- We typically assume the random errors balance out – they average zero.
- Then this is equivalent to assuming the mean of Y, denoted $E(Y)$, equals the deterministic component.

Straight-Line Regression Model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Y = response variable (dependent variable)
X = predictor variable (independent variable)
ε = random error component

β_0 = Y-intercept of regression line
β_1 = slope of regression line

Note that the deterministic component of this model is $E(Y) = \beta_0 + \beta_1 X$

Typically, in practice, β_0 and β_1 are unknown parameters. We estimate them using the sample data.

Response Variable (Y): Measures the major outcome of interest in the study.
Predictor Variable (X): Another variable whose value explains, predicts, or is associated with the value of the response variable.

Fitting the Model (Least Squares Method)

If we gather data (*X*, *Y*) for several individuals, we can use these data to estimate β_0 and β_1 and thus estimate the linear relationship between *Y* and *X*.

First step: Decide if a straight-line relationship between *Y* and *X* makes sense.

Plot the bivariate data using a scattergram (scatterplot).

Once we settle on the "best-fitting" regression line, its equation gives a predicted Y-value for any new X-value.

How do we decide, given a data set, which line is the best-fitting line?