

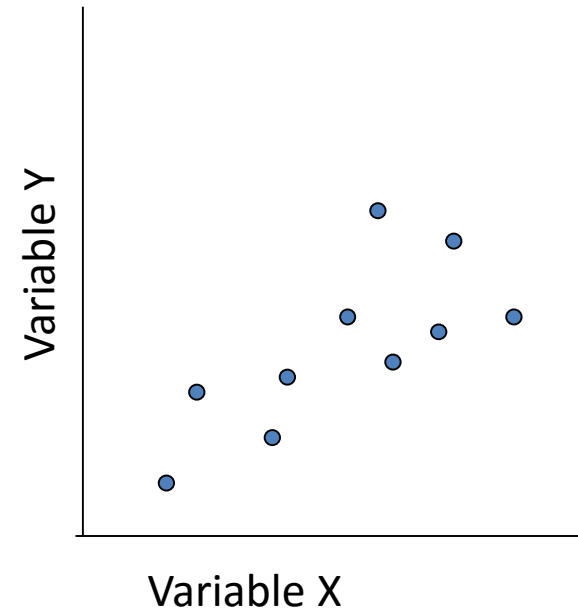
Modelo de Regresión Lineal Simple

GRÁFICOS DE DISPERSIÓN: Permite representar la evolución conjunta de ambas variables

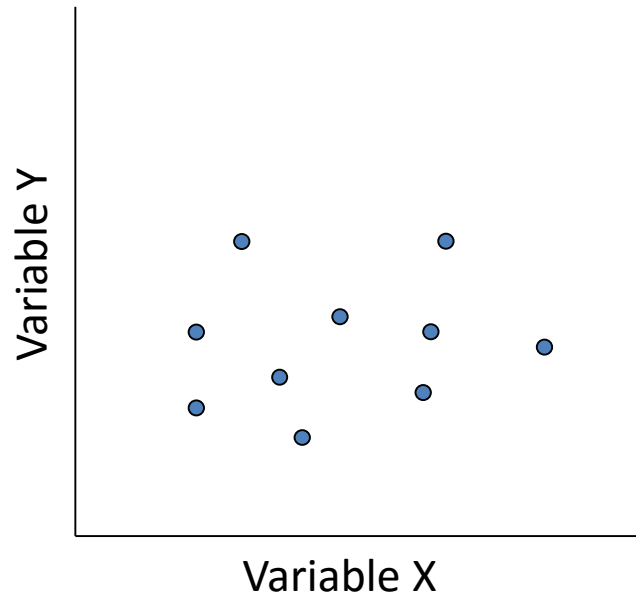
Dadas dos variables (Y , X) tomadas sobre el mismo elemento de la población, el diagrama de dispersión es simplemente un gráfico de dos dimensiones, donde en un eje (la abscisa) se sitúa una variable, y en el otro eje (la ordenada) se sitúa la otra variable

Si las variables están correlacionadas, el gráfico mostraría algún nivel de correlación (tendencia) entre las dos variables. Si no hay ninguna correlación, el gráfico presentaría una figura sin forma, una nube de puntos dispersos en el gráfico.

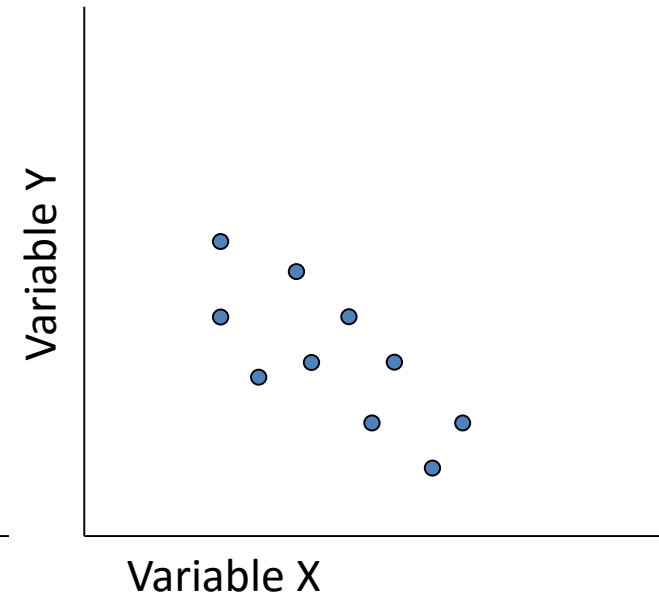
Representación gráfica de una relación



Relación lineal positiva



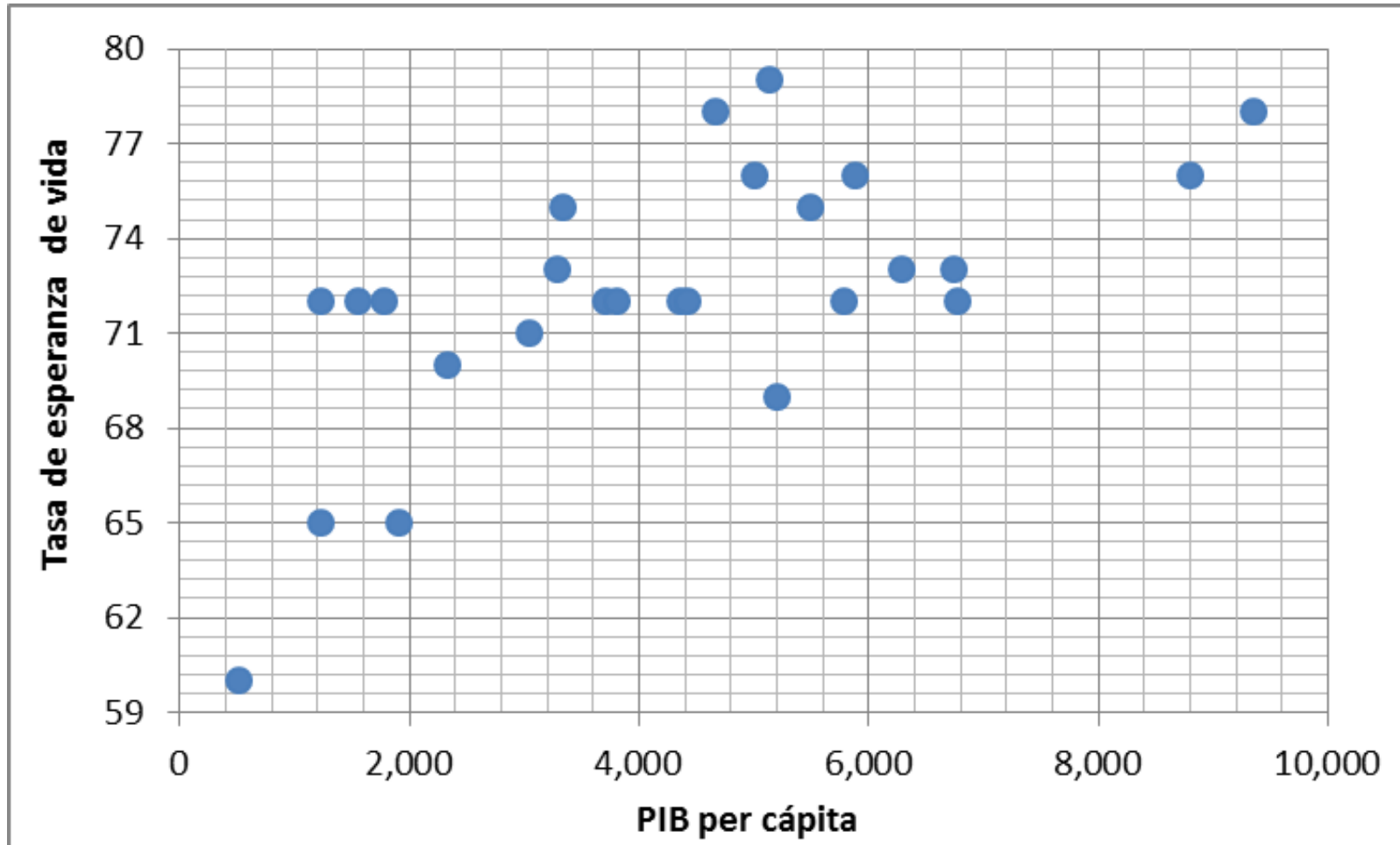
Sin relación



Relación lineal negativa

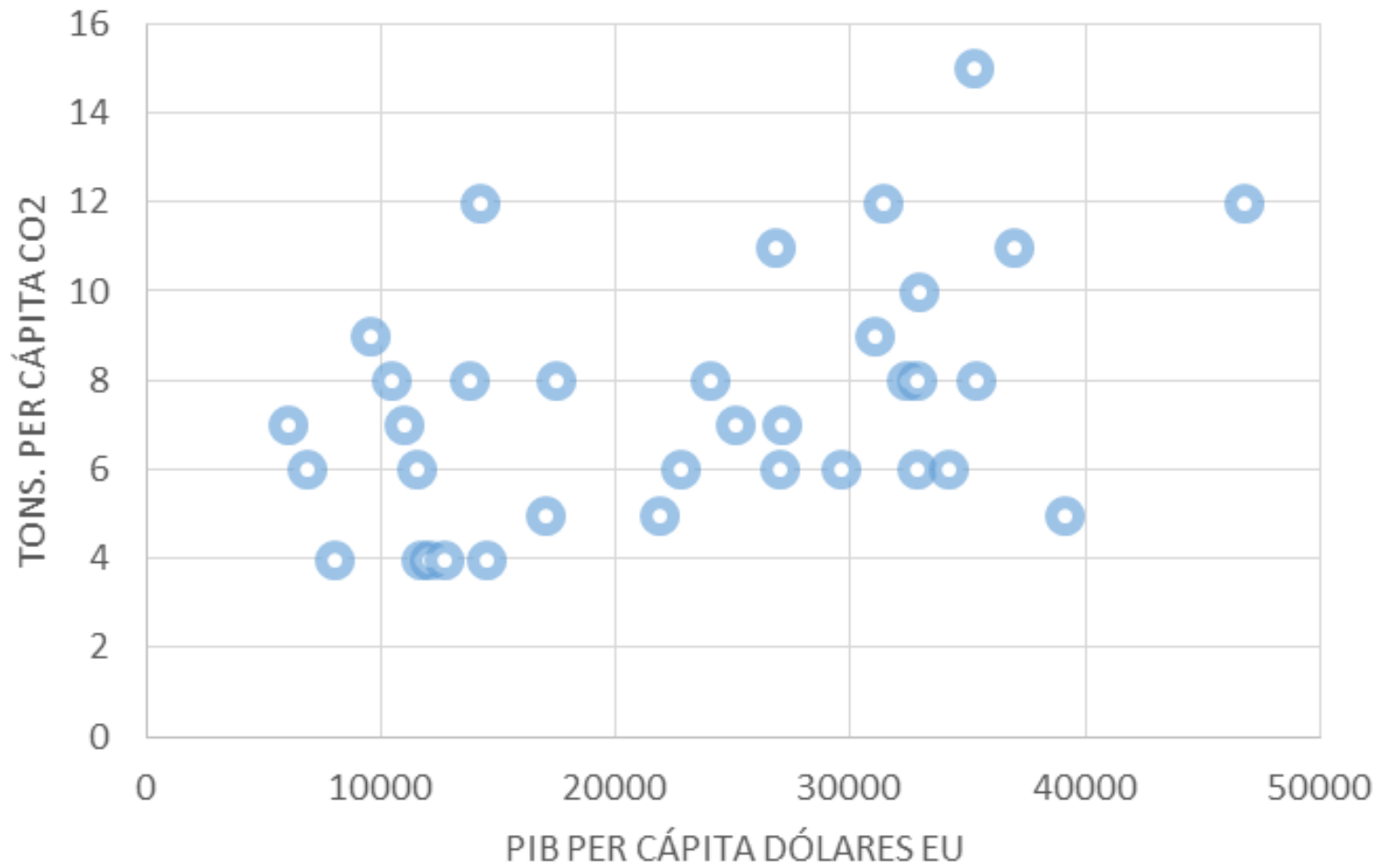
Diagrama de dispersión

Tasa de esperanza de vida vs PIB per cápita



26 países de América Latina

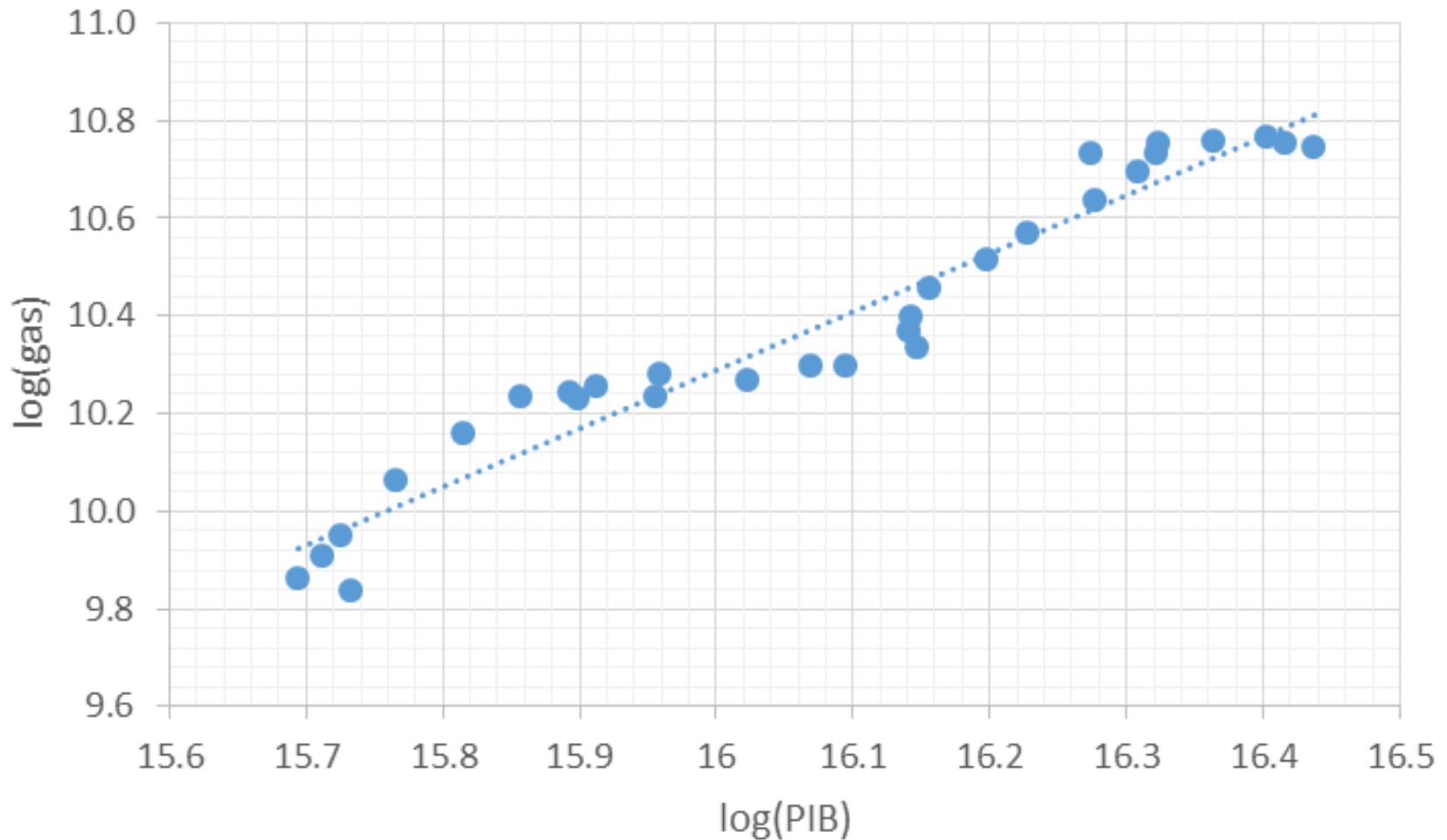
CO2 percapita vs PIB percapita



35 países

Diagrama de dispersión

Consumo de gasolina vs PIB



Variables en logaritmo natural

Una primera aproximación, al Modelo Estadístico General es en el marco de un modelo de regresión entre dos variables es:

$$Y_i = \alpha + \beta X_i \quad i = 1, 2, \dots, n$$

Donde:

Y se denomina como la variable dependiente

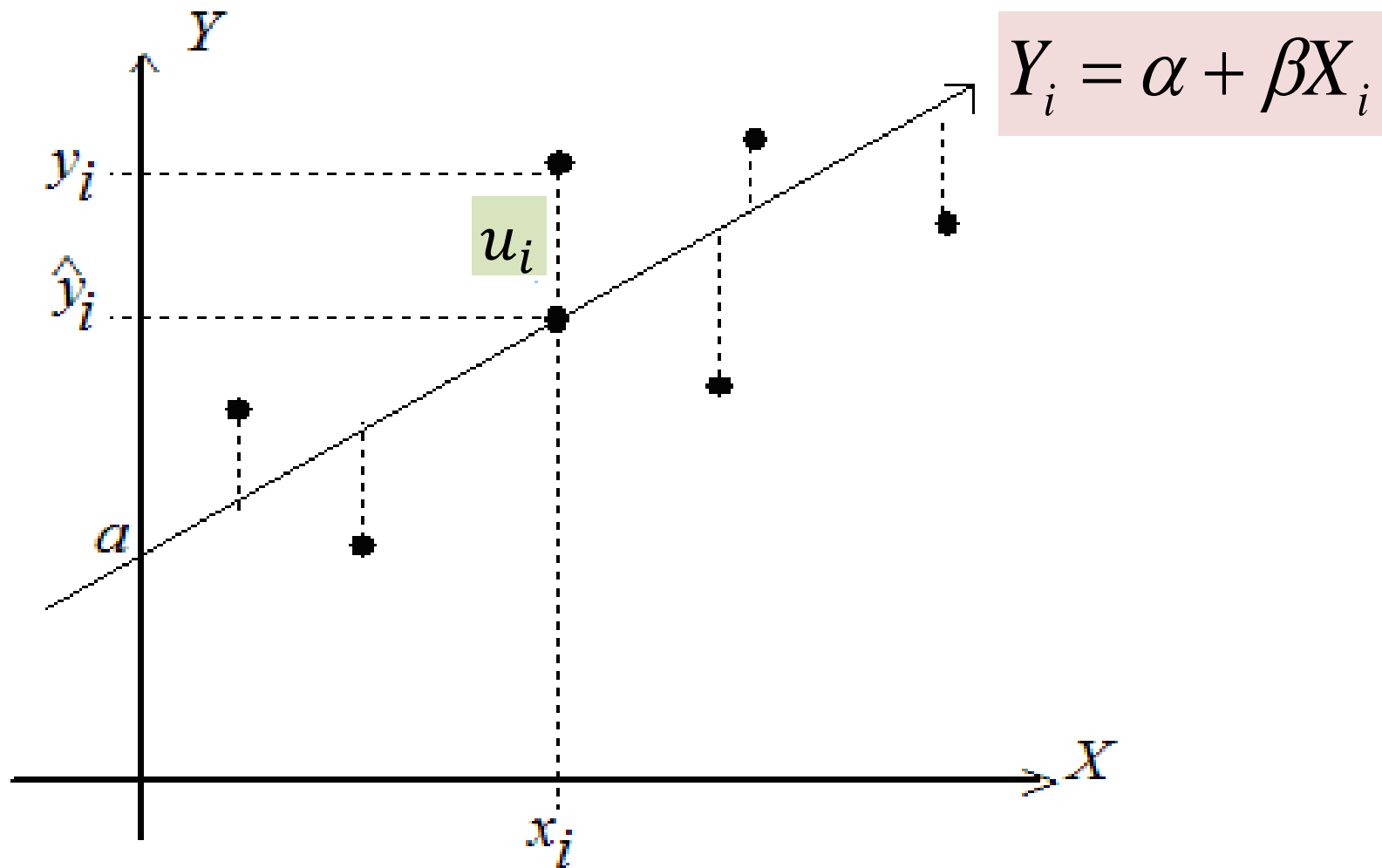
X es la variable explicativa o independiente

α y β los parámetros del modelo a estima

El modelo establece que un cambio en una unidad de X produce u ocasiona un cambio en la variable Y, medido por el parámetro β

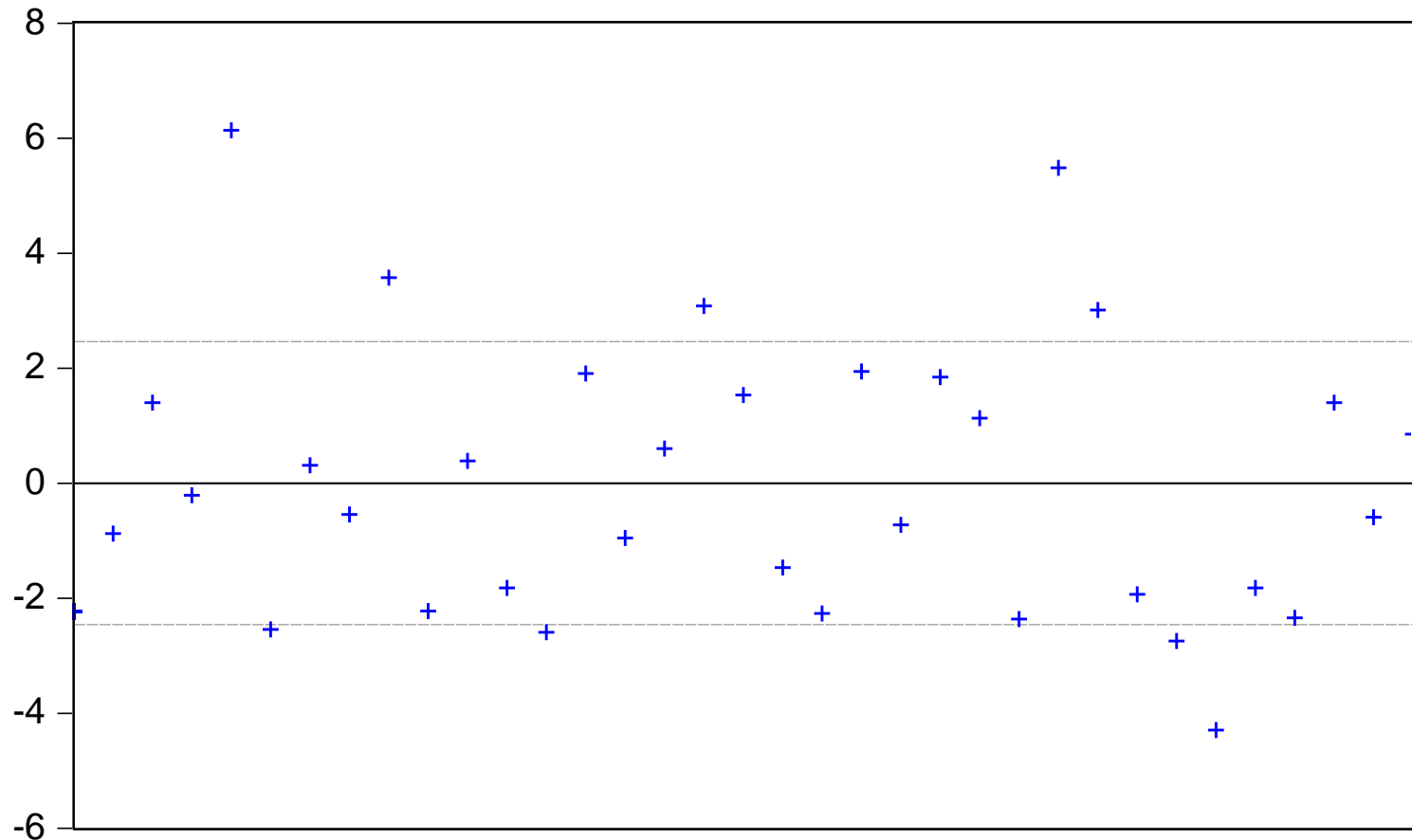
Sin embargo, la media poblacional de Y es una característica desconocida de la distribución, por lo tanto la pendiente es desconocida. Así el supuesto básico del modelo de regresión es que una realización muestral de la variable Y puede ser expresada como una combinación lineal de las observaciones de X incluyendo un el componente denominado término de error:

$$Y_i = \alpha + \beta X_i + u_i \quad i = 1, 2, \dots, n$$

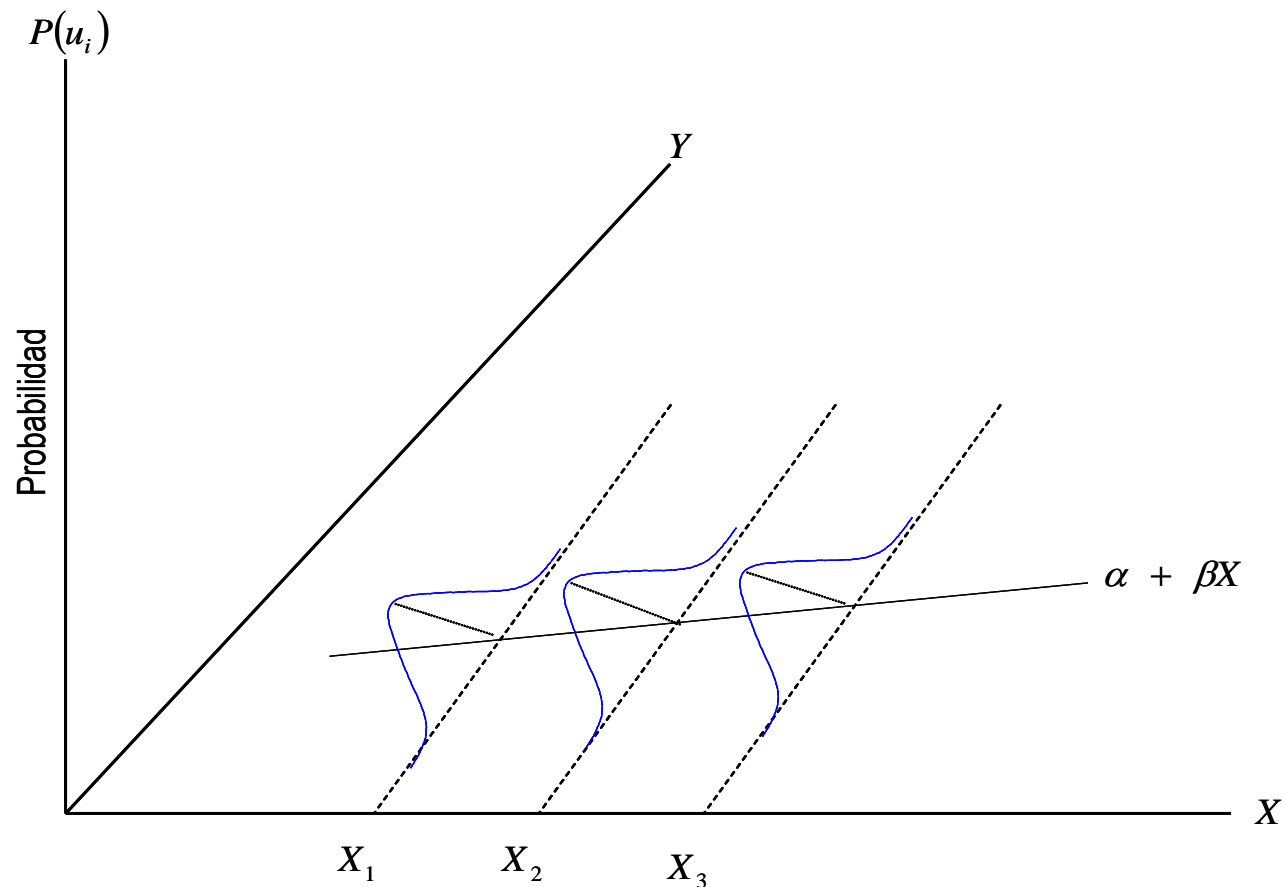


La diferencia entre el valor observado de la variable y la recta estimada se denomina error

Las desviaciones en cada observación generan una serie de errores, la cual se considera como una variable aleatoria alrededor del cero, con valores positivos y negativos



Para cada valor de X existe una distribución de probabilidad del término de error y por consiguiente una distribución de probabilidad para la variable Y



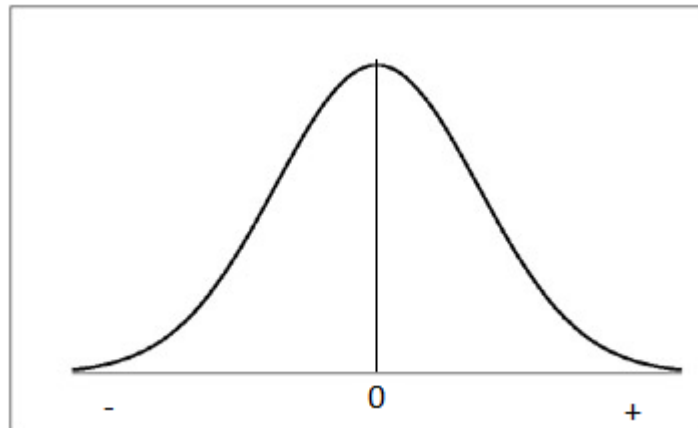
Asumiendo que el término de error presenta una distribución de probabilidad se realizan entonces ciertos supuestos sobre dicha distribución

1. El valor esperado del término de error es igual a cero. $E(u_i) = 0$, para todo $i = 1, \dots, N$. El término aleatorio tiene esperanza igual a cero para todas las observaciones. Este supuesto implica que en promedio la relación entre Y y variable X es exactamente lineal, aunque las realizaciones particulares de los u_i 's pueden ser distintas de cero.

2. Homocedasticidad o varianza constante.

$Var(u_i) = \sigma^2; i = 1, \dots, N$. La varianza del término aleatorio es constante para todas las observaciones. Esto se conoce como supuesto de homoscedasticidad

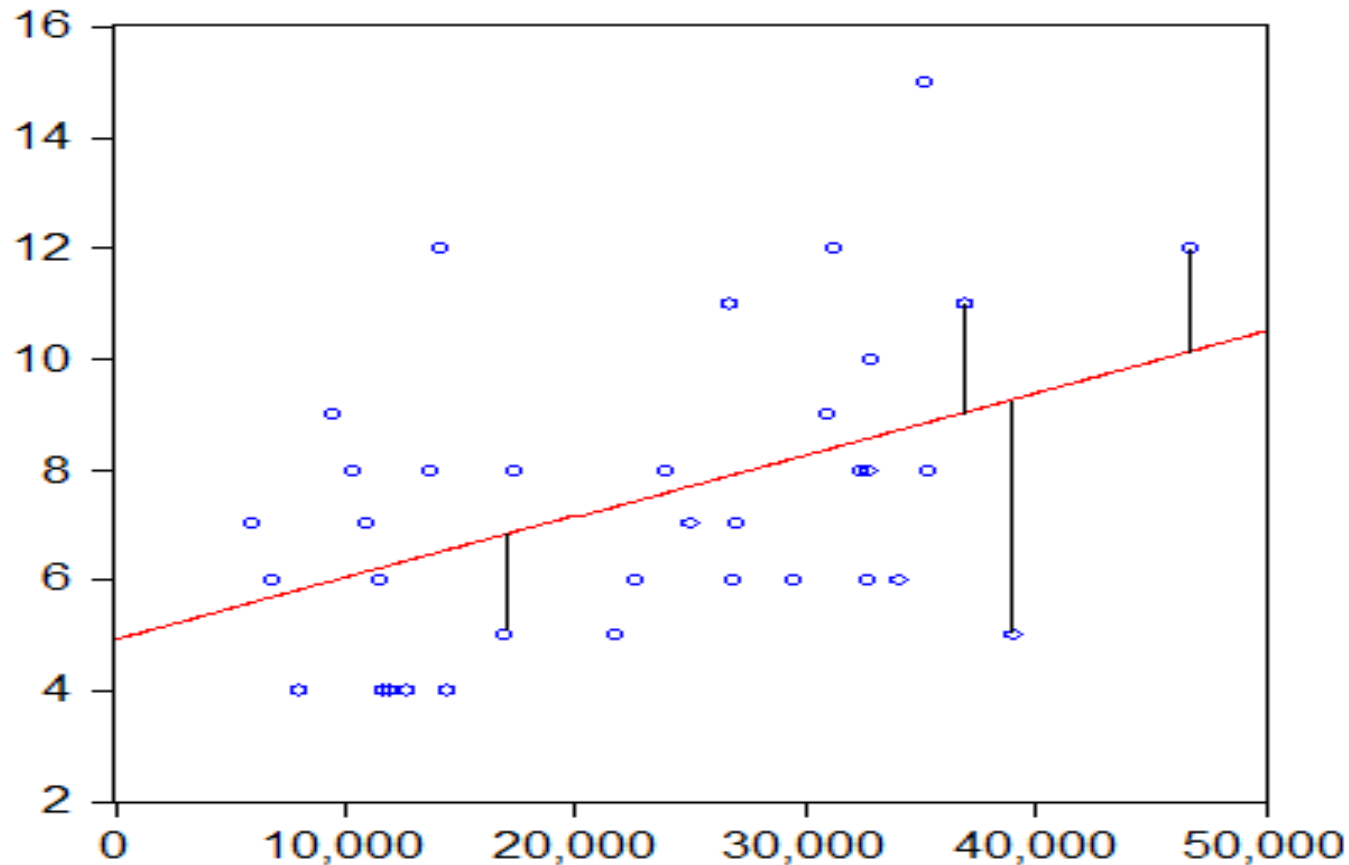
3. Normalidad. Los errores se distribuyen como una función de densidad de probabilidad normal, con media cero y varianza constante




4. u_i , es independiente de u_j , para todo $i \neq j$.

Método de estimación de Mínimos Cuadrados Ordinarios (MCO)

Se puede observar que la mayoría de los puntos no pasan por la línea recta, lo cual se identifica como el término de error





La forma de obtener una estimación de los parámetros del modelo es por medio del MÉTODO DE MÍNIMOS CUADRADOS ORDINARIOS (MCO)

Modelo lineal

$$(1) Y_i = \alpha + \beta X_i + u_i$$

Se despeja el término de error

$$(2) u_i = Y_i - \alpha - \beta X_i$$

La distancia esta definida como el término de error al cuadrado, y la distancia total como la suma de los errores al cuadrado (SEC)

$$(3) \sum_{i=1}^N u_i^2 = \sum_{i=1}^N (Y_i - \alpha - \beta X_i)^2$$

La ecuación (3) representa la función objetivo a minimizar, dado que la función depende de los parámetros poblacionales, que son desconocidos.

Así, dada una muestra de la variable dependiente y, el método de mínimos cuadrados ordinarios (MCO), considera los valores muestrales de los parámetros, los cuales se definen como los estimadores de MCO

$$(4) \quad S(\hat{u}) = \sum_{i=1}^N (Y_i - \hat{\alpha} - \hat{\beta}X_i)^2$$

Se resuelve como un problema de optimización, mediante un sistema de dos ecuaciones y dos incógnitas. Donde las incógnitas son los valores de α y β , y dada la información muestra se obtienen los valores estimados

$$\frac{\partial S(\hat{u})}{\partial \hat{\alpha}} = 2 \sum_{i=1}^N (Y_i - \hat{\alpha} - \hat{\beta} X_i)(-1) = 0$$

$$\frac{\partial S(\hat{u})}{\partial \hat{\beta}} = 2 \sum_{i=1}^N (Y_i - \hat{\alpha} - \hat{\beta} X_i)(-X_i) = 0$$

$$\frac{\partial S(\hat{u})}{\partial \hat{\alpha}} = -2 \sum_{i=1}^N Y_i + 2 \sum_{i=1}^N \hat{\alpha} + 2 \sum_{i=1}^N \hat{\beta} X_i = 0$$

$$\frac{\partial S(\hat{u})}{\partial \hat{\beta}} = -2 \sum_{i=1}^N Y_i X_i + 2 \sum_{i=1}^N \hat{\alpha} X_i + 2 \sum_{i=1}^N \hat{\beta} X_i^2 = 0$$

Es un sistema de dos ecuaciones con dos incógnitas. Despejando $\hat{\alpha}$ de la primera ecuación, se obtiene el siguiente resultado

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

Sustituyendo $\hat{\alpha}$ en la segunda ecuación

$$-2 \sum_{i=1}^N Y_i X_i + 2 \sum_{i=1}^N (\bar{Y} - \hat{\beta} \bar{X}) X_i + 2 \sum_{i=1}^N \hat{\beta} X_i^2 = 0$$

$$- \sum_{i=1}^N Y_i X_i + \sum_{i=1}^N (\bar{Y} - \hat{\beta} \bar{X}) X_i + \sum_{i=1}^N \hat{\beta} X_i^2 = 0$$

Reordenando la ecuación

$$- \sum_{i=1}^N Y_i X_i + \sum_{i=1}^N \bar{Y} X_i - \hat{\beta} \sum_{i=1}^N \bar{X} X_i + \sum_{i=1}^N \hat{\beta} X_i^2 = 0$$

Agrupando términos semejantes:

$$-\sum_{i=1}^N Y_i X_i + \bar{Y} \sum_{i=1}^N X_i + \hat{\beta} \left(\sum_{i=1}^N X_i^2 - \bar{X} \sum_{i=1}^N X_i \right) = 0$$

El estimador $\hat{\beta}$ se puede calcular como:

$$\hat{\beta} = \frac{\sum_{i=1}^N Y_i X_i - \bar{Y} \sum_{i=1}^N X_i}{\left(\sum_{i=1}^N X_i^2 - \bar{X} \sum_{i=1}^N X_i \right)} = \frac{\sum_{i=1}^N (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^N (X_i - \bar{X})^2} = \frac{Cov(YX)}{Var(X)}$$

El modelo de regresión lineal establece que Y_i puede ser aproximada por una función lineal de X_i ,

Los valores de $\hat{\alpha}$ y $\hat{\beta}$ permiten “estimar” los valores de la ordenada al origen y la pendiente