

# STATISTICS

wikipedia.com - statistics is a mathematical science pertaining to the collection, analysis, interpretation or explanation, and presentation of data

## intro

more specifically, within the information maximisation -via data forecasting- part, there are two options

- a- direct application of regression methods  
[regression analysis]
- b- specific model application

group b- theories intend to extend the (really restricted) group a- methods applicability  
[more than 2 independent variables]

## a- regression methods

within this subgroup, just linear and nonlinear regressions are included  
[not PLS regressions]

advantage → acceptable accuracy and "relatively controllable uncertainty"  
[no arbitrary user intervention]

disadvantage → maximum number of independent variables is too low  
[max. 2 independent variables]

best solution? [trendingBot point of view]

1. virtually, any information can be adapted to the aforementioned duality of independent/dependent variables
2. ideas to bear in mind
  - NEVER extrapolate
  - predictive character only under certain conditions  
[i.e., minimum number of repetitions, goodness of the fit, etc.]
  - weightings/user defined parameters only under extreme circumstances  
["no equation" should be as valid as any numerical result]

but...

...no behaviour (no one worthy to be predicted) can be described by attending at a so low number of variables

## b- forecasting methods

within this subgroup, we include all the statistical methods whose (implicit) leitmotiv is extending the applicability of (group a-) regressions

[forecasts for situations involving more than 2 independent variables]

### b-1 time series analysis

wikipedia.com -

in statistics, signal processing, and many other fields, a time series is a sequence of data points, measured typically at successive times, spaced at (often uniform) time intervals - time series analysis comprises methods that attempt to understand such time series

there are many models specifically designed to maximise time series, that is, to understand the described behaviour and thus to predict future events on the basis of this information

#### 1. linear dependence

[~ linear regressions]

three main types

- autoregressive (AR) models
- integrated (I) models
- moving average (MA) models

additionally to these ones, there are still two combinations [autoregressive moving average (ARMA) models & autoregressive integrating moving average (ARIMA) models] and one generalisation [autoregressive fractionally integrated moving average (ARFIMA) models] of them

#### 2. non-linear dependence or autoregressive conditional heteroskedasticity models

[~ nonlinear regressions]

- generalised autoregressive conditional heteroskedacity (GARCH) models
- threshold autoregressive conditional heteroskedacity (TARCH) models
- exponential generalised autoregressive conditional heteroskedacity (EGARCH) models

...

trendingBot point of view

all these models have two characteristics in common

- a.- account for just two variables (dependent vs. independent)
- b.- try to understand stochastic (= random) processes

a.- why not applying conventional regression methods?

statistics' answer -> random essence has to be accounted for (!?)

- b.1.- stochastic/random ~ impossible to be predicted [see b-5] - ... then?
- b.2.- a weighted (based on sensible assumptions) regression method shouldn't be defined as stochastic, if the weightings are applied on a regular and consistent basis
- b.3.- probably, the randomness might be removed, for the case a more adequate set of variables would be chosen

CONCLUSION 1 –

time-series-analysing methods can be defined as extensions of conventional regression methods to stochastic behaviours (!)

CONCLUSION 2 –

trendingBot's result for any (stochastic) time series = "trend not found"

## b-2 extrapolation methods

wikipedia.com - in mathematics, extrapolation is the process of constructing new data points outside a discrete set of known data points

although there are no essential differences between extrapolation and interpolation methods [regressions], the expected accuracy from their results do differ quite appreciably; this fact and the main intention underlying the present classification [highlighting the opposition probable/predictable vs. random/unpredictable] are the only reasons explaining this specific subtype, outside the regression methods

nobody doubts that an increase in the uncertainty is the immediate consequence from any extrapolating process, however the logical attitude resulting from this idea seems to be not so clear; or at least this is what anyone could understand after noticing the wide variety of extrapolation types

- linear
- polynomial
- conic
- french curve

and even methods developed specifically for computer coding

- Richardson extrapolation
- Aitken extrapolation

trendingBot point of view

extrapolating has to be considered as the last resource and, in any case, to be clearly identified, in order to avoid any association extrapolation-interpolation

lame example

raw data -  $X$  (independent)  $\in [5,10]$  and  $Y$ (dependent)  $\in [10,20]$

- $Y$  values, for any  $X$  within the aforementioned range, can be predicted - 7.5 -> 15
- $Y$  values, for any  $X$  outside it, can only be roughly estimated - 15 -> 30

thus, predicting implies uncertainty but, usually, a more or less controllable one (a sensible set of minimum conditions has to be established in order to guarantee the predictive character) - (roughly) estimating implies uncontrollable uncertainty to be used just as a preliminary idea and the result from it never called "prediction"

## b-3 partial least squares (PLS) regression

wikipedia.com -

in statistics, the method of partial least squares regression (PLS-regression) bears some relation to principal component analysis; instead of finding the hyperplanes of minimum variance, it finds a linear model describing some predicted variables in terms of other observable variables

this method is recommended for cases where standard regressions show instability

[e.g., more predictors than observations or multicollinearity among predictors]

drawbacks

- partial solutions - mathematical-expressions-based [latent variables] outputs, rather than directly applicable expressions [equations]
- no predictive capabilities - qualitative results [most influential predictor over measurements, interdependence among predictors, etc.], instead of quantitative ones [equations]

trendingBot point of view

as soon as the number of independent variables increases beyond certain limit [2 can be taken as a good estimation], standard regression methods are not reliable enough and traditional statistics preferred to consider roughly-estimating approaches, rather than trying a different solution to the problem

lame example

40 values for 5 independent variables [X\_a, X\_b, X\_c, X\_d, X\_e], affecting a dependent one [Y\_1] (and, eventually, two additional dependent variables)

### 1. PLS regression [PLS path modelling]

- X\_c is the most influential variable over Y\_1
- all the variables, except X\_a & X\_e, are positively correlated with Y\_1
- from Y\_1, Y\_2 and Y\_3, it can be stated that every fluctuation in X\_b is compensated by the addition of X\_a & X\_c [evolution (among different phenomena) of any interest?]

### 2. trendingBot

NOTE: best trends = showing the lowest mean error after being applied to the original data

	expected error [%]
$Y_1 = X_a^{0.42} + 5.21 * X_c - X_e$	5.0
$Y_2 = X_c^{-1.3} * X_c - X_a$	3.6
$Y_3 = X_c - X_e / 2$	8.1

## b-4 principal component analysis

wikipedia.com -

principal component analysis (PCA) involves a mathematical procedure that transforms a number of possibly correlated variables into a smaller number of uncorrelated variables called principal components

its basic structure, equivalently to the one from PLS regressions, consists in two matrices: X [independent variables] & Y [dependent variable(s)]

the differences between both methods are direct consequence from the mathematical models used to relate these matrices

- linear model → PLS regression
- hyperplanes of minimum variance → principal component analysis

the aforementioned distinction is not relevant for the present study and thus the PLS regression section [b-3] describable enough

## b-5 probability-related methods [randomness]

wikipedia.com -

probability is the likelihood or chance that something is the case or that an event will occur

the most relevant theories are

- (generalised) method of moments
- bayesian methods
- predictive modelling
- method of instrumental variables (IV)
- 2SLS/3SLS
- seemingly unrelated regression

hence these methods do not predict future behaviours on the basis of past ones [effects on the dependent variable from variations in the independent one(s)], but the probability of an event [= invariant phenomenon = not describable as a result of the interaction between independent/dependent variables] to occur

trendingBot point of view (I)  
overview

any forecasting method

- predict definable & more-or-less-certain situations
- indicated for any “repetitive” behaviour

probability

- some certainty to pure randomness
- indicated for indefinable and/or random behaviours

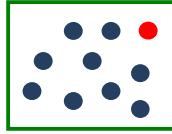
both theories are applicable to different fields

trendingBot	can not deal with random behaviours
probability	can not understand/describe phenomena, just estimate how likely they would happen under specific conditions

thus and from the point of view of the current classification, neither probability should be included within the forecasting methods or any method defined as such should deal with randomness

trendingBot point of view (II)  
detailed analysis

PROBLEM 1 – too simple behaviour [no description]



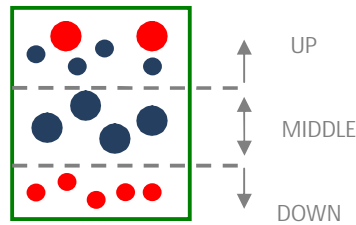
probability

- ideal for true/false-based problems [e.g., PROBLEM 1]
- probability of finding a red circle [true] = 10% (1/10)

trendingBot

- actuator description [independent] / action [dependent]
- PROBLEM 1 can not be solved
  - no description / 2 possible actions [dependent variables]

PROBLEM 2 – predictable behaviour



probability

- PROBLEM 2 → not-directly-applicable answers
- two possible actions [true/false] but no behaviour description [= multiple answers]
- solution
  - $P(\text{blue,up})=50\%$  -  $P(\text{blue,small,up})=100\%$  -  $P(\text{small, up})=44\%$  - ...
  - $P(\text{blue,small})/\text{all}=27\%$  -  $P(\text{blue,small})/\text{smalls}=44\%$  - ...
  - etc.

trendingBot

- PROBLEM 2 → extremely simple
- just one possible action for the described behaviour
  - [e.g., small blue circles up & down = “incoherent” (further variables required)]
- solution
  - pre-step → conversion into [arbitrarily-assigned] numerical values

property	non-numerical	numerical [arbitrary]
size	big	1
	small	2
colour	blue	1
	red	2
position	up	1
	medium	2
	down	3

independent variables [actuator description]	size	2	2	2	2	1	1	1	1	1	2	2	2	2
	colour	1	1	1	1	2	2	1	1	1	2	2	2	2
dependent variable [action]	position	1	1	1	1	1	1	2	2	2	2	3	3	3

✗ redundant information

b. valid trend → position = 4.33·size·colour+0.67·(size·colour)<sup>2</sup> [mean error < 0.001%] ✓

applicability

a.- probability –

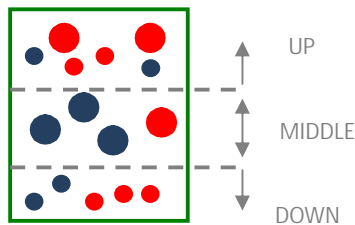
- lots of calculations in order to “fully understand” this behaviour
- any further “action” [e.g., middle-up] requires a new set of calculations
- no direct result / sequential comparison required
  - where is blue/small? → trial-and-error process
 

P(down) = 0%	}	→ UP
P(middle) = 0%		
- binary sequential process [is it here? Yes/no – and here? ...] ≠ behaviour understanding
  - small/blue up & down = non-understandable (no reason why this could happen (no one present within the given variables)) → probability considers this as logical as any other option

b.- trendingBot –

- straightforward understanding
- neutral to any complexity increase via “actions” [e.g., XYZ local information]
- direct result → where is blue/small? UP
- actual behaviour understanding
  - small/blue up & down = trend not found = no (understandable) behaviour

PROBLEM 2b – random behaviour



analysis

- big/blue circles follow an understandable behaviour → trendingBot is applicable
- rest of circles are randomly located → trendingBot concludes “trend not found”
  - probability can deal with this problem
  - probability will not make any difference between big/blue and the rest

trendingBot point of view (III)  
recommended proceeding

run trendingBot

- a. there is a valid trend => behaviour can be predicted  
[IF a set of minimum conditions is met (number of observations, max. error, etc.)]
- b. "trend not found"
  - b.1. look for further variables [e.g., density] removing the (apparent) randomness
  - b.2. further variables not found → perform a probabilistic analysis  
it can be stated that under the given circumstances randomness occurs