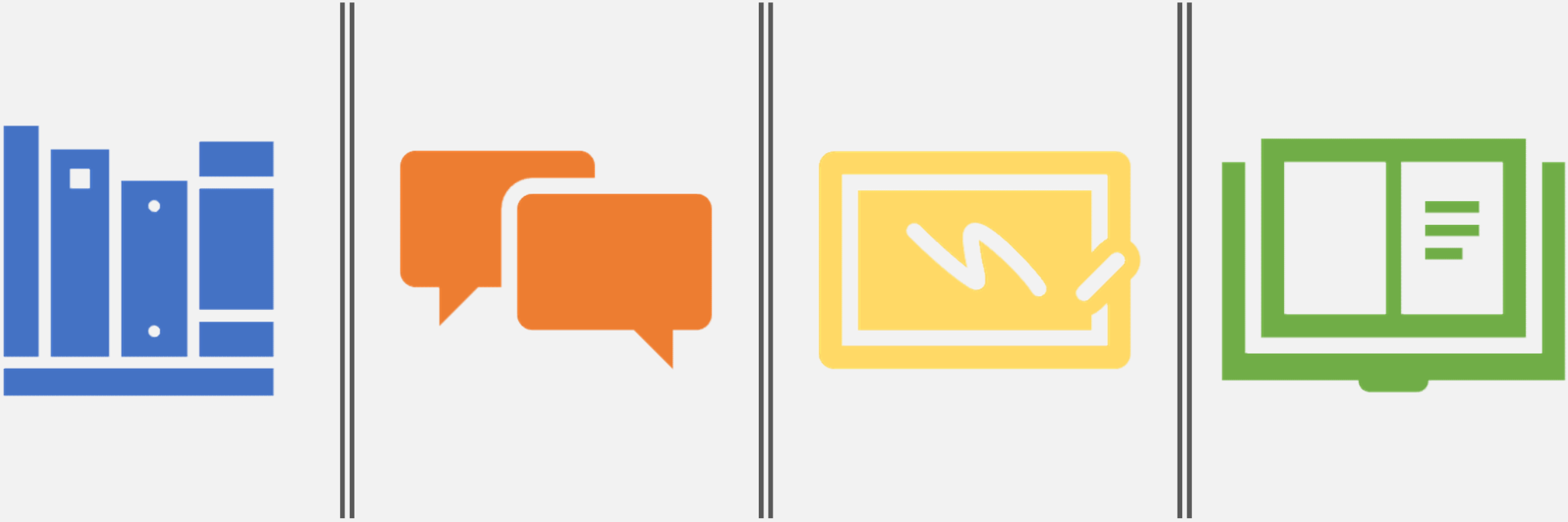
COMPUTATIONAL STATISTICS

AKASH NAKASHE



MULTIPLE LINEAR REGRESSION

Regression Predictive Modelling

Regression predictive modelling is the task of approximating a mapping function (f) from input variables (X) to a continuous output variable (y).

A continuous output variable is a real-value, such as an integer or floating point value. These are often quantities, such as amounts and sizes.

For example, a house may be predicted to sell for a specific dollar value, perhaps in the range of $100,000 to $200,000.

- A regression problem requires the prediction of a quantity.A regression can have real valued or discrete input variables.
- A problem with multiple input variables is often called a multivariate regression problem.
- A regression problem where input variables are ordered by time is called a time series forecasting problem.
- Because a regression predictive model predicts a quantity, the skill of the model must be reported as an error in those predictions.

Classification Predictive Modelling

Classification predictive modelling is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y).

The output variables are often called labels or categories. The mapping function predicts the class or category for a given observation.

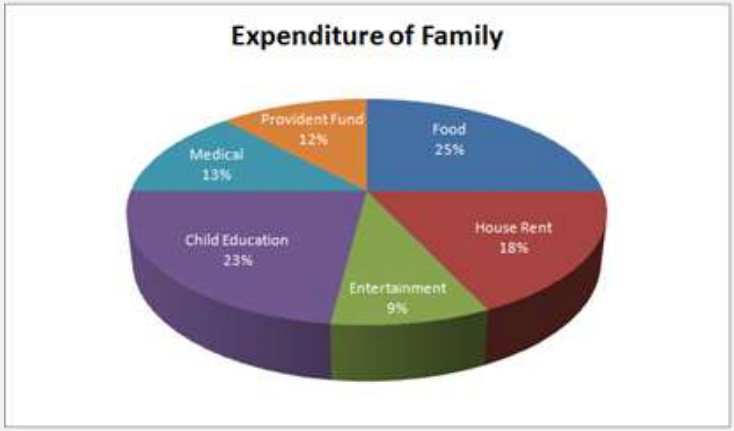
For example, an email of text can be classified as belonging to one of two classes: “spam“ and “*not spam*“.

- A classification problem requires that examples be classified into one of two or more classes.
- A classification can have real-valued or discrete input variables.
- A problem with two classes is often called a two-class or binary classification problem.
- A problem with more than two classes is often called a multi-class classification problem.
- A problem where an example is assigned multiple classes is called a multi-label classification problem.

Example 1

I have data of my monthly spending, monthly income and the number of trips per month for the last three years. Now I need to answer the following questions:

- What will be my monthly spending for next year?
- Which factor(monthly income or number of trips per month) is more important in deciding my monthly spending?
- How monthly income and trips per month are correlated with monthly spending?



Example 2

In the credit card industry, a financial company may be interested in minimizing the risk portfolio and wants to understand the top five factors that cause a customer to default. Based on the results the company could implement specific EMI options so as to minimize default among risky customers.

Example 3

A company wanted to be able to estimate or predict how much fuel they needed to transport building materials to their oil wells so that they could line them with concrete. The data provided was:

- Number of wells
- Depth of wells
- Distance to wells
- Weight of materials
- Ton kilometers
- Fuel costs

Example 4

An ecommerce company wants to measure

- the impact of product price,
- product promotions, and
- holiday seasonality on product sales.

- A product sales manager can discover which predictors included in the analysis will have significant impact on *product sales*.
- For the predictors with the most impact, the team can make important strategic decisions to meet product sales targets.
- For instance, if promotions and holiday seasons are significant factors, these factors should be given more focus when devising a marketing strategy.

Example 5

Predicting Gross Movie Revenue

- Success or failure of a movie can depend on many factors: star-power, release date, Critics review, budget, rating, plot and the highly unpredictable human reactions.
- Predicted revenues can be used for planning both the production and distribution stages.

Example 6

A TV industry analyst wants to build a statistical model for predicting the number of subscribers that a cable station can expect

Y = Number of cable subscribers

X_1 = Advertising rate which the station charges local advertisers for one minute of prime time space,
X_2 = Kilowatt power of the station's non-cable signal,
X_3 = Number of families living in the station's area of dominant influence,
X_4 = Number of competing stations

Purpose

Multiple regression analysis has three main uses:

- You can look at the strength of the effect of the independent variables on the dependent variable.
- You can use it to ask how much the dependent variable will change if the independent variables are changed.
- You can also use it to predict trends and future values.

Correlation

Definition:

- Correlation is a statistical measure (expressed as a number) that describes the size and direction of a relationship between two or more variables.
- Two variables are said to be correlated if change in one variable affects the change in other variable, and the relation between them is known as correlation.
- If two variables vary together they are said to be correlated.
- For example, the fuel consumed by a car is correlated to the number of miles travelled. The exact relationship is not easily found because so many other factors are involved such as speed of travel, condition of the engine, tyre pressures and so on. What is important is that as the mileage increases so does the fuel used – there is a correlation.

Correlation

Uses:

- Prediction: If there is a relationship between two variables, we can make predictions about one from another.
- Validity: Concurrent validity (correlation between a new measure and an established measure).
- Reliability: Test-retest reliability (are measures consistent).
- Inter-rater reliability (are observers consistent).
- Theory verification
- Predictive validity.

Correlation

Types of correlations:

- Positive Correlation: Two variables are said to be positively correlated if they deviates in the same direction. e.g. height & weight, income & expenditure

- Negative Correlation: Two variables are said to be negatively correlated if they deviates in the opposite directions. e.g. volume and pressure of a perfect gas, price and demand

- No Correlation: Two variables are said to be uncorrelated or statistically independent if there is no relation between them.

Simple Linear Regression

- Regression analysis is a statistical technique for investigating and modeling the relationship between variables.
- Regression analysis is the process of constructing a mathematical model or function that can be used to predict or determine one variable by another variable or other variables.
- The most elementary regression model is called simple regression or bivariate regression involving two variables in which one variable is predicted by another variable.
- In simple regression, the variable to be predicted is called the dependent variable and is designated as **y**. The predictor is called the independent variable, or explanatory variable, and is designated as **x**. In simple regression analysis, only a straight-line relationship between two variables is examined.
- In math courses, the slope-intercept form of the equation of a line often takes the form

$$y=mx+b$$

Where

m = slope of the line

b = y intercept of the line

Simple Linear Regression

In statistics, the slope-intercept form of the equation of the regression line through the population points is

$$\hat{y} = \beta_0 + \beta_1 x$$

Where

$\hat{y}$ = the predicted value of y

β_0 = the population y intercept

β_1 = the population slope

It is known as Deterministic models or mathematical models that produce an “exact” output for a given input.

Multiple Linear Regression

- By extending the simple regression model to a multiple regression model with two independent variables, it is possible to determine the multiple regression equation for any number of unknowns.

- Multiple regression analysis is similar in principle to simple regression analysis.

- However, it is more complex conceptually and computationally.

- Extending this notion to multiple regression gives the general equation for the probabilistic multiple regression model.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots\ldots\ldots\ldots \beta_k x_k + \in$$

Multiple Linear Regression

Where

- k = the number of independent variables
- β_k = the partial regression coefficient for independent variable k
- β_3 = the partial regression coefficient for independent variable 3
- β_2 = the partial regression coefficient for independent variable 2
- β_1 = the partial regression coefficient for independent variable 1
- β_0 = the regression constant
- y = the value of the dependent variable

- In virtually all research, these values are estimated by using sample information.
- Shown here is the form of the equation for estimating y with sample information.

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots\ldots\ldots\ldots b_k x_k + \in$$

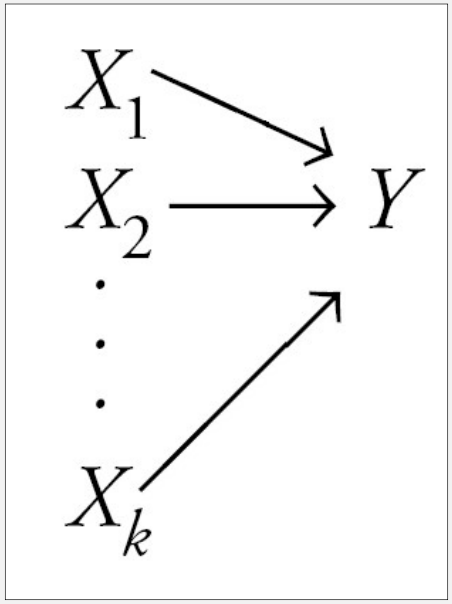
Multiple Linear Regression

Where

- k = the number of independent variables
- b_k = the estimate of the partial regression coefficient for independent variable k
- b_3 = the estimate of the partial regression coefficient for independent variable 3
- b_2 = the estimate of the partial regression coefficient for independent variable 2
- b_1 = the estimate of the partial regression coefficient for independent variable 1
- b_0 = the estimate of the regression constant
- y = the predicted value of y

Multiple Linear Regression

Multiple regression simultaneously considers the influence of multiple explanatory variables on a response variable Y



The intent is to look at the independent effect of each variable while “adjusting out” the influence of potential confounders

Case 1

- Suppose that we are statistical consultants hired by a client to provide advice on how to improve sales of a particular product.
- The Advertising data set consists of the sales of that product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: TV, radio, and newspaper.
- It is not possible for our client to directly increase sales of the product. On the other hand, they can control the advertising expenditure in each of the three media.
- Therefore, if we determine that there is an association between advertising and sales, then we can instruct our client to adjust advertising budgets, thereby indirectly increasing sales.
- In other words, our goal is to develop an accurate model that can be used to predict sales on the basis of the three media budgets.

Case 1

- In this setting, the advertising budgets are input variables while sales input is an output variable. So **X1** might be the TV budget, **X2** the radio budget, and **X3** the newspaper budget. The inputs go by different names, such as predictors, independent variables, features, predictor independent variable feature or sometimes just variables. The output variable—in this case, sales—is variable often called the response or dependent variable, and is typically denoted response dependent variable using the symbol **Y**.

Case 1

Here are a few important questions that we might seek to address:

1. Is there a relationship between advertising budget and sales?
    - Our first goal should be to determine whether the data provide evidence of an association between advertising expenditure and sales.
    - If the evidence is weak, then one might argue that no money should be spent on advertising!

2. How strong is the relationship between advertising budget and sales?
    - Assuming that there is a relationship between advertising and sales, we would like to know the strength of this relationship.
    - In other words, given a certain advertising budget, can we predict sales with a high level of accuracy? This would be a strong relationship.
    - Or is a prediction of sales based on advertising expenditure only slightly better than a random guess? This would be a weak relationship.

Case 1

Here are a few important questions that we might seek to address:

3. Which media contribute to sales?
    - Do all three media—TV, radio, and newspaper—contribute to sales, or do just one or two of the media contribute?
    - To answer this question, we must find a way to separate out the individual effects of each medium when we have spent money on all three media.

4. How accurately can we estimate the effect of each medium on sales?
    - For every dollar spent on advertising in a particular medium, by what amount will sales increase?
    - How accurately can we predict this amount of increase?

Case 1

Here are a few important questions that we might seek to address:

5. How accurately can we predict future sales?
    - For any given level of television, radio, or newspaper advertising, what is our prediction for sales, and what is the accuracy of this prediction?

6. Is the relationship linear?
    - If there is approximately a straight-line relationship between advertising expenditure in the various media and sales, then linear regression is an appropriate tool.

    - If not, then it may still be possible to transform the predictor or the response so that linear regression can be used.

In our case the MLR equation will become:

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} + \epsilon.$$

Assumptions

Assumption 1: Linearity

There must be a linear relationship between the outcome variable and the independent variables. Scatterplots can show whether there is a linear or curvilinear relationship.

Assumption 2: Auto Correlation

Linear regression analysis requires that there is little or no autocorrelation in the data. Autocorrelation occurs when the ***residuals*** are not independent from each other. For instance, this typically occurs in stock prices, where the price is not independent from the previous price.

Assumption 3: Homoscedasticity

The error terms must have constant variance. This phenomenon is known as ***homoscedasticity***. The presence of non-constant variance is referred to heteroscedasticity. The scatter plot is good way to check whether the data are homoscedastic (meaning the residuals are equal across the regression line).

Assumptions

Assumption 4: Outliers/influential cases:

As with simple linear regression, it is important to look out for cases which may have a disproportionate influence over your regression model.

Assumption 5: Multicollinearity:

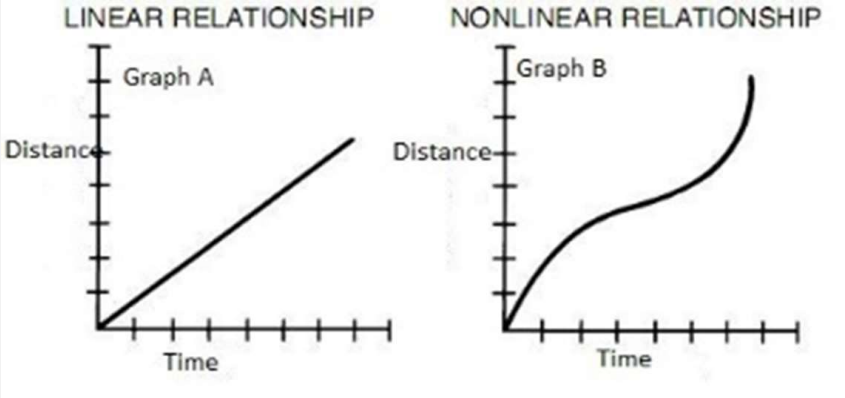
Multicollinearity exists when two or more of the explanatory variables are highly correlated. This is a problem as it can be hard to disentangle which of them best explains any shared variance with the outcome. It also suggests that the two variables may actually represent the same underlying factor. We can check Multicollinearity using VIF(variance inflation factor).

Assumption 6: Independence of Error

Residuals should be normally distributed. This can be checked by visualizing Q-Q Normal plot.

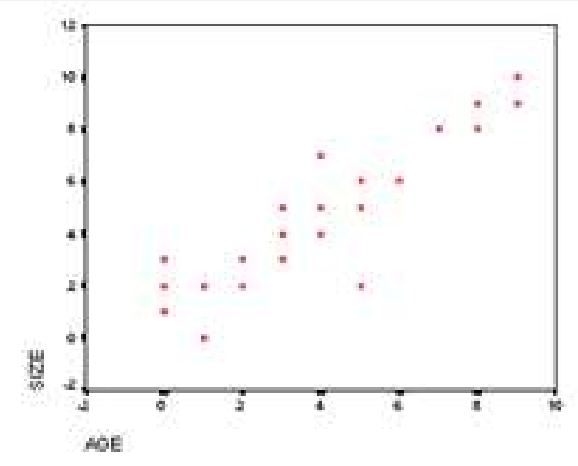
Assumption 1: Linearity

- First, multiple linear regression requires the relationship between the independent and dependent variables to be linear.
- The linearity assumption can best be tested with *scatterplots*.



- If data is given in pairs then the scatter diagram of the data is just the points plotted on the xy-plane.
- The scatter plot is used to visually identify relationships between the first and the second entries of paired data.

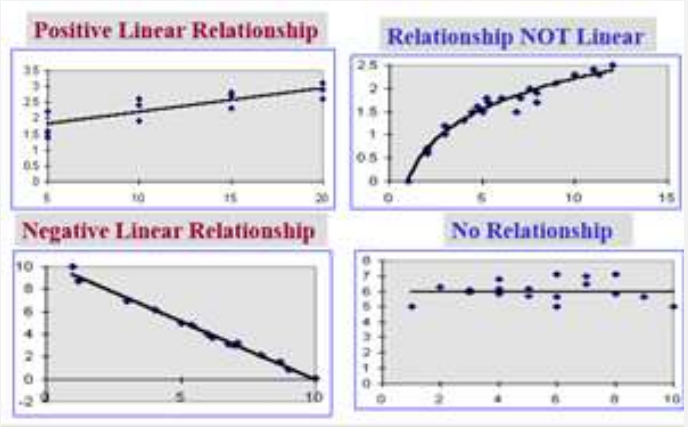
Assumption 1: Linearity



- The scatter plot above represents the age vs. size of a plant. It is clear from the scatter plot that as the plant ages, its size tends to increase. If it seems to be the case that the points follow a linear pattern well, then we say that there is a high linear correlation, while if it seems that the data do not follow a linear pattern, we say that there is no linear correlation. If the data somewhat follow a linear path, then we say that there is a moderate linear correlation.

- Given a scatter plot, we can draw the line that best fits the data

Assumption 1: Linearity



- A scatter plot matrix is a grid (or matrix) of scatter plots used to visualize bivariate relationships between combinations of variables.

- Each scatter plot in the matrix visualizes the relationship between a pair of variables, allowing many relationships to be explored in one chart.

Assumption 1: Linearity

Example: A soft drink bottler is trying to predict delivery times for a driver. He has collected data on the delivery time, the number of cartons delivered and the distance the driver walked. He wants to see the relationship between these three variables. We will use the scatter plot matrix to do this.

Delivery Time	X1 (Cartons)	X2 (Distance)
16.68	7	560
11.5	3	220
12.03	3	340
14.88	4	80
13.75	6	150
18.11	7	330
8	2	110
17.83	7	210
79.24	30	1460
21.5	5	605
40.33	16	688
21	10	215
13.5	4	255
19.75	6	462
24	9	448
29	10	776
15.35	6	200
19	7	132
9.5	3	36
35.1	17	770
17.9	10	140
52.32	26	810
18.75	9	450
19.83	8	635
10.75	4	150