A parametric framework for multidimensional linear regression

While Ordinary Least Squares regression establishes fundamental concepts in data analysis, it requires the independent variable to be error-free and the error term to be constant. Dr Stanley Luck, statistics consultant and founding member of Vector Analytics LLC in the US, has developed an innovative parametric framework for multidimensional linear regression. This provides a more general framework for establishing measurement error regression, even if data for both variables is subject to error.

Dr Stanley Luck, statistics consultant and founding member of Vector Analytics LLC, Delaware, was motivated to investigate the applied algebraic foundations of data analysis after working on a collaborative research and development project involving the identification of beneficial agronomic variation in maize. This project involved the application of genome-wide association studies (GWAS) and expression quantitative trait loci (eQTL) methods to identify genetic variants. After performing high-dimensional searches of the data, Luck observed that the results from the classification and regression tree (CART) analyses did not correspond well with the GWAS.

Luck uncovered extensive research literature, and his applied algebraic investigation of the merits of various effect size measures and their associated statistical methodologies has already been recorded in two recent journal publications. In this third phase of his research into the foundations of data analysis, he investigated the issue of fitting a multidimensional line to data that are subject to stochastic error (ie, a random effect that may result in an outcome that is not expected, even though both the model and parameters are correct). This led to his developing a novel parametric framework for multidimensional linear regression.

**PARAMETRIC VS CARTESIAN REPRESENTATION**

A curve can be defined using a Cartesian equation, an equation in terms of x and y only. Alternatively, a parametric equation can be used where both x and y are functions of a third variable (usually t). Luck demonstrates how employing a parametric representation, rather than a Cartesian equation, enabled him to obtain a more general framework for linear regression that also takes the experimental error in all variables into account. Using the chain rule, he transformed the ordinary linear regression method to the parametric representation \((x(t), y(t))\), with t corresponding to an element of a convex set (a convex set is made up of points so that the line joining any two points in the set lies entirely within that set, so the set is connected).

**MEASUREMENT ERROR**

Measurement error refers to a sub-discipline of statistics supported by extensive literature and a long history. Luck relates how the wide range of opinions about both the statistical framework and methodology of measurement error suggest that the standard textbook treatment of linear regression may be incomplete. Moreover, the confusion surrounding the fundamental role of measurement error models and Weighted Least Squares optimisation in partitioning the effects of errors contributes to the problem of irreproducibility in data analysis.

**LINEAR REGRESSION**

Ordinary Least Squares regression is a common statistical technique for modelling a two-dimensional linear relationship between an independent variable, x, and a dependent variable, y. It produces the straight line that minimises the sum of the squares (the least squares) of the difference between the observed and predicted values.

Luck explains that while Ordinary Least Squares regression serves a definitive role in establishing fundamental concepts in data analysis, it requires the independent variable to be error-free and the variance of the residual, or error term, to be constant, or homoscedastic.

If the Ordinary Least Squares assumption of constant variance in the errors is violated, the Weighted Least Squares method can be used. This is an extension of Ordinary Least Squares regression where non-negative weights are applied to the data points. The error-free condition, however, is a requirement of the Moore-Penrose inverse algorithm that is used to estimate the parameters of the Weighted Least Squares regression model. Furthermore, if the independent variable is subject to error, the Ordinary Least Squares regression estimate for the slope is reduced, causing the attenuation of the Pearson correlation coefficient that measures the strength of the linear relationship between two variables. This has spurred Luck’s longstanding research effort to develop a more general framework for establishing measurement error linear regression framework is bivariate because it is based on the Cartesian representation \(y = f(x)\), since the dependent variable y is explained by the independent variable x.

Applying the chain rule to this linear relationship led Luck to discover a novel parametric framework for linear regression.

**LINEAR REGRESSION**

The chain rule. The chain rule is used to differentiate a function of a function, or compound function of the form \(f(g(x))\). He also notes that the standard linear regression framework is bivariate because it is based on the Cartesian representation \(y = f(x)\), since the dependent variable y is explained by the independent variable x.
Behind the Research

Dr Stanley Luck

Research Objectives

Dr Luck developed a novel parametric framework for multidimensional linear regression.

Detail

Address
Vector Analytics LLC
3911 Concord Plz #7361
Wilmington, DE 19803
USA

Bio
Dr Stanley Luck is a statistics consultant and founding member of Vector Analytics LLC. He develops algorithms for effect size statistics. Formerly, he worked in Genetic Discovery research at E I DuPont de Nemours, Inc. and collaborated in the application of GWAS and eQTL methods for identifying beneficial agronomic traits for maize.

Collaborators
Dr Luck thanks many former colleagues in the Genetic Discovery Group at DuPont for helpful discussions about parametric framework for multidimensional linear regression.

References


Personal Response

What do you envisage to be the next stage in your research into the foundations of data analysis?

The next stage of my research involves the development of data analysis methods for complex systems, such as nursing homes, organisms, and economic systems. I expect that accounting for the degrees of freedom, unbalanced sample sizes, and measurement error is necessary for obtaining reproducible results in data analysis for such systems. However, there is an additional mathematical complication arising from the fact that the performance of such systems is determined by complex interactions between many components. Furthermore, there will be alternative ways to optimise performance; there is not a unique well-defined solution. Instead, the objective is to explore the space of solutions using data analysis methods such as CART, to obtain functional information that helps in the development of engineering models for predicting improved performance. The specification of cost-benefit trade-offs between different forms of variation is required and the criteria for substantive significance for effect size will vary depending on the particular application. There is no ‘one effect size fits all’ approach in data analysis for complex systems.