Geodesic least squares regression with applications to astrophysical data

G. Verdoolaege

Department of Applied Physics, Ghent University, B-9000 Ghent, Belgium
Laboratory for Plasma Physics, Royal Military Academy, B-1000 Brussels, Belgium

Abstract

Regression analysis on astrophysical data is a relatively challenging activity and it has long been recognised that many of the assumptions of ordinary least squares regression (GLS) are not valid in several applications in the field. Accordingly, various techniques from the domain of frequentist statistics and Bayesian probability theory have been proposed to address the shortcomings of GLS. We here present a new regression method called "geodesic least squares regression" (GLS), which has recently been developed and applied in the field of magnetic confinement fusion (MCF). The main difference with standard techniques is that, for the case of a single response variable, the distribution of the response conditioned on a prior marginalized by the regression model (marginal distribution) may be different from the distribution conditional on an actual measurement (observed distribution). Then, instead of minimising the difference between observed and predicted values of the variable, which is the goal of the standard GLS, we aim at minimizing the distance between the modelled and observed distribution. To achieve this, the model is fitted to the measured data set, but now the model is equipped with the Fisher information metric, the method can handle errors in all variables, is robust against data outliers and the uncertainty in the regression model, and can be used with arbitrary distribution models and regression functions. After introducing GLS and demonstrating its advantages on a synthetic data set, we show results of fitting MCF scaling laws as well as the baryonic Tully-Fisher relation in astronomy.

Motivation

- In many areas of science, regression analysis is used:
  - As an aid to build and validate theoretical models from data and to find parametric dependencies
  - As a statistical tool to formulate scaling laws for the purpose of extrapolation

- Ordinary least squares regression (OLS) is the workhorse
- Often, multiple assumptions underlying OLS are not fulfilled
- There may be various reasons:
  - Considerable measurement uncertainty: statistical and systematic
  - Uncertainty on response (dependent), y and predictor (independent, x) variables
  - Model uncertainty: linear, power law, semi-empirical,
  - Heterogeneous data and error bars, correlations, non-Gaussian probability distributions
  - Atypical observations (outliers)
  - Near-collinearity of predictor variables
  - Data transformations, e.g.
    \( \log y = \log b_0 + b_1 \log x_1 + \ldots + b_n \log x_n \)
  - Inferior regression analysis counters other efforts!
- A flexible, robust and user-friendly regression tool is needed

Geodesic least squares regression (GLS)

- OLS
  - Formulate model with parameters \( \hat{\beta} = [\beta_0, \ldots, \beta_d]^T \)
  - Take n measurements \( y_i, x_i \) and \( \hat{\beta} = [\hat{\beta}_0, \ldots, \hat{\beta}_d]^T \)
- GLS
  - Least squares minimization:
    \[ \sum_{i=1}^{n} || y_i - \hat{y}_i ||^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \beta_1 x_i)^2 \]
  - GLS
    - Modulated distribution (e.g. Gaussian likelihood)
    - Modified least square minimization:
      \[ \sum_{i=1}^{n} || y_i - \hat{y}_i ||^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \beta_1 x_i)^2 \]

Information geometry

- Geometric approach to probability theory [3]
- A family of probability density functions (PDFs) forms a metric space, or manifold
- Fisher information is the metric tensor
- Rao geodesic distance (GD) is the shortest distance between points (PDFs)

- Example: Gaussian manifold
  - \( p_N(x | \mu, \sigma) = \frac{1}{\sqrt{2\pi} \sigma} \exp \left[ -\frac{(x - \mu)^2}{2 \sigma^2} \right] \)
  - GD squared distance:
    \[ d_G^2 = \frac{1}{2} \int_0^\infty \left[ \frac{\partial^2}{\partial \mu^2} \log p_N(x | \mu, \sigma) \right]^2 dx \]

- The pseudospace (or Fréchet) is a model for the manifold of univariate Gaussian distribution, respecting true geometry (d in red, \( d_G \) in blue)

Scalings in magnetic confinement fusion

H-mode power threshold

- Controlled nuclear fusion: clean, safe, limitless energy
- Magnetic confinement fusion: tokomaks (ITER), stellarators, ...
- High confinement mode (H-mode): threshold \( P_{TH} \) on input power
- Scaling with classical power law:
  \[ P_{TH} = P_b h \]
- Power law is an average plasma density (10^19 m^-3)
- \( s \) plasma surface area (m^2)
- ITPA H-mode threshold database (4): 645 measurements from 7 tokomaks
- Logarithmic variables assumed Gaussian: single standard deviation
- Relative error from database

5. Synthetic data: outliers

- Linear regression: \( \hat{\beta}_0 = \beta_0 + \beta_1 r + \beta_2 x + \beta_3 s \)
- Artificial data set (a)
  - Gaussian noise: 40% on \( r \), 10% on \( r \), 3% on \( s \), 15% on \( P_{TH} \)

2. Synthetic data: logarithmic space

- Logarithmic transformation:
  \[ \eta = \log \hat{\beta}_0 + \log \hat{\beta}_1 + \log \hat{\beta}_2 + \log \hat{\beta}_3 + \log S \]
- Gaussian noise: 20% on \( r \), 5% on \( r \), 15% on \( s \), 15% on \( P_{TH} \)

3. Real data: comparison of loglinear with nonlinear

Baryonic Tully-Fisher relation in astronomy

-Relation between rotational velocity and baryonic mass of galaxies
- Various purposes:
  - Distance indicator
  - Constraints on galaxy formation models
- Test for alternatives to ΛCDM (e.g. MONO) via slope and scatter

- Data from McGaugh [5] (gas-rich galaxies)
- Compare with least trimmed squares (LTS) [6,7]

Conclusion

- Geodesic least squares regression is flexible and robust
- GLS is simple but powerful due to strong mathematical foundations
- GLS offers unified solution to various regression problems
- Probability distributions more informative for regression
- Future development: more accurate error bars on GLS estimates and predictions
- GLS will be implemented in a public software package

References