Which statistical model should you choose?

A guide to choose a statistical modeling tool according to the situation

The choice of a statistical model is not straightforward. It is erroneous to think that every data set has its own adapted model. If you are new to statistical modelling, this easy and short tutorial may be useful before exploring the following grid.

Every modelling tool answers specific questions. For example, glycaemia linked to a specific diabetes can be explained by a qualitative variable (sex for example). In this situation, the ANOVA model can be used. We may also use age data (quantitative variable) to see if there is a linear increasing or decreasing trend of glycaemia according to the age of the patients, using the same data. In this situation we would use linear regression.

The choice of a statistical model can also be guided by the shape of the relationships between the dependent and explanatory variables. A graphical exploration of these relationships may be very useful. Sometimes these shapes may be curved, so polynomial or nonlinear models may be more appropriate than linear ones.

The choice of a model can also be intimately tied to the very specific question you are investigating. For example, the estimation of the Vmax and Km parameters of the Michaelis-Menten enzyme kinetics implies the consideration of the specific Michaelis-Menten equation linking reaction rate (dependent variable) to substrate concentration (explanatory variable) in a nonlinear way.

If the purpose of the study is only to make predictions from a large set of variables, then solutions other than parametric models may be considered. The possibly correlated explanatory variables. The use of Partial Least Squares regression is very popular in chemometrics, where outputs are often predicted by a large spectrum of wavelengths.

What number of parameters should be included in the model?

Once you choose the appropriate modelling tool, in many situations you may ask how many parameters you should include in the model. The higher the number of parameters you include, the better the fit of the model to the data (i.e. the lower the residuals which implies a higher R^2 statistic). So should the number of parameters in the model be maximized in a way that residuals are extremely minimized? Not really. A model which fits the data too much will be too representative of the particular sample that is used, and the generalization to the whole population will be less accurate.

each other, the model with the lowest index has the best quality in the set. The interpretation of these indices does not make sense in an absolute context, in other words, when only one model is taken into consideration.

The grid

The grid below will help you choose a statistical model that may be appropriate to your situation (types and numbers of dependent and explanatory variables). The grid also includes a column with an example in each situation.

Conditions of validity of parametric models are listed in the paragraph following the grid.

The displayed solutions are the most commonly used tools in statistics. They are all available in XLSTAT. The list is not exhaustive. Many other solutions exist.

Dependent variable	Explanatory variable(s)	Example	Parametric models	Conditions of validity	Other solutions
One quantitative variable	One qualitative variable (= factor) with two levels	Effect of contamination (yes / no) on the concentration of a trace element in a plant	One-way ANOVA with two levels	1 ; 2 ; 3 ; 4	Mann-Whitney test
	One qualitative variable with k levels	Effect of the site (4 factories) on the concentration of a trace element in a plant	One-way ANOVA	1 ; 2 ; 3 ; 4	Kruskal-Wallis test
	Several qualitative variables with several levels	Combinatory effects of site (4 factories) and plant species on the concentration of a compound in plant tissue	Multi-way ANOVA (factorial designs)	1 ; 2 ; 3 ; 4	
	One quantitative variable	Effect of temperature on the concentration of a protein	Simple linear regression ; nonlinear models (depends on the	1 - 3	nonparametric regression (*);quantile regression; regression trees (*);

			shape of the relationship between the dependent / explanatory variable)		Random Forest(*)
	Several quantitative variables	Effect of the concentration of several contaminants on plant biomass	Multiple linear regression ; nonlinear models	1 - 6	PLS regression (*); Lasso; Ridge; Elastic Net
	Mixture of qualitative / quantitative variables	Combinatory effects of sex and age on glycaemia associated to a type of diabetes	ANCOVA	1 - 6	PLS regression (*); quantile regression; regression trees (*); Random Forest(*); Lasso; Ridge; Elastic Net
Several quantitative variables	Qualitative &/or quantitative variable(s)	Effect an environmental variables matrix on the transcriptome	MANOVA	1 ; 4 ; 7 ; 8	Redundancy analysis; PLS regression (*)
One qualitative variable	Qualitative &/or quantitative variable(s)	Dose effect on survival / death of mouse individuals	Logistic regression (binomial or ordinal or multinomial)	5 ; 6	PLS-DA (*); Discriminant Analysis (*); classification trees (*); classification Random Forest(*)
One count variable (with many zero's)	Qualitative &/or quantitative variable(s)	Dose effect on the number of necroses in mice	Log-linear regression (Poisson)	5 ; 6	

(*) solutions designed more for prediction

Conditions of validity

INSTALLATION & LICENSING

GETTING STARTED

TUTORIALS & GUIDES

Statistical Guides
Data management
Descriptive Statistics
Data Visualization
Exploratory Data Analysis
Modeling data
Hypothesis testing
Machine Learning
Sensory analysis
Marketing analysis
Conjoint Analysis
Text Mining
Decision Aid
Time Series Analysis
Monte Carlo Simulations
Power Analysis
Statistical Process Control
Design of Experiments
Survival analysis
Lab data analysis
Multiblock data analysis
Path modeling
XLSTAT AI
R in Excel

Conditions of validity

1. Individuals are independent.
2. Variance is homogeneous.
3. Residuals follow a normal distribution.
4. At least 20 individuals (recommended).
5. Absence of multicollinearity (if the purpose is to estimate model parameters).
6. No more explanatory variables than individuals.
7. Multivariate normality of residuals.
8. Variance is homogeneous within every dependent variable. Correlations across dependent variables are homogeneous.

INCLUDED IN

TABLE OF CONTENTS

Similar articles

Free Case Studies and White Papers
How to interpret goodness of fit statistic...
Webinar XLSTAT: Sensory data analysis ...
What is statistical modeling?
Comparison of Supervised Machine Lear...
Which clustering method should you cho...

Help Center

Try XLSTAT now!

INSTALLATION & LICENSING

GETTING STARTED

TUTORIALS & GUIDES