Regression

NPR

August 4, 2017

1 Introduction

In statistical modeling, regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors').

2 From Wikipedia

In statistical modeling, regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors'). More specifically, regression analysis helps one understand how the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed. Most commonly, regression analysis estimates the conditional expectation of the dependent variable given the independent variables that is, the average value of the dependent variable when the independent variables are fixed. Less commonly, the focus is on a quantile, or other location parameter of the conditional distribution of the dependent variable given the independent variables. In all cases, the estimation target is a function of the independent variables called the regression function. In regression analysis, it is also of interest to characterize the variation of the dependent variable around the regression function which can be described by a probability distribution. A related but distinct approach is necessary condition analysis[1] (NCA), which estimates the maximum (rather than average) value of the dependent variable for a given value of the independent variable (ceiling line rather than central line) in order to identify what value of the independent variable is necessary but not sufficient for a given value of the dependent variable.

Regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. Regression analysis is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships. In restricted circumstances, regression analysis can be used to infer causal relationships between the independent and dependent variables. However this can lead to illusions or false relationships, so caution is advisable;[2] for example, correlation does not imply causation.

Many techniques for carrying out regression analysis have been developed. Familiar methods such as linear regression and ordinary least squares regression are parametric, in that the regression function is defined in terms of a finite number of unknown parameters that are estimated from the data. Nonparametric regression refers to techniques that allow the regression function to lie in a specified set of functions, which may be infinite-dimensional.

The performance of regression analysis methods in practice depends on the form of the data generating process, and how it relates to the regression approach being used. Since the true form of the data-generating process is generally not known, regression analysis often depends to some extent on making assumptions about this process. These assumptions are sometimes testable if a sufficient quantity of data is available. Regression models for prediction are often useful even when the assumptions are moderately violated, although they may not perform optimally. However, in many applications, especially with small effects or questions of causality based on observational data, regression methods can give misleading results.[3][4]

In a narrower sense, regression may refer specifically to the estimation of continuous response variables, as opposed to the discrete response variables used in classification.[5] The case of a continuous output variable may be more specifically referred to as metric regression to distinguish it from related problems.[6]

2.1 Models

Linear regression; Simple regression; Ordinary least squares; Polynomial regression; General linear model; Generalized linear model; Discrete choice; Logistic regression; Multinomial logit; Mixed logit; Probit; Multinomial probit; Ordered logit; Ordered probit; Poisson; Multilevel model; Fixed effects; Random effects; Mixed model; Nonlinear regression; Nonparametric Semiparametric Robust Quantile Isotonic Principal components Least angle Local Segmented Errors-in-variables

2.2 Estimation

Least squares Ordinary least squares Linear (math) Partial Total Generalized Weighted Non-linear Non-negative Iteratively reweighted Ridge regression Least absolute deviations Bayesian Bayesian multivariate

2.3 Background

Regression model validation Mean and predicted response Errors and residuals Goodness of fit Studentized residual GaussMarkov theorem

3 Regression and Data Science

https://medium.com/towards-data-science/simple-and-multiple-linear-regression-in-python-c928425168f9