Machine Learning

Introduction to Correlation and Regression

Introduction to Correlation & Regression

- **Correlation** is a statistical method used to determine whether a relationship between variables exists or not.
- **Regression** is a statistical method used to describe the nature of the relationship between variables, that is, positive or negative, linear or nonlinear.

Introduction to Correlation & Regression

- There are two types of relationships:
  - *Simple relationship* and
  - *Multiple relationship*

- In a **simple relationship,** there are two variables—an **independent variable,** also called an explanatory variable or a predictor variable, and a **dependent variable,** also called a response variable.

- **For example**, a manager may wish to see whether the number of years the sales people have been working for the company has anything to do with the amount of sales they make.

Introduction to Correlation & Regression

- In a **multiple relationship,** called *multiple regression,* two or more independent variables are used to predict one dependent variable.

- For example, an educator may wish to investigate the relationship between a student's success in college and factors such as the number of hours devoted to studying, the student's GPA, and the student's high school background.

Introduction to Correlation & Regression

Simple relationships can also be positive or negative.

- A **positive relationship** exists when both variables increase or decrease at the same time. For instance, a person's height and weight are related; and the relationship is positive, since the taller a person is, generally, the more the person weighs.

- In a **negative relationship,** as one variable increases, the other variable decreases, and vice versa. For example, if you measure the strength of people over 60 years of age, you will find that as age increases, strength generally decreases.

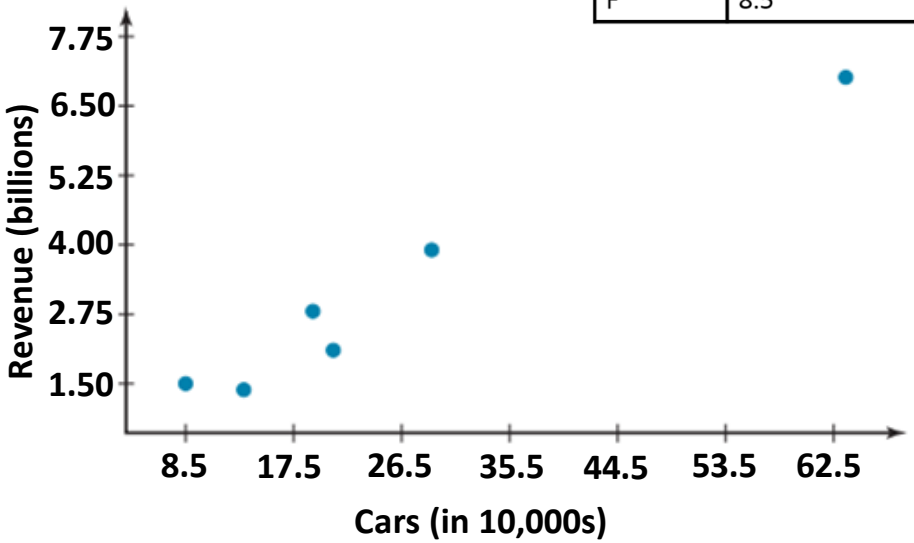
Machine Learning

Scatter Plots & Correlation

Scatter Plots & Correlation

A **scatter plot** is a graph of the ordered pairs (x, y) of numbers consisting of the independent variable x and the dependent variable y.

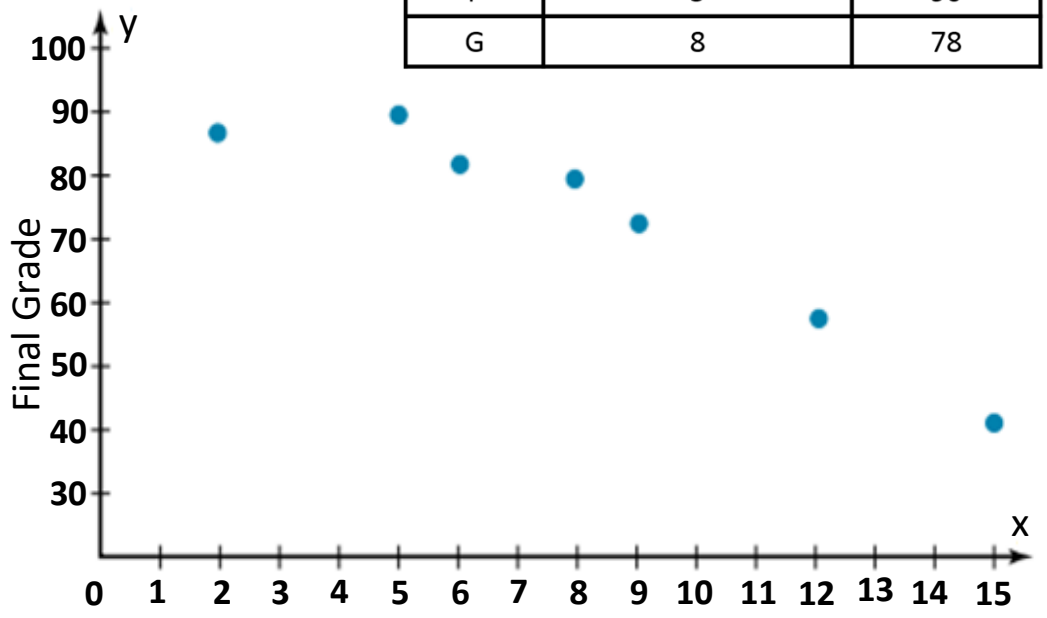
Company	Cars (in ten thousands)	Revenue (in billions)
A	63.0	$7.0
B	29.0	3.9
C	20.8	2.1
D	19.1	2.8
E	13.4	1.4
F	8.5	1.5



Scatter Plots and Correlation

Example:

Student	Number of absences x	Final grade y (%)
A	6	82
B	2	86
C	15	43
D	9	74
E	12	58
F	5	90
G	8	78

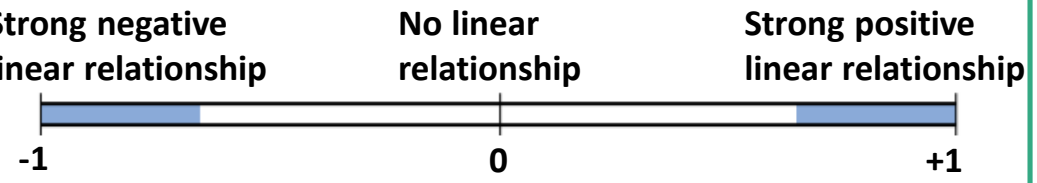


Machine Learning

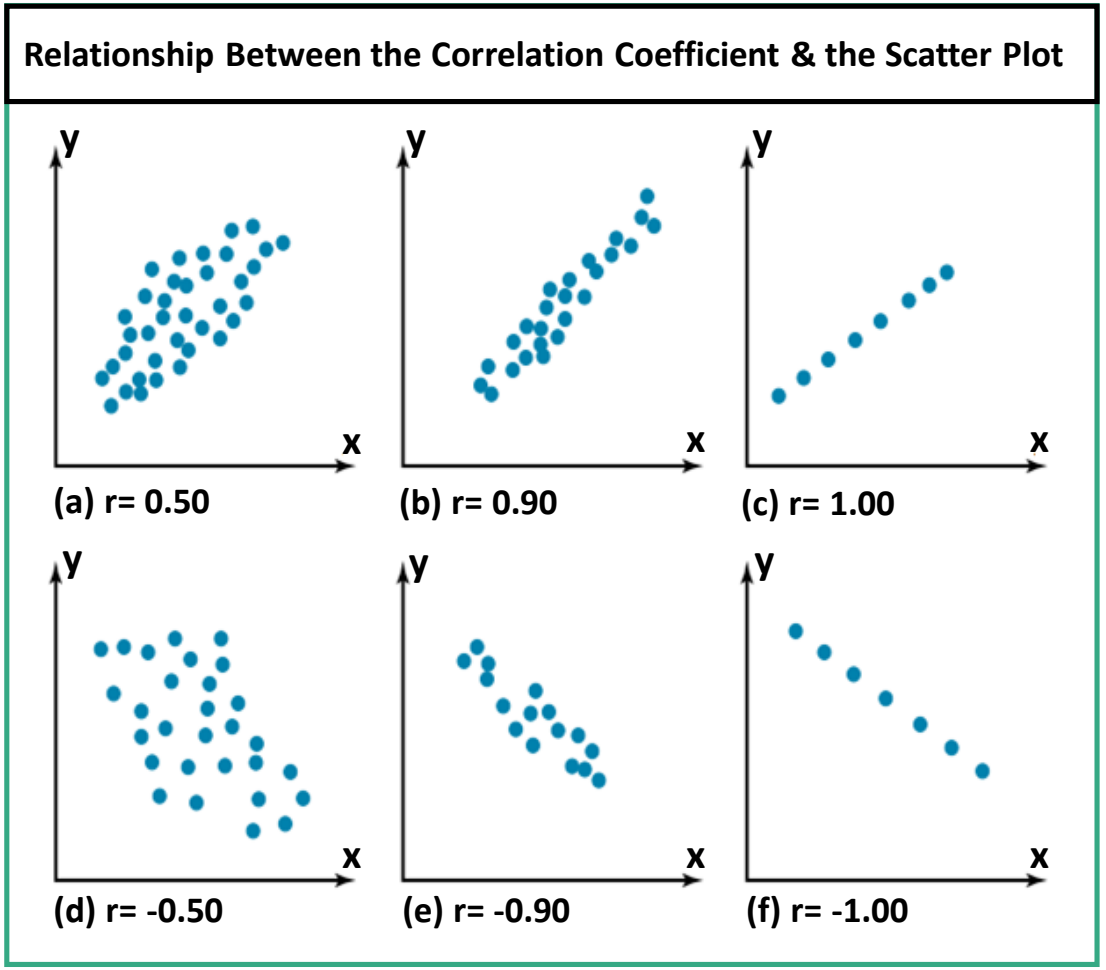
Correlation Coefficient

Correlation Coefficient

The **correlation coefficient** computed from the sample data measures the strength and direction of a linear relationship between two variables. The symbol for the sample correlation coefficient is *r.* The symbol for the population correlation coefficient is ρ (Greek letter rho).



Correlation Coefficient



Correlation Coefficient

Formula for the **correlation coefficient r**

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

Where *n* is the number of data pairs.

Correlation Coefficient

Example:

Student	Number of absences *x*	Final grade y (%)	Xy	x^2	y^2
A	6	82	492	36	6,724
B	2	86	172	4	7,396
C	15	43	645	225	1,849
D	9	74	666	81	5,476
E	12	58	696	144	3,364
F	5	90	450	25	8,100
G	8	78	624	64	6,084
	$\sum x=57$	**$\sum y=511$**	**$\sum xy=3745$**	**$\sum x^2=579$**	**$\sum y^2=38,993$**

Substitute in the formula and solve for r.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

$$= \frac{(7)(3745) - (57)(511)}{\sqrt{[(7)(579) - (57)^2][7(38{,}993) - (511)^2]}} = -0.944$$

Machine Learning

Regression: Line of Best Fit

Regression: Line of Best Fit

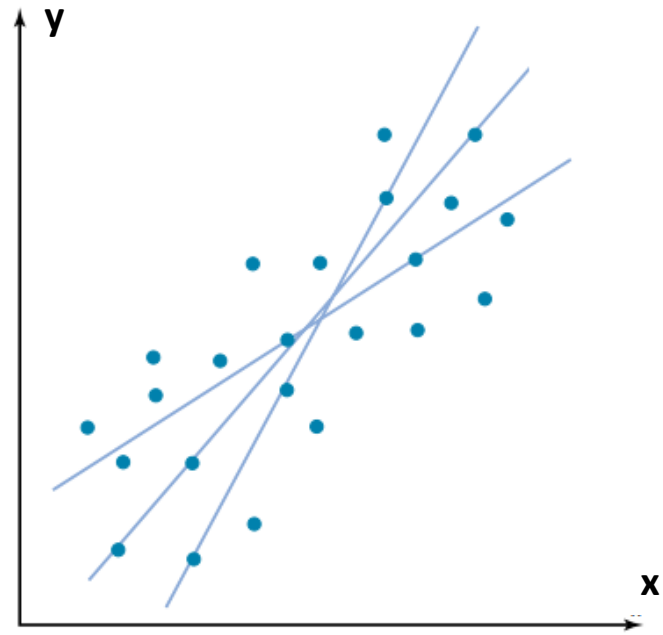
- In studying relationships between two variables, collect the data and then construct a scatter plot to determine the nature of the relationships (positive linear or negative linear or curvilinear, or no detectable relationships).

- After calculating correlation coefficient is to test the significance of the relationship.

- If the value of the correlation coefficient is significant, the next step is to determine the equation of the regression line, which is the data's line of best fit.

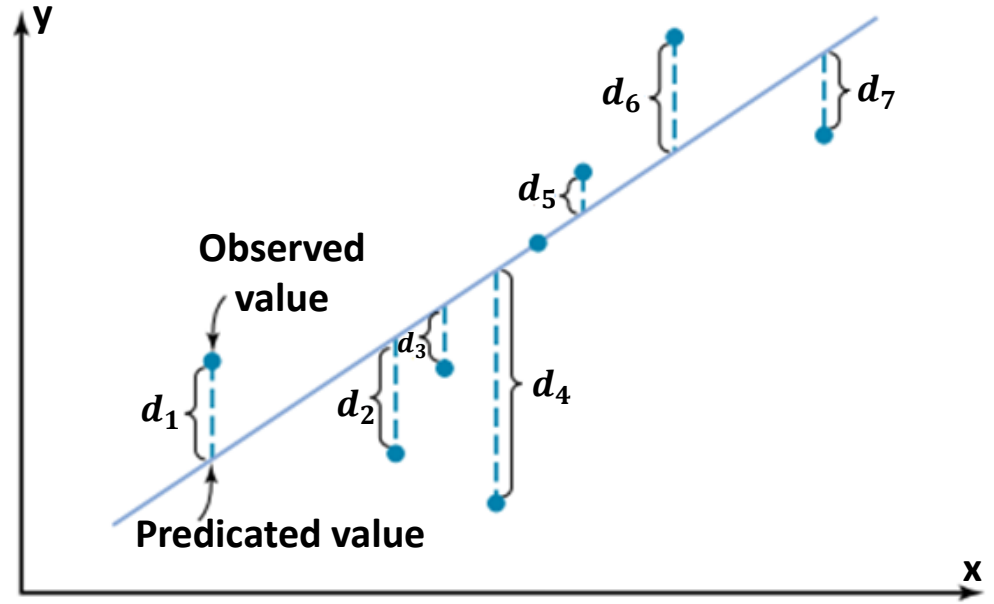
Regression: Line of Best Fit

The reason to need a line of best fit is that the values of y will be predicted from the values of x; hence, the closer points are to the line, the better the fit and the prediction will be.



Regression: Line of Best Fit

The reason to need a line of best fit is that the values of y will be predicted from the values of x; hence, the closer the points are to the line, the better the fit and the prediction will be.

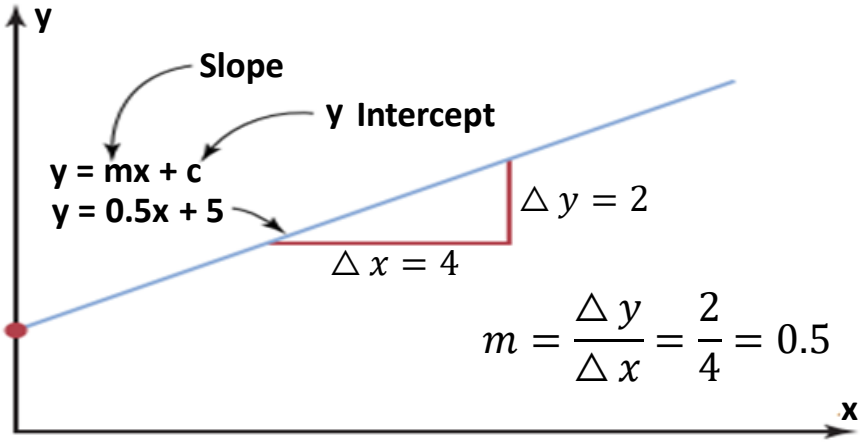


Machine Learning

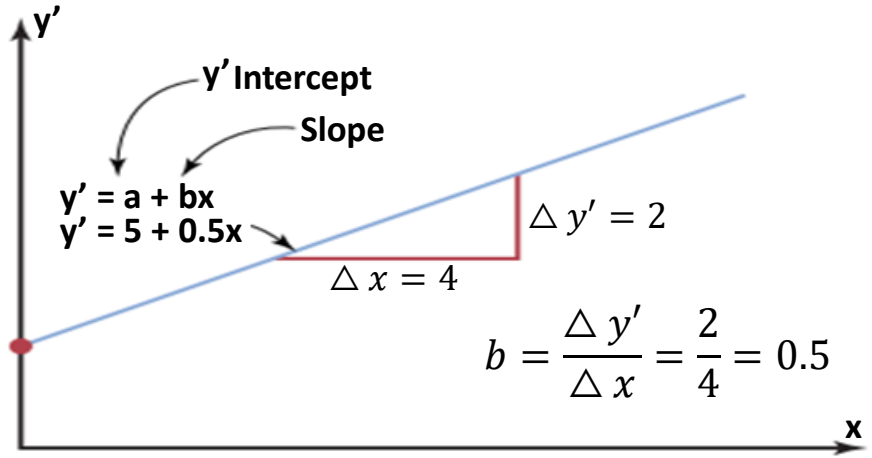
Regression Line Equation

Regression Line Equation

A line as represented in Algebra and in Statistics



(a) Algebra of a line



(b) Statistical notation for a regression line

Regression Line Equation

Formulas for the Regression Line

$$\mathbf{y' = a + bx}$$

where,

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

Where a is the y' intercept and b is the slope of the line.

Regression Line Equation

Example:

Student	Number of absences *x*	Final grade y (%)	Xy	x^2	y^2
A	6	82	492	36	6,724
B	2	86	172	4	7,396
C	15	43	645	225	1,849
D	9	74	666	81	5,476
E	12	58	696	144	3,364
F	5	90	450	25	8,100
G	8	78	624	64	6,084
	$\sum x=57$	**$\sum y=511$**	**$\sum xy=3745$**	**$\sum x^2=579$**	**$\sum y^2=38,993$**

Substitute in the formula and solve for ***r***:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

$$= \frac{(7)(3745) - (57)(511)}{\sqrt{[(7)(579) - (57)^2][7(38{,}993) - (511)^2]}} = -0.944$$

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2} = \frac{(511)(579) - (57)(3745)}{(7)(579) - (57)^2} = 102.493$$

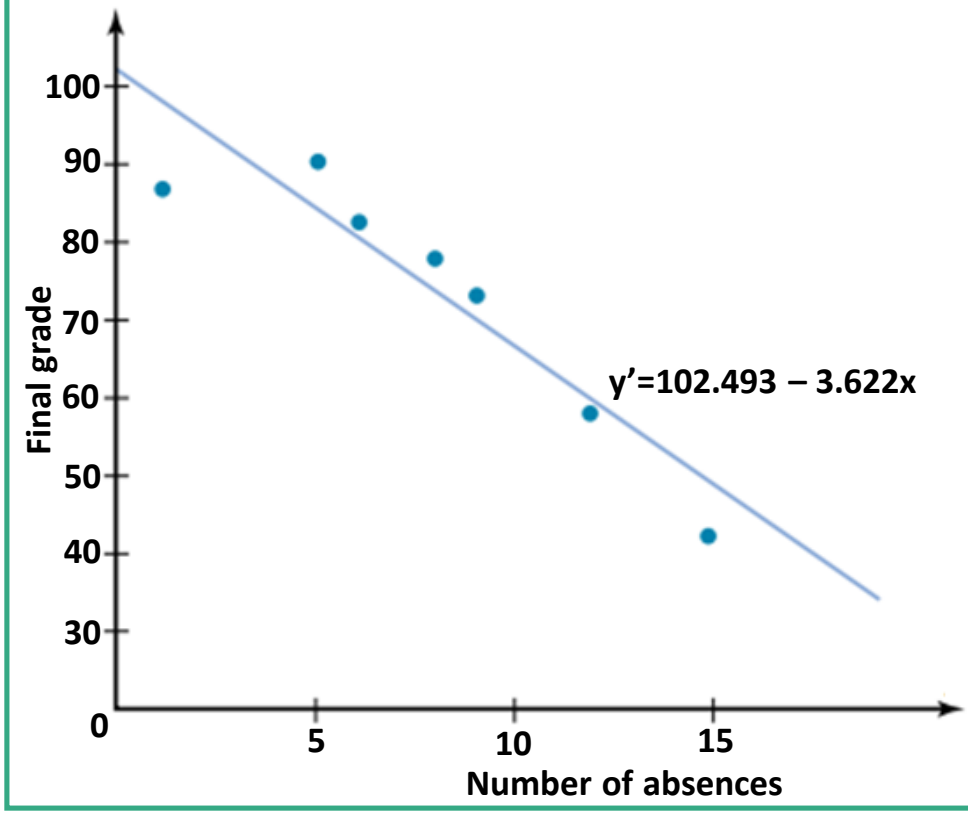
$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} = \frac{(7)(3745) - (57)(3745)}{(7)(579) - (57)^2} = -3.622$$

Hence, the equation of the regression line y'=a + bx is

y'=102.493 – 3.622x

Regression Line Equation

Regression Line for the Example:



Regression Model

- The regression model is $y = \beta_0 + \beta_1 x + \varepsilon$
- Data about x and y are obtained from a sample.
- From the sample of values of x and y, estimates b0 of β0 and b1 of β1 are obtained using the least squares or another method.
- The resulting estimate of the model is $\hat{y} = b_0 + b_1 x$
- The symbol is termed “y hat” and refers to the predicted values of the dependent variable y that are associated with values of x, given the linear model.

Machine Learning

Multiple Regression

Multiple Regression

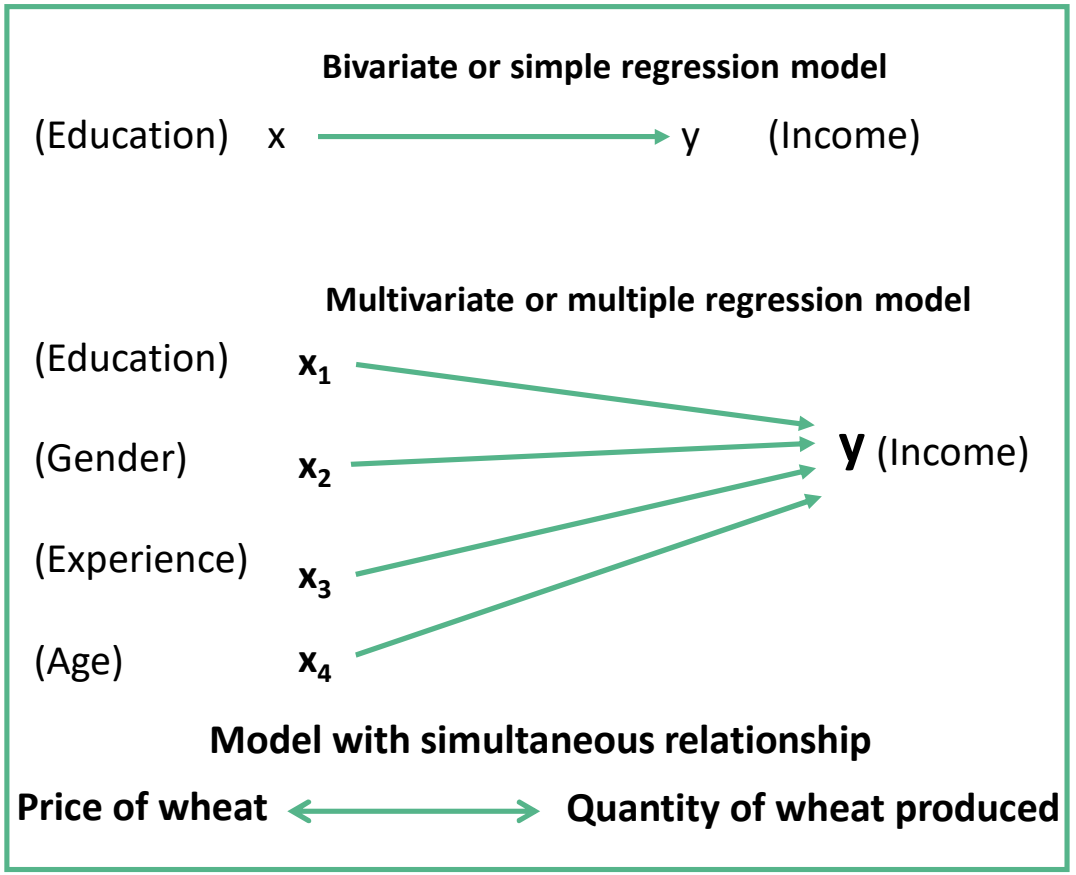
Multiple Regression

In **multiple regression,** there are several independent variables and one dependent variable, and the equation is

$$y' = a + b_1x_1 + b_2x_2 + \bullet \bullet \bullet + b_kx_k$$

where x_1, x_2 , … , x_k are the independent variables.

Bivariate & Multivariate Models



Multiple Regression

Example:

Student	GPA x_1	Age x_2	State board score y
A	3.2	22	550
B	2.7	27	570
C	2.5	24	525
D	3.4	28	670
E	2.2	23	490

The multiple regression equation obtained from the data is

$$\mathbf{y' = -44.81 + 87.64x_1 + 14.533x_2}$$

If GPA=3.0 and Age=25, then predicted State board score =

$$\mathbf{y' = -44.81 + 87.64\ (3.0) + 14.533\ (25)}$$
$$\mathbf{= 581.44 \text{ or } 581}$$

So the predicted State board score is 581.

Machine Learning

First Order Linear Model

First Order Linear Model

- $Y = b_0 + b_1X + \varepsilon$
  - Y = dependent variable
  - X = independent variable
  - b_0 = y-intercept
  - b_1 = slope of the line
  - ε = error variable

- The quantity ε is a random variable assumed to be normally distributed with $E(\varepsilon) = 0$ and $V(\varepsilon) = \sigma^2$.

- Without ε, any observed pair (X, Y) would fall exactly on the line $Y = b_0 + b_1X$ called the true regression line.

Let us go for a practical demonstration…

Machine Learning

Polynomial Regression

Polynomial Regression

- A form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modelled as an nth degree polynomial in x

$$y = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_m x_i^m + \epsilon_i (i = 1,2, \ldots, n)$$

- Used to describe nonlinear phenomena such as the growth rate of tissues, the distribution of carbon isotopes in lake sediments, and the progression of disease epidemics
- Can be expressed in matrix form: $\vec{y} = X\vec{\beta} + \vec{\epsilon}$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^m \\ 1 & x_2 & x_2^2 & \ldots & x_2^m \\ 1 & x_3 & x_3^2 & \cdots & x_3^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^m \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_m \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ . \\ . \\ . \\ \epsilon_n \end{bmatrix}$$

Machine Learning

Errors & Residuals

Errors & Residuals

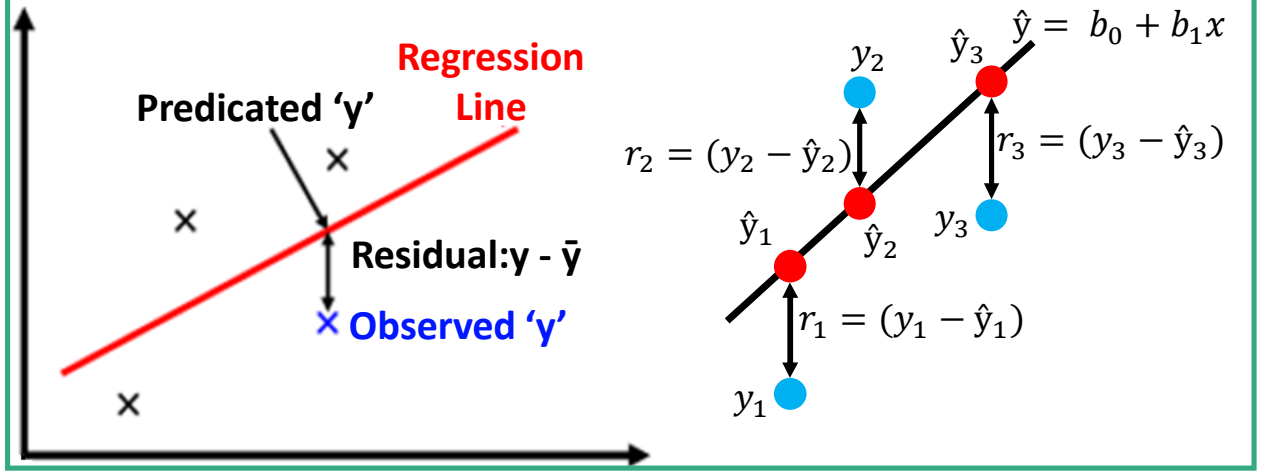
- A statistical **error** is the difference between an **observation** and its **expected** value which is based on the entire population. Usually values for the **entire population** are unobservable, e.g. mean height of all human beings
- A **residual** is the difference between an **observation** and its **estimated** value, e.g. mean height of a randomly chosen sample of human beings
- Residual Sum of Squares (RSS), also known as the Sum of Squared Residuals (SSR) is the sum of the squares of residuals.
- It is a measure of the discrepancy between the data and the estimation model and is used as an optimality criterion in parameter selection and model selection.
- In a standard linear simple regression model, $y_i = a + bx_i + \varepsilon_i$ $RSS = \sum_{i=1}^{n}(\varepsilon_i)^2 = \sum_{i=1}^{n}(y_i - (\alpha + \beta x_i))^2$ where α is the estimated value of a and β is the estimated value of the slope b.
- Minimizing the RSS function is a building block of supervised learning algorithms, and in the field of machine learning this function is referred to as the **cost function**.

Graphical Example Explaining Residual

- When you perform simple linear regression (or any other type of regression analysis), you get a line of best fit.
- The data points usually dot sometimes fall exactly on this regression equation line; they are scattered around.
- A **residual** is the vertical distance between a data point and the regression line. Each data point has one residual. They are positive if they are above the regression line and negative if below the line. If the line actually passes through the point, the residual at that point is zero.

Graphical Example Explaining Residual

- Say, the line has equation
  y=0.8x + 5
  Points are: (x1, y1) = (1, 5),
  (x2, y2) = (2, 7),
  (x3, y3) = (3, 6)
  The corresponding residuals are: r1(-0.8), r2(0.4) and r3(-1.4)



R-Square & Adjusted R Square

- R-Square determines how much of the total variation in Y (dependent variable) is explained by the variation in X (independent variable).

$$R - Square = 1 - \frac{\sum(y_{actual} - y_{predicated})^2}{\sum(Y_{actual} - Y_{mean})^2}$$

- The value of R-square is always between 0 and 1, where 0 means that the model does not explain any variability in the target variable (Y) and 1 meaning it explains full variability in the target variable.
- The Adjusted R-Square is a modified form of R-Square that has been adjusted for the number of predictors in the model. It incorporates model's degree of freedom. The adjusted R-Square only increases if the new term improves the model accuracy.

$$R^2 adjusted = 1 - \frac{(1 - R^2)\,(N - 1)}{N - P - 1}$$

where R^2 = sample R^2 value, N = total sample size and p = Number of predictors

Machine Learning

Regularization

Regularization

- Overfitting refers to a model that corresponds too closely or exactly to a particular set of training data, but fails to fit new data or predict future observations reliably.
- It is a statistical model that contains more parameters than can be justified by the data.
- Regularization is a very important technique in machine learning to prevent overfitting.
- Mathematically speaking, it adds a regularization term in order to prevent the coefficients to fit so perfectly to overfit.
- **L1-norm loss function** minimizes the sum of the absolute differences (S) between the target value (Y_i) and the estimated values ($f(x_i)$):

$$S = \sum_{i=1}^{n} |y_i - f(x_i)|$$

- **L2-norm loss function** minimizes the sum of the squares of the differences (S) between the target value (Y_i) and the estimated values ($f(x_i)$):

$$S = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

- Regularization adds a regularizer R(f) to a loss function.

$$\min_{f} \sum_{i=1}^{n} V(f(x_i), y_i) + \lambda R(f)$$

- Where V is the underlying loss function, e.g. L1 or L2 and λ = parameter which controls the importance of the regularizer

Machine Learning

Ridge & LASSO

Ridge & LASSO

- Ridge and LASSO (Least Absolute Shrinkage and Selection Operator) regression are powerful techniques generally used for creating compact models in presence of a “large” number of features. Problem with large number of features:
  - Enhance the tendency of a model to overfit (as low as 10 variables might cause overfitting)
  - Cause computational challenges. With modern systems, this situation might arise in case of millions or billions of features

Ridge & LASSO

- Both Ridge and LASSO work by **penalizing** the magnitude of coefficients of features along with minimizing the error between predicted and actual observations. The key difference is in how they assign penalty to the coefficients.
- **Ridge Regression:**
  Performs L2 regularization, i.e. adds penalty equivalent to **square of the magnitude** of coefficients
  Minimization objective = Least Squares Objective + α * (sum of square of coefficients)
- **LASSO Regression**:
  Performs L1 regularization, i.e. adds penalty equivalent to **absolute value of the magnitude** of coefficients
  Minimization objective = Least Squares Objective + α * (sum of absolute value of coefficients)

Ridge & LASSO (Cont.)

- Both Ridge and LASSO try to penalize the Beta coefficients so that we can get the important variables (all in case of Ridge and few in case of LASSO).
- If we take $\alpha = 0$, it will become Ridge and if $\alpha = 1$ it is LASSO.
- The major advantage of Ridge regression is coefficient shrinkage and reducing model complexity. So It is majorly used to prevent overfitting but not very useful in reducing number of features.
- Along with shrinking coefficients, LASSO performs feature selection as well - some of the coefficients become exactly zero, which is means that the feature is excluded from the model. So it is useful for modelling cases where the number of features are in millions or more

Let us go for a practical demonstration…