Topic #9: Regression

Regression analysis models the relationship between one or more response variables (also called dependent variables, explained variables, predicted variables, or regressands) (usually named Y), and the predictors (also called independent variables, explanatory variables, control variables, or regressors,) usually named X1,...,Xp). Multivariate regression describes models that have more than one response variable.

Types of regression

Simple linear regression and multiple linear regression are related statistical methods for modeling the relationship between two or more random variables using a linear equation. Simple linear regression refers to a regression on two variables while multiple regression refers to a regression on more than two variables. Linear regression assumes the best estimate of the response is a linear function of some parameters (though not necessarily linear on the predictors).

Nonlinear regression models

If the relationship between the variables being analyzed is not linear in parameters, a number of nonlinear regression techniques may be used to obtain a more accurate regression.

Linear models

Predictor variables may be defined quantitatively (i.e., continuous) or qualitatively (i.e., categorical). Categorical predictors are sometimes called factors. Although the method of estimating the model is the same for each case, different situations are sometimes known by different names for historical reasons:

- If the predictors are all quantitative, we speak of multiple regression.

- If the predictors are all qualitative, one performs analysis of variance.
- If some predictors are quantitative and some qualitative, one performs an analysis of covariance.

The linear model usually assumes that the dependent variable is continuous. If least squares estimation is used, then if it is assumed that the error component is normally distributed, the model is fully parametric. If it is not assumed that the data are normally distributed, the model is semi-parametric. If the data are not normally distributed, there are often better approaches to fitting than least squares. In particular, if the data contain outliers, robust regression might be preferred.

If two or more independent variables are correlated, we say that the variables are multicollinear. Multicollinearity results in parameter estimates that are unbiased and consistent, but inefficient.

If the regression error is not normally distributed but is assumed to come from an exponential family, generalized linear models should be used. For example, if the response variable can take only binary values (for example, a Boolean or Yes/No variable), logistic regression is preferred. The outcome of this type of regression is a function which describes how the probability of a given event (e.g. probability of getting "yes") varies with the predictors.

Regression and Bayesian statistics

Maximum likelihood is one method of estimating the parameters of a regression model, which behaves well for large samples. However, for small amounts of data, the estimates can have high variance or bias. Bayesian methods can also be used to estimate regression models. A prior is placed over the parameters, which incorporates everything known about the parameters. (For example, if one parameter is known to be non-negative, a non-negative distribution can be assigned to it.) A posterior distribution is then obtained for the parameter vector. Bayesian methods have the advantages that they use all the information that is available. They are exact, not asymptotic, and thus work well for small data sets if some contextual information is available to be used in the prior. Some practitioners use maximum a posteriori (MAP) methods, a simpler method than full Bayesian analysis, in which the parameters are chosen that maximize the posterior mode. MAP methods are related to Occam's Razor: there is a preference for simplicity among a family of regression models (curves) just as there is a preference for simplicity among competing theories.