Gauss propuso que se considerase $\varepsilon$ una variable aleatoria. Por otra parte, se ha considerado asumir que

$$\mathbb{E}(\varepsilon)=0 \qquad VAR(\varepsilon)=\sigma^2 \quad \text{y que}$$

$$(RL) \cdots \quad Y_i=\beta_0+\beta_1 X_i+\varepsilon_i \quad ; \; i=1,2,\ldots,N.$$

Si se asume que la variable explicativa X tiene valores que son fijados en forma determinísta por quienes estudien el experimento, entonces podemos considerar que $\beta_0+\beta_1 X_i$ no es un término aleatorio en el modelo y por tanto el carácter aleatorio de $\varepsilon_i$ se hereda de forma directa a $Y_i$, de forma que

$$\begin{aligned}\mathbb{E}(Y_i)&=\mathbb{E}(\beta_0+\beta_1 X_i+\varepsilon_i)\\&=\mathbb{E}(\beta_0+\beta_1 X_i)+\mathbb{E}(\varepsilon_i)\\&=\beta_0+\beta_1 X_i+0=\beta_0+\beta_1 X_i\,;\;\forall i=1,\ldots,N.\end{aligned}$$

También

$$\begin{aligned}VAR(Y_i)&=VAR(\beta_0+\beta_1 X_i+\varepsilon_i)\\&=VAR(\varepsilon_i)=\sigma^2\,;\;\forall i=1,\ldots,N\end{aligned}$$

Para simplificar el modelo, también se considera que el error en la medición $Y_i$ no tiene que ver, en un sentido estocástico, con el error para la medición $Y_j$, para cada $i \neq j$. Aunque la correlación cero <u>no</u> implica independencia de variables aleatorias, por el momento asumiremos que $\text{cov}(\varepsilon_i, \varepsilon_j) = 0 \ \forall \ i \neq j$.

Como consecuencia de este supuesto

$$\text{COV}(Y_i, Y_j) = \text{cov}(\beta_0 + \beta_1 X_i + \varepsilon_i, \beta_0 + \beta_1 X_j + \varepsilon_j)$$

$$= \text{cov}(\beta_0 + \beta_1 X_i, \beta_0 + \beta_1 X_j + \varepsilon_j) + \text{cov}(\varepsilon_i, \beta_0 + \beta_1 X_j + \varepsilon_j)$$

$$= \text{cov}(\varepsilon_i, \beta_0 + \beta_1 X_j + \varepsilon_j)$$

$$= \text{cov}(\varepsilon_i, \varepsilon_j) + \text{cov}(\varepsilon_i, \beta_0 + \beta_1 X_j)$$

$$= \text{cov}(\varepsilon_i, \varepsilon_j) = 0 \quad ; \ \forall i \neq j$$

Usaremos estas propiedades para evaluar la calidad de los coeficientes ajustados por mínimos cuadrados $\hat{\beta}_0$ y $\hat{\beta}_1$. Recordemos que

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \quad \text{y} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

Como ya mencionamos antes, los valores de la variable explicativa se consideran como constantes conocidas, de forma que $\sum_{i=1}^{N}(x_i-\bar{x})^2$ tambien es una constante conocida para nosotros. Por otra parte $\sum_{i=1}^{N}(y_i-\bar{y})(x_i-\bar{x})$ involucra valores de la variable de respuesta, dado el modelo (RL), estos valores son realizaciones de las variables aleatorias $Y_1, \dots, Y_N$ y por ende, $\hat{\beta}_1$ también se puede pensar como una variable aleatoria. A continuación analizaremos a $\hat{\beta}_1$ con más detalle, observemos que

$$
\begin{aligned}
\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y}) &= \sum_{i=1}^{N}(x_i-\bar{x})y_i - \sum_{i=1}^{N}(x_i-\bar{x})\bar{y} \\
&= \sum_{i=1}^{N}(x_i-\bar{x})y_i - \bar{y}\sum_{i=1}^{N}(x_i-\bar{x}) \\
&= \sum_{i=1}^{N}(x_i-\bar{x})y_i - \bar{y}\cdot 0 \\
&= \sum_{i=1}^{N}(x_i-\bar{x})y_i .
\end{aligned}
$$

De forma que $\hat{\beta}_1 = \sum_{i=1}^{N} a_i Y_i$, donde $a_i = \dfrac{X_i-\bar{X}}{\sum_{i=1}^{N}(X_i-\bar{X})^2}$; $i=1,2,\dots,N$. Lo anterior nos dice que $\hat{\beta}_1$ es una función lineal de las variables aleatorias $Y_1,\dots,Y_N$, como consecuencia de esto se tiene que

$$\mathbb{E}(\hat{\beta}_1) = \mathbb{E}\left(\sum_{i=1}^{N} a_i Y_i\right) = \sum_{i=1}^{N} \mathbb{E}(a_i Y_i)$$

$$= \sum_{i=1}^{N} a_i \, \mathbb{E}(Y_i) = \sum_{i=1}^{N} a_i(\beta_0 + \beta_1 X_i)$$

$$= \sum_{i=1}^{N} a_i \beta_0 + \sum_{i=1}^{N} a_i \beta_1 X_i$$

$$= \beta_0 \sum_{i=1}^{N} a_i + \beta_1 \sum_{i=1}^{N} a_i X_i .$$

Pero $\sum_{i=1}^{N} a_i = \frac{1}{\sum_{i=1}^{N}(X_i-\bar{X})^2} \cdot \sum_{i=1}^{N}(X_i - \bar{X}) = 0$ y además

$$\sum_{i=1}^{N} a_i X_i = \frac{1}{\sum_{i=1}^{N}(X_i-\bar{X})^2} \sum_{i=1}^{N}(X_i-\bar{X}) X_i$$

$$= \frac{1}{\sum_{i=1}^{N}(X_i-\bar{X})^2} \left\{ \sum_{i=1}^{N}(X_i-\bar{X})(X_i-\bar{X}) + \sum_{i=1}^{N}(X_i-\bar{X})\bar{X} \right\}$$

$$= \frac{1}{\sum_{i=1}^{N}(X_i-\bar{X})^2} \cdot \sum_{i=1}^{N}(X_i-\bar{X})^2 = 1$$

$$\therefore \quad \mathbb{E}(\hat{\beta}_1) = \beta_1$$

"$\hat{\beta}_1$ es un estimador insesgado y lineal de $\beta_1$"

Por otra parte, debido a que $\mathrm{COV}(Y_i, Y_j) = 0 \;\; \forall \, i \neq j$

$$\mathrm{VAR}(\hat{\beta}_1) = \mathrm{VAR}\left(\sum_{i=1}^{N} a_i Y_i\right) = \sum_{i=1}^{N} a_i^2 \, \mathrm{VAR}(Y_i)$$

$$VAR(\hat{\beta}_1) = \sum_{i=1}^{N} \frac{(x_i-\overline{x})^2}{\left(\sum_{i=1}^{N}(x_i-\overline{x})^2\right)^2} \sigma^2$$

$$= \sigma^2 \frac{1}{\left(\sum_{i=1}^{N}(x_i-\overline{x})^2\right)^2} \sum_{i=1}^{N}(x_i-\overline{x})^2$$

$$= \frac{\sigma^2}{\sum_{i=1}^{N}(x_i-\overline{x})^2}$$

De este cálculo, observamos que: La precisición de la estimación aumenta en la medida en que se incrementa el valor de $\sum_{i=1}^{N}(x_i-\overline{x})^2$ (1). Entonces, cuando es posible elegir los valores de la variable explicativa, en el diseño del experimento, resulta conveniente que estos valores sean tales que su varianza sea lo más grande posible.

El caso de $\hat{\beta}_0$ es similar

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X} = \overline{Y} - \left(\sum_{i=1}^{N} a_i Y_i\right)\overline{X}$$

$$= \sum_{i=1}^{N}\frac{1}{N}Y_i - \left(\sum_{i=1}^{N} a_i Y_i\right)\overline{X} = \sum_{i=1}^{N}\left(\frac{1}{N} - a_i\overline{X}\right)Y_i$$

(1) Si $\sum_{i=1}^{N}(x_i-\overline{x})^2$ crece, entonces $VAR(\hat{\beta}_1) = \mathbb{E}\left((\hat{\beta}_1-\beta_1)^2\right)$ decrece y en promedio la "distancia" $(\hat{\beta}_1-\beta_1)^2$ es más pequeña.

$$\hat{\beta}_0 = \sum_{i=1}^{N} b_i Y_i \,, \quad \text{donde} \quad b_i = \frac{1}{N} - a_i \bar{X}$$

Por tanto, $\hat{\beta}_0$ es también, un estimador lineal de $\beta_0$.

Como variable aleatoria tenemos que $\hat{\beta}_0$ satisface

$$\begin{aligned}
\mathbb{E}(\hat{\beta}_0) &= \sum_{i=1}^{N} b_i \mathbb{E}(Y_i) \\
&= \sum_{i=1}^{N} b_i (\beta_0 + \beta_1 X_i) \\
&= \sum_{i=1}^{N} b_i \beta_0 + \sum_{i=1}^{N} b_i \beta_1 X_i \\
&= \beta_0 \sum_{i=1}^{N} b_i + \beta_1 \sum_{i=1}^{N} b_i X_i \\
&= \beta_0 \sum_{i=1}^{N} \left(\frac{1}{N} - a_i \bar{X}\right) + \beta_1 \sum_{i=1}^{N} \left(\frac{1}{N} - a_i \bar{X}\right) X_i .
\end{aligned}$$

Pero como $\sum_{i=1}^{N} a_i X_i = 1$ entonces

$$\sum_{i=1}^{N} \left(\frac{1}{N} - a_i \bar{X}\right) X_i = \sum_{i=1}^{N} \frac{1}{N} X_i - \bar{X} \sum_{i=1}^{N} a_i X_i = \bar{X} - \bar{X} = 0 .$$

Por otra parte $\sum_{i=1}^{N} a_i = 0$ implica que

$$\sum_{i=1}^{N} \left(\frac{1}{N} - a_i \bar{X}\right) = 1 - \bar{X} \sum_{i=1}^{N} a_i = 1 .$$ Por tanto

$$\mathbb{E}(\hat{\beta}_0) = \beta_0 \times 1 + \beta_1 \times 0 = \beta_0 ,$$

de manera que $\hat{\beta}_0$ es un estimador lineal e insesgado de $\beta_0$. Por otra parte (debido a que $COV(Y_i, Y_j) = 0 \quad \forall i \neq j$)

$$VAR(\hat{\beta}_0) = \sum_{i=1}^{N} b_i^2 VAR(Y_i)$$

$$= \sum_{i=1}^{N} b_i^2 \sigma^2$$

$$= \sigma^2 \sum_{i=1}^{N} b_i^2$$

Pero $\sum_{i=1}^{N} b_i^2 = \sum_{i=1}^{N} \left(\frac{1}{N} - a_i\bar{X}\right)^2 = \sum_{i=1}^{N} \left(\frac{1}{N^2} - \frac{2a_i\bar{X}}{N} + a_i^2\bar{X}^2\right)$

$$= \frac{1}{N} + \bar{X}^2 \sum_{i=1}^{N} a_i^2 = \frac{1}{N} + \bar{X}^2 \sum_{i=1}^{N} \left\{ \frac{(X_i - \bar{X})^2}{\left(\sum_{i=1}^{N}(X_i-\bar{X})^2\right)^2} \right\}$$

$$= \frac{1}{N} + \bar{X}^2 \frac{1}{\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

$$= \frac{\frac{1}{N}\sum_{i=1}^{N}(X_i-\bar{X})^2 + \bar{X}^2}{\sum_{i=1}^{N}(X_i-\bar{X})^2} = \frac{\frac{1}{N}\left(\sum_{i=1}^{N} X_i^2 - N\bar{X}^2\right) + \bar{X}^2}{\sum_{i=1}^{N}(X_i-\bar{X})^2}$$

$$= \frac{\frac{1}{N}\sum_{i=1}^{N} X_i^2 - \bar{X}^2 + \bar{X}^2}{\sum_{i=1}^{N}(X_i-\bar{X})^2} = \frac{1}{N} \frac{\left(\sum_{i=1}^{N} X_i^2\right)}{\sum_{i=1}^{N}(X_i-\bar{X})^2}$$

$$\therefore \; VAR(\hat{\beta}_0) = \sigma^2\left(\sum_{i=1}^{N} X_i^2\right) \Big/ N\left(\sum_{i=1}^{N}(X_i-\bar{X})^2\right)$$

Esta varianza es más pequeña, a medida que los valores de la variable explicativa produzcan un