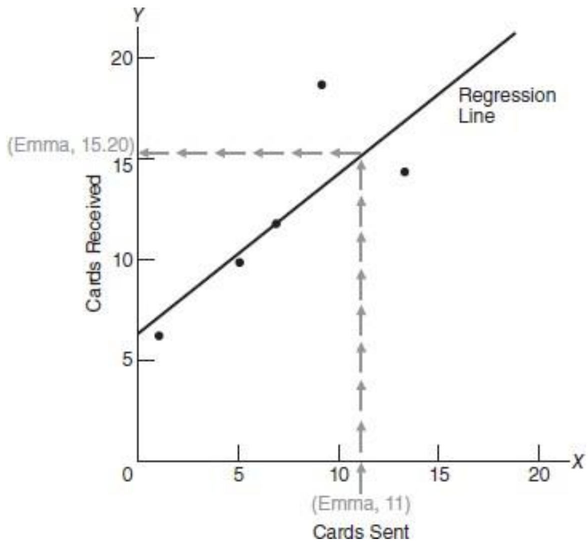
REGRESSION

A regression is a statistical technique that relates a dependent variable to one or more independent (explanatory) variables. A regression model is able to show whether changes observed in the dependent variable are associated with changes in one or more of the explanatory variables. Regression captures the correlation between variables observed in a data set, and quantifies whether those correlations are statistically significant or not.

A Regression Line

A regression line is a line that best describes the behaviour of a set of data. In other words, it's a line that best fits the trend of a given data.



The purpose of the line is to describe the interrelation of a dependent variable (Y variable) with one or many independent variables (X variable). By using the equation obtained from the regression line an analyst can forecast future behaviours of the dependent variable by inputting different values for the independent ones.

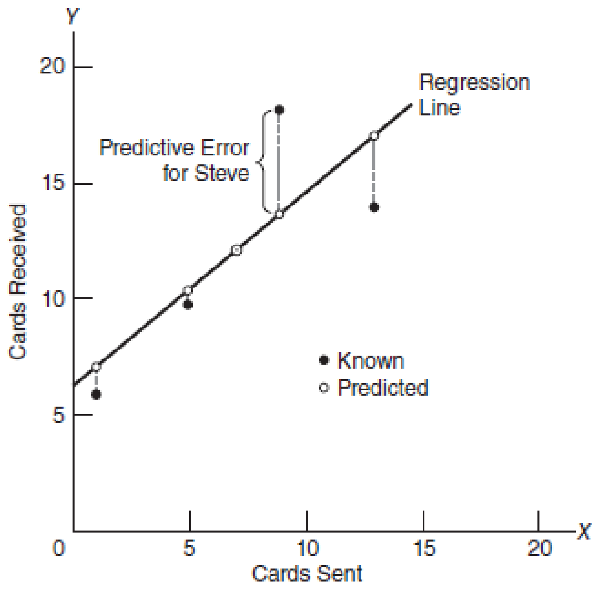
Types of regression

The two basic types of regression are

- Simple linear regression Simple linear regression uses one independent variable to explain or predict the outcome of the dependent variable Y.
- Multiple linear regression Multiple linear regressions use two or more independent variables to predict the outcome.

Predictive Errors

Prediction error refers to the difference between the predicted values made by some model and the actual values.



LEAST SQUARES REGRESSION LINE

The placement of the regression line minimizes not the total predictive error but the total squared predictive error, that is, the total for all squared predictive errors. When located in this fashion, the regression line is often referred to as the least square's regression line.

The Least Squares Regression Line is the line that minimizes the sum of the residuals squared. The residual is the vertical distance between the observed point and the predicted point, and it is calculated by subtracting ˆy from y.

Formula

y' = bx+a b – slope , a – y intercept

$$b = \frac{N\,\Sigma(xy) - \Sigma x\,\Sigma y}{N\,\Sigma(x^2) - (\Sigma x)^2}$$

$$b = \frac{\Sigma y - m\,\Sigma x}{N}$$

Example

"x"	"y"
2	4
3	5
5	7
7	10
9	15

Step 1: For each (x, y) calculate x^2 and xy:

x	y	x^2	xy
2	4	4	8
3	5	9	15
5	7	25	35
7	10	49	70
9	15	81	135

Step 2: Sum x, y, x2 and xy (gives us Σx, Σy, Σx2 and Σxy):

Σx: 26 Σy: 41 Σx²: 168 Σxy: 263

Step 3: Calculate Slope b

$$b = \frac{N\,\Sigma(xy) - \Sigma x\,\Sigma y}{N\,\Sigma(x^2) - (\Sigma x)^2}$$

$$= \frac{5 \text{ x } 263 - 26 \text{ x } 41}{5 \text{ x } 168 - 262}$$

$$= \frac{1315 - 1066}{840 - 676}$$

$$= \frac{249}{164}$$

b = 1.5183.

Step 4: Calculate Intercept a

$$a = \frac{\Sigma y - b\,\Sigma x}{N}$$

$$= \frac{41 - 1.5183 \text{ x } 26}{5}$$

a = 0.3049.

Step 5: y' = bx+a

y' = 1.518x + 0.305

x	y	y = 1.518x + 0.305	error
2	4	3.34	−0.66
3	5	4.86	−0.14
5	7	7.89	0.89
7	10	10.93	0.93
9	15	13.97	−1.03

To predict the y value, we can assume any value for x.

Assume x = 8.

Then y = 1.518 x 8 + 0.305

= 12.45