Regression problem deals with the relationship between the frequency distribution of one (dependent) variable and another (independent) variable(s) which is (are) held fixed at each of several values.

The technical term "regression" has become part of the language of statistics due to the work of Sir Francis Galton. In the 1880's, Galton laid the foundations of modern correlation techniques in a study of the relationship between the average heights of children and the heights of their parents.

In statistics, "regression" means simply average relationship, and thus between two variables, y and x,

\[
\text{Regression of variable } y \text{ on variable } x
\]

implies a relationship between

- the average of the values of the variable y for a given value of the variable x,

and

- the value of the variable x.
Note that the presumption of being dependent and independent for respective variables $y$ and $x$, does not necessarily mean that a causality (a cause-effect or an input-output relationship) must exist between them, even though they may highly correlated. Because,

- both variables may be affected by the same cause, or
- two variables may be interdependent, or
- one variable is the cause, although not necessarily the sole cause, of the other.

Note also that

**a regression of $y$ on $x$**

can only be used to estimate the mean value of $y$ for a given $x$, and should not be used to estimate the mean value for $x$ for a given $y$. In other words,

**a regression of $y$ on $x$**

does not immediately imply

**a regression of $x$ on $y$**

However, there may exists a regression of $x$ on $y$ but in a totally different nature.

In the practice of engineering experimentation, the regression analysis is used to estimate the “best” empirical constants, with their respective confidence limits, to fit a mathematical model to a set of measurement data.

Once a mathematical model is so established between the dependent variable $y$ and the independent variable(s) $x$, it can then be used to *predict $y$ for new values of $x*, by treating $x$ as a continuous variable in the implied interpolation process involved.
**Linear Regression:**
It is the regression in which the model used is linear; i.e.,
\[ y = ax + b \]

**Curvilinear Regression:**
It is the regression in which the model used is a polynomial in \( x \); i.e.,
\[ y = f(x) \]

**Nonlinear Regression:**
It is the regression in which the model used is nonlinear (polynomial or not); e.g.,
\[ y = ax + be^{-cx} \]

**Multivariate (Multiple) Regression:**
It is the regression in which there exist multiple independent variables used in a linear model; e.g.,
\[ y = a_1 x_1 + a_2 x_2 + b \]

or in a nonlinear model; e.g.,
\[ y = a(1-x_1^2)x_2 + bx_3 \]
**Method of Least Squares:**

It is a general method used in a very broad class of engineering problems like
- curve fitting,
- parameter estimation,
- system identification, and
- static and dynamic optimization

Its major advantage over other techniques is that it results in a set of linear algebraic equations in terms of unknown model parameters if these parameters appear linearly in the mathematical model.

**Example:** Let $x_1, x_2, \ldots, x_n$ constitute $n$ measurements. A best estimate $x_{\text{best}}$ of these measurements is asked in the sense that the quantity

$$E = \sum_{i=1}^{n} (x_i - x_{\text{best}})^2$$

is minimized.

The minimization of $E$ with respect to $x_{\text{best}}$ requires that

$$\frac{dE}{dx_{\text{best}}} = \frac{d}{dx_{\text{best}}} \sum_{i=1}^{n} (x_i - x_{\text{best}})^2 = -2 \sum_{i=1}^{n} (x_i - x_{\text{best}}) = 0$$

$$\sum_{i=1}^{n} x_i - n x_{\text{best}} = \sum_{i=1}^{n} x_i - nx_{\text{best}} = 0$$

which is nothing but the *arithmetic mean* of the data set given.
Example (Linear Regression): Let \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\) constitute \(n\) paired data points. A best linear regression of \(y\) on \(x\) as

\[
y = ax + b
\]

is asked (i.e., the best estimates of \(a\) and \(b\) are asked) in the sense that the quantity

\[
E = \sum_{i=1}^{n} [y_i - (ax_i + b)]^2
\]

is minimized.

Note that the quantity

\[
y_i - (ax_i + b)
\]

is the vertical distance between the data point \(i\) and the regressed value of \(y\) (=\(ax_i + b\)) at \(x_i\) in the \(y\) versus \(x\) plane.

The minimization of \(E\) with respect to \(a\) and \(b\) requires that

\[
\frac{dE}{da} = 0 \quad \text{and} \quad \frac{dE}{db} = 0
\]
The solution of last two equations in terms of $a$ and $b$ gives:

\[
a = \frac{n \sum_{i=1}^{n} x_i y_i - \left( \sum_{i=1}^{n} x_i \right) \left( \sum_{i=1}^{n} y_i \right)}{n \sum_{i=1}^{n} x_i^2 - \left( \sum_{i=1}^{n} x_i \right)^2}
\]

\[
b = \frac{\left( \sum_{i=1}^{n} y_i \right)\left( \sum_{i=1}^{n} x_i^2 \right) - \left( \sum_{i=1}^{n} x_i y_i \right)\left( \sum_{i=1}^{n} x_i \right)}{n \sum_{i=1}^{n} x_i^2 - \left( \sum_{i=1}^{n} x_i \right)^2}
\]

or by defining

\[
\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i
\]
\[
\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i
\]

\[
\overline{x^2} = \frac{1}{n} \sum_{i=1}^{n} x_i^2
\]
\[
\overline{xy} = \frac{1}{n} \sum_{i=1}^{n} x_i y_i
\]

\[
a = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2}
\]

and

\[
b = \frac{\overline{x^2} \cdot \overline{y} - \overline{x} \cdot \overline{xy}}{\overline{x^2} - \overline{x}^2}
\]

or

\[
\begin{bmatrix} a \\ b \end{bmatrix} = \frac{1}{\overline{x^2} - \overline{x}^2} \begin{bmatrix} -\overline{x} & 1 \\ \overline{x^2} - \overline{x} & \overline{xy} \end{bmatrix} \begin{bmatrix} \overline{y} \\ \overline{xy} \end{bmatrix}
\]

Note also that

\[
b = \overline{y} - a \overline{x}
\]
Example: Let the following data show the result of an experiment as 6 pairs of values.

<table>
<thead>
<tr>
<th>i</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_i$</td>
<td>0.5</td>
<td>1.0</td>
<td>1.5</td>
<td>2.0</td>
<td>2.5</td>
<td>3.0</td>
</tr>
<tr>
<td>$y_i$</td>
<td>0.71</td>
<td>1.33</td>
<td>1.68</td>
<td>1.88</td>
<td>2.31</td>
<td>2.48</td>
</tr>
</tbody>
</table>

It is desired to obtain the linear regression of $y$ on $x$.

The computed values are:

$$
\bar{x} = 1.75 \ (s_x=0.935) \quad \bar{y} = 1.732 \ (s_y=0.652) ;
$$

$$
\bar{x^2} = 3.792 \quad \bar{xy} = 3.530
$$

to give $a = 0.685$ and $b = 0.533$

for a linear regression expression of $y$ on $x$:

$$
y = 0.685 \ x + 0.533
$$

<table>
<thead>
<tr>
<th>$x_i$</th>
<th>0.5</th>
<th>1.0</th>
<th>1.5</th>
<th>2.0</th>
<th>2.5</th>
<th>3.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_i$</td>
<td>0.71</td>
<td>1.33</td>
<td>1.68</td>
<td>1.88</td>
<td>2.31</td>
<td>2.48</td>
</tr>
<tr>
<td>$y_{ti}=ax_i+b$</td>
<td>0.876</td>
<td>1.218</td>
<td>1.561</td>
<td>1.903</td>
<td>2.246</td>
<td>2.588</td>
</tr>
<tr>
<td>$e_i=y_{ti}-y_i$</td>
<td>0.166</td>
<td>-0.112</td>
<td>-0.119</td>
<td>0.023</td>
<td>-0.065</td>
<td>0.108</td>
</tr>
</tbody>
</table>
Correlation Coefficient:

It is a measure of the degree of linear correlation existing between \( y \) and \( x \). It is defined as:

\[
r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1)S_x S_y}
\]

or

\[
r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}
\]

or

\[
r = \frac{\bar{xy} - \bar{x} \cdot \bar{y}}{\sqrt{(\bar{x}^2 - \bar{x}^2)(\bar{y}^2 - \bar{y}^2)}}
\]

or if the regression coefficient \( a \) has already been calculated

\[
r = a \frac{S_x}{S_y}
\]
The correlation coefficient \( r \) always lies between -1 and +1. If and only if all data points lie on the regression line, then \( r = \pm 1 \). If \( r = 0 \), the regression does not explain anything about the variation of \( y \), and the regression line is horizontal; that is \( y = b = \bar{y} \).

The original definition of the correlation coefficient can be interpreted as

\[
\frac{r^2}{\text{variation due to regression}} = \frac{\text{total variation}}{\text{variation due to regression}}
\]

where \( r^2 \) is referred to as the coefficient of determination.

A negative correlation coefficient \( (r<0) \) is simply an indication of an inverse correlation between \( y \) and \( x \), leading to a negative slope \( (a<0) \) for the regression line.

For the last example, \( r^2=0.967 \) \( (r=0.983) \), indicating the existence of a very strong correlation between \( y \) and \( x \). It can also be commented that, the 96.7 % of the total dispersion of data points \( y_i \) from their overall mean value \( \bar{y} \) can be explained by the existing regression between \( y \) and \( x \).
Standard Deviation of Data From the Regression Line (Standard Error of Estimate):

The deviation of each data point from the regression line is written as:

\[ y_i - y_{ti} = y_i - (ax_i + b) \]

Then, as a measure of the vertical scatter (dispersion) of data, the standard error of estimate is defined as:

\[
s_{y,x} = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} (y_i - y_{ti})^2}
\]

Note that \( s_{y,x}^2 \) can be considered as the estimate of the variance of \( y \) left unexplained by the regression of \( y \) on \( x \). It is defined with \( n-2 \) rather than \( n-1 \) in the denominator because two degrees of freedom are used in estimating \( a \) and \( b \).

The followings are two equivalent expressions for \( s_{y,x} \) used in its calculation if a regression is already done:

\[
s_{y,x} = \sqrt{\frac{n-1}{n-2} \left( s_y^2 - a^2 s_x^2 \right)}
\]

\[
s_{y,x} = \sqrt{\frac{n-1}{n-2} s_y^2(1 - r^2)}
\]

For the last example, \( s_{y,x} = 0.133 \).
**Standard Error of Slope:**

\[ s_a = \frac{s_{y,x}}{s_x \sqrt{n - 1}} = \frac{s_{y,x}}{\sqrt{n}} \cdot \frac{1}{\sqrt{x^2 - x^2}} \]

For the last example \( s_a = 0.064 \)

**Standard Error of \( y \)-Intercept:**

\[ s_b = s_{y,x} \sqrt{\frac{1}{n} + \frac{x^2 - x^2}{s_x^2(n - 1)}} = s_{y,x} \cdot \frac{x^2}{\sqrt{n} \cdot \sqrt{x^2 - x^2}} \]

For the last example \( s_b = 0.124 \)

**Standard Error of Mean Value of \( y_t \) \( @ \) any \( x \):**

\[ s_{y_t} = s_{y,x} \cdot \sqrt{1 + \frac{(x - x)^2}{x^2 - x^2}} \]

For the last example \( s_{y_t} = 0.096 \) \( @ \) \( x = 3 \)

**Standard Error of Estimation of \( y_t \) \( @ \) any \( x \):**

\[ s_{y_t} = s_{y,x} \cdot \sqrt{1 + n + \frac{(x - x)^2}{x^2 - x^2}} \]

For the last example \( s_{y_t} = 0.164 \) \( @ \) \( x = 3 \)

**Confidence Limits:**

Assume Gaussian distribution with a degree of confidence 90 % \( \Rightarrow z = 1.645 \)
Confidence of Slope: $\Delta a = \pm z s_a$
For the last example: $a = 0.685 \pm 1.645 \times 0.064$

\[
\begin{array}{c|c|c|c|c|c|c}
0.0 & 0.5 & 1.0 & 1.5 & 2.0 & 2.5 & 3.0 \\
\hline
0.0 & 0.5 & 1.0 & 1.5 & 2.0 & 2.5 & 3.0 \\
\end{array}
\]

Confidence of Slope

Confidence of $y$-Intercept: $\Delta b = \pm z s_b$
For the last example: $b = 0.533 \pm 1.645 \times 0.124$

\[
\begin{array}{c|c|c|c|c|c|c}
0.0 & 0.5 & 1.0 & 1.5 & 2.0 & 2.5 & 3.0 \\
\hline
0.0 & 0.5 & 1.0 & 1.5 & 2.0 & 2.5 & 3.0 \\
\end{array}
\]

Confidence of $y$-Intercept
**Confidence of Mean Value of \(y_t\) at \(x(=3)\):**

\[
\Delta y_t = \pm z \ s_{yt}
\]

For the last example:

\[
y_{yt} = 2.588 \pm 1.645 \times 0.096
\]

\[
y_{yt} = 2.588 \pm 0.158
\]

**Confidence of Estimation of \(y_t\) at \(x(=3)\):**

(i.e., the size of the band at \(x(=3)\) which is expected to include 90% of the new data points)

\[
\Delta y_e = \pm z \ s_{ye}
\]

For the last example:

\[
y_{ye} = 2.588 \pm 1.645 \times 0.164
\]

\[
y_{ye} = 2.588 \pm 0.270
\]