Supongamos que cierta variable respuesta Y depende linealmente de dos variables regresoras x_1 y x_2, de manera que se satisface el modelo:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \quad i = 1, \ldots, n,$$

donde los errores ϵ_i cumplen las hipótesis habituales. Se ajusta por mínimos cuadrados el modelo $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1}$, sin tener en cuenta la segunda variable regresora. Demuestra que el estimador $\hat{\beta}_1$ es, en general, sesgado y determina bajo qué condiciones se anula el sesgo.

Solución:

$$\hat{\beta}_1 = \frac{S_{x_1 y}}{S_{x_1 x_1}} = \frac{1}{S_{x_1 x_1}}(x_{11} - \bar{x}_1, \ldots, x_{n1} - \bar{x}_1) \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \mathbf{B}\mathbf{y},$$

donde $\mathbf{y} = (y_1, \ldots, y_n)'$,

$$\mathbf{B} = \frac{1}{S_{x_1 x_1}}(\mathbf{x}_1 - \bar{x}_1 \mathbf{1}_n)' \qquad \text{y} \qquad \mathbf{x}_1 = \begin{pmatrix} x_{11} \\ \vdots \\ x_{n1} \end{pmatrix}.$$

Por tanto, si denotamos

$$\mathbb{X} = \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{11} & x_{12} \\ \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} \end{pmatrix} \qquad \text{y} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

tenemos que

$$\mathbb{E}(\hat{\beta}_1) = \mathbf{B}\mathbb{E}(\mathbf{y}) = \mathbf{B}\mathbb{X}\boldsymbol{\beta} = \left(0, 1, \frac{S_{x_1 x_2}}{S_{x_1 x_1}}\right)\boldsymbol{\beta} = \beta_1 + \beta_2 \frac{S_{x_1 x_2}}{S_{x_1 x_1}}. \tag{1}$$

Observemos que $S_{x_1 x_2}/S_{x_1 x_1}$ es el coeficiente de regresión lineal de x_2 sobre x_1. Por (1), $\hat{\beta}_1$ es insesgado si y sólo si $\beta_2 = 0$ o si $S_{x_1 x_2} = 0$ (es decir, x_1 y x_2 son incorreladas).