Robust analysis of trends in noisy tokamak confinement data using geodesic least squares regression[a)]

G. Verdoolaege,[1,2,b)] A. Shabbir,[1,3] and G. Hornung[1]
[1)] *Department of Applied Physics, Ghent University, B-9000 Ghent, Belgium*
[2)] *Laboratory for Plasma Physics, Royal Military Academy, B-1000 Brussels, Belgium*
[3)] *Max Planck Institute for Plasma Physics, Boltzmannstr. 2, 85748 Garching, Germany*

(Dated: 4 July 2016)

Regression analysis is a very common activity in fusion science for unveiling trends and parametric dependencies, but it can be a difficult matter. We have recently developed the method of geodesic least squares regression (GLS) that is able to handle errors in all variables, is robust against data outliers and uncertainty in the regression model, and can be used with arbitrary distribution models and regression functions. We here report on first results of application of GLS to estimation of the multi-machine scaling law for the energy confinement time in tokamaks, demonstrating improved consistency of the GLS results compared to standard least squares.

I. INTRODUCTION

Fitting a trend line or a nonlinear relation to a data set is extremely common in almost all areas of science, and fusion is no exception. In fusion research, applications include capturing parametric dependencies of a variable of interest, extrapolation towards new operational regimes or future machines, estimating the parameters of a theoretical model from a data set, as well as various diagnostic applications such as calibration. However, in the vast majority of cases, practitioners use ordinary least squares regression (OLS) to perform the fit. While OLS has the advantage of simplicity and availability in any software package for scientific analysis, it should be stressed that OLS relies on a number of assumptions that are often not fulfilled in real-world data.[1] Geodesic least squares regression (GLS) is a new method that is also simple and fast, yet is much more general than OLS and many other common regression techniques. In this paper, after briefly sketching the working of GLS, we apply the method to estimate the classic scaling law for the energy confinement time in tokamaks.[2] We obtain results that are consistent, whether the relation is fitted in logarithmic space or as a nonlinear power law, in contrast to OLS.

II. REGRESSION METHODOLOGY

The idea behind OLS regression for a single response variable y is to estimate the parameters β_k $(k = 0, \ldots, p)$ of the regression model by minimizing the difference between, on the one hand, the prediction of the values of y, given n measurements x_{ij} of the m predictor variables x_j, and, on the other hand, the actually observed values y_i $(i = 1, \ldots, n,\ j = 1, \ldots, m)$. However, this only takes into account the statistical error on y, whereas the x_j may have non-negligible uncertainty as well. Furthermore, the errors may be different from one measurement to another (e.g. when they originate from different diagnostics or devices). A way around this is to consider the more general maximum likelihood method (ML), which maximizes the probability distribution of the response variable conditional on the predictor variables. For the remainder of the paper we will assume normally distributed uncertainties, reducing ML to the following optimization problem:

$$\{\hat{\beta}_k\} = \arg\max_{\{\beta_k\}} p_{\mathrm{m}},$$

$$p_{\mathrm{m}} \equiv \frac{1}{\sqrt{2\pi}\sigma_{\mathrm{m}}} \exp\left\{ -\sum_{i=1}^{n} \frac{\left[y_i - f\left(\{x_{ij}\}, \{\beta_k\}\right) \right]^2}{2\sigma_{\mathrm{m}}^2} \right\}.$$

Here, f is the regression function (possibly nonlinear), while the measurements y_i are assumed to be mutually independent, and similar for the x_{ij}. The standard deviation σ_{m} in general describes uncertainty on the response *and* the predictor variables. We refer to σ_{m} as the standard deviation of the *modeled distribution* p_{m}, since it depends on the regression model: the uncertainty on the x_j propagates through f.

There is one flaw in this reasoning, which is shared by most regression methods, including many of the more sophisticated. It assumes that σ_{m} is indeed the correct standard deviation on the y_i, leaving no room for unforseen sources of uncertainty. Still, such additional uncertainties often occur, e.g. due to outliers in the data, plasma fluctuations or transients, uncertainty in the regression model, etc. The GLS regression method accommodates these situations by considering, apart from

[a)] Contributed paper published as part of the Proceedings of the 21st Topical Conference on High-Temperature Plasma Diagnostics, Madison, Wisconsin, June, 2016.

[b)] geert.verdoolaege@ugent.be

p_m, another distribution for the dependent variable that makes as few assumptions about the data as possible. We call this the *observed distribution* p_o, and here we will assume only that it is a normal distribution centered on each measurement y_i, with some unknown standard deviation σ_o that is to be estimated from the data. As such, every measurement y_i is actually treated as a probability distribution and GLS aims to minimize the overall difference between the modeled and observed distributions, just like OLS minimizes the overall difference between the modeled and observed values of y. As a distance measure between probability distributions we choose the *geodesic distance* (GD) rooted in information geometry, which is a geometric approach to probability theory.[1] Hence, the $p+2$ parameters $\beta_0, \ldots, \beta_p, \sigma_\mathrm{o}$ are estimated via the following optimization problem:[1]

$$\{\hat{\beta}_k,\, \hat{\sigma}_\mathrm{o}\} = \underset{\{\beta_k, \sigma_\mathrm{o}\}}{\arg\min}\, \mathrm{GD}^2 \left[\prod_{i=1}^{n} p_\mathrm{o}(y|y_i, \sigma_\mathrm{o}), \prod_{i=1}^{n} p_\mathrm{m}(y|\{x_{ij}\}, \{\beta_k\}) \right].$$

It has been demonstrated that, despite its simplicity, GLS consistently outperforms several other regression methods in various challenging regression tasks.[1]

III. CONFINEMENT SCALING

We next apply GLS to the classic multi-machine scaling law for the confinement time τ_E (s) for the ELMy H-mode in tokamaks, in terms of engineering variables:

$$\tau_\mathrm{E} = \beta_0\, I_\mathrm{p}^{\beta_I}\, B_\mathrm{t}^{\beta_B}\, \bar{n}_\mathrm{e}^{\beta_n}\, P_\mathrm{l}^{\beta_P}\, R^{\beta_R}\, \kappa^{\beta_\kappa}\, \epsilon^{\beta_\epsilon}\, M_\mathrm{eff}^{\beta_M}.$$

Here, I_p is the plasma current (MA), B_t the vacuum toroidal magnetic field (T), $\bar{n}_\mathrm{e}$ the central line-averaged electron density (10^{19} m^{-3}), P_l the loss power (MW), R the plasma major radius (m), ϵ the inverse aspect ratio, κ the elongation and M_eff the effective atomic mass. The version of this scaling law that is currently mostly quoted is IPB98(y,2),[2], the coefficients of which were estimated using OLS regression on a logarithmic scale, as shown in Table I. It should be noted that, in estimating this scaling law, the coefficient for the B_t scaling was fixed at 0.15, in accordance with observations at individual machines.[3]

We now invoke GLS and compare it with the classic OLS approach using a more recent version of the confinement database, namely version DB3 13f,[4], limited to the standard set. This consists of 1310 entries from 9 machines, together with (simple) error estimates for each of the variables. We use these relative errors to derive the standard deviations that are required in the GLS method. The error bar on a specific quantity, in particular the measured confinement time, can be different from one machine to another. Therefore, for each machine we need a parameter representing the observed standard deviation.

TABLE I. Estimates of the regression coefficients and ITER predictions ($\hat{\tau}_\mathrm{E}$) in log-linear scaling of the energy confinement time (constrained: OLSc, GLSc; unconstrained: OLSu, GLSu).

	β_0	β_I	β_B	β_n	β_P	β_R	β_κ	β_ϵ	β_M	$\hat{\tau}_\mathrm{E}$ (s)
IPB98	0.056	0.93	0.15	0.41	-0.69	1.97	0.78	0.58	0.19	4.9
OLSc	0.053	0.88	0.15	0.45	-0.66	2.12	0.22	0.43	0.20	3.3
GLSc	0.053	0.84	0.15	0.48	-0.74	2.26	0.28	0.58	0.37	3.1
OLSu	0.049	0.78	0.32	0.44	-0.67	2.23	0.38	0.58	0.19	4.2
GLSu	0.048	0.65	0.43	0.49	-0.76	2.52	0.62	0.86	0.28	4.0

We first impose the same constraint $\beta_B \equiv 0.15$, but then we also derive the unconstrained coefficients. Furthermore, we start the analysis in the logarithmic domain, where a linear regression model is usually imposed on the data. The results are given in Table I, together with the predicted confinement time in ITER under the conditions $I_\mathrm{p} = 15$ MA, $B_\mathrm{t} = 5.3$ T, $\bar{n}_\mathrm{e} = 10.3 \times 10^{19}$ m^{-3}, $P_l = 87$ MW, $R = 6.2$ m, $\kappa = 1.7$, $\epsilon = 0.32$, $M = 2.5$. Uncertainty estimates on the coefficients will be reported elsewhere, in a future, more comprehensive study. As expected, the dependence on B_t in the unconstrained analysis is somewhat stronger compared to the constrained analysis. It was also observed before that the constraint has an influence on the coefficient of I_p.[3] In addition, the coefficients of the geometrical quantities R, κ and ϵ turn out to be relatively difficult to estimate, depending both on the data set and analysis method. When considering the unconstrained estimates only, the geometrical dependencies estimated by GLS are considerably stronger than those given by OLS. The dependence on magnetic field, loss power and effective mass is also somewhat stronger with GLS.

Although conducting the analysis of a power law in the logarithmic domain by means of linear regression is a very common and convenient procedure in many areas of science, it is not necessarily an optimal strategy. When the data are duely considered as samples from an underlying probability distribution, one notices that the operation of taking the logarithm leads to a skewing of the distribution. In general it is safer to work in the original data space, hence performing nonlinear regression on the power law. OLS is perfectly capable of estimating the exponents in the power law, but its results may differ from those emerging from the loglinear analysis. Indeed, Table II shows the estimates by OLS and GLS using the original data without logarithmic transformation, both in the constrained and unconstrained case. When comparing with the results from linear regression, several parameters estimated by OLS have changed considerably, particularly those for the geometrical quantities. In contrast, the parameters estimated by GLS are similar, whether derived from the logarithmic or the original data and the predictions for ITER are the same.

TABLE II. Estimates of the regression coefficients and ITER predictions ($\hat{\tau}_E$) in nonlinear scaling of the energy confinement time (constrained: OLSc, GLSc; unconstrained: OLSu, GLSu).

	β_0	β_I	β_B	β_n	β_P	β_R	β_κ	β_ϵ	β_M	$\hat{\tau}_E$ (s)
OLSc	0.065	0.88	0.15	0.43	-0.80	2.33	0.73	0.52	-0.16	2.6
GLSc	0.055	0.81	0.15	0.46	-0.73	2.32	0.31	0.62	0.35	3.1
OLSu	0.058	0.68	0.49	0.47	-0.82	2.57	1.00	0.84	-0.25	3.5
GLSu	0.049	0.59	0.49	0.48	-0.74	2.57	0.70	0.96	0.25	4.0

TABLE III. Estimates by nonlinear GLS of the modeled and observed standard deviations on the power threshold, expressed as percentage errors on τ_E, for the various tokamaks in the database.

	ASDEX	AUG	C-Mod	DIII-D	JET
σ_m (%)	29	21	24	20	22
σ_o (%)	32	25	26	25	26

	JFT-2M	JT-60U	PBXM	PDX
σ_m (%)	26	29	26	37
σ_o (%)	28	57	29	43

The reason for the more consistent performance of GLS across different representations of the data lies in the flexibility embodied within the σ_o parameters. By increasing its estimate of the observed standard deviations for the various machines, GLS can accommodate deviations from the proposed regression model (e.g. outliers or secondary lobes of data points), without compromising its estimate of the actual trend in the data. A similar flexibility is not shared by OLS. A detailed study of the features in the data causing this behavior is beyond the scope of this paper, but we can get an idea of which machines contribute data far from the main regression surface by comparing the estimates of σ_o for each tokamak with σ_m averaged over the measurements coming from that machine. This comparison for the unconstrained nonlinear analysis is shown in Table III (the log-linear case is similar). One notices that for all machines the observed standard deviation is higher than the modeled one, indicating the presence of additional sources of uncertainty not taken into account by the regression model. For some machines this can become as high as 30% or even more, but JT-60U is essentially the only device for which there is a major discrepancy between the modeled and observed standard deviations. Hence, it would be worthwhile to investigate the data from this machine in more detail.

Finally, Figure 1 shows histograms for the predicted confinement time for ITER, obtained by performing the nonlinear OLS and GLS regression analysis 100 times, each time generating a random permutation of the database with repetition ('bootstrapping' in statistics terminology). The histograms are relatively well separated, with GLS predicting an average confinement time of 4 seconds, while OLS predicts only 3.5 seconds.

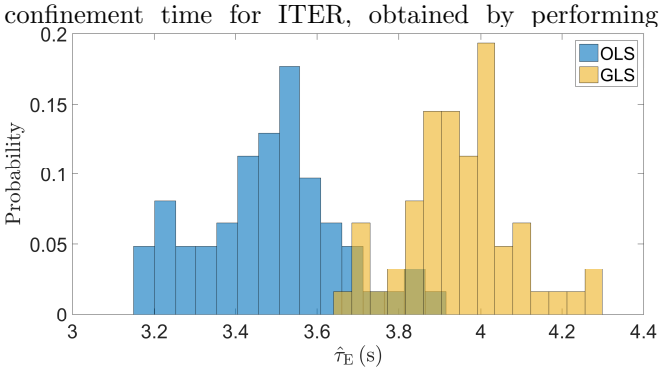

FIG. 1. Histograms of the confinement time predicted by nonlinear OLS and GLS on 100 permutations of the database with repetition.

IV. CONCLUSION

We have applied the method of geodesic least squares regression (GLS) to estimate the classic scaling law for the energy confinement in tokamaks, using the multi-machine database. In comparison with ordinary least squares, GLS performs more consistently, yielding similar results on the logarithmic and linear scales. GLS can also be used to indicate subsets of the data that deviate from the main trend. In future work, GLS will be used for continued analysis of the energy confinement scaling in tokamaks, as well as for various other regression analyses in fusion science.

ACKNOWLEDGMENTS

The authors wish to acknowledge the ITPA Topical Group on Transport and Confinement for maintaining and kindly providing the data in the H-mode Confinement database.

[1] G. Verdoolaege and J.-M. Noterdaeme, Nucl. Fusion **55**, 113019 (2015).
[2] ITER Physics Basis, Nucl. Fusion **39**, 2175 (1999).
[3] J. Christiansen *et al.*, Nucl. Fusion **32**, 291 (1992).
[4] D. McDonald *et al.*, Nucl. Fusion **47**, 147 (2007).