{\beta} = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} = 0.86741$$