Journal of Statistical Planning and Inference 168 (2016) 90–96

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Statistical calibration and exact one-sided simultaneous tolerance intervals for polynomial regression

Yang Han [a,b], Wei Liu [a,*], Frank Bretz [c,d], Fang Wan [a], Ping Yang [e]

[a] S3RI and School of Mathematics, University of Southampton, Southampton SO17 1BJ, UK
[b] University of Exeter Medical School, Exeter, EX1 2LU, UK
[c] Novartis Pharma AG, Basel, 4002, Switzerland
[d] Shanghai University of Finance and Economics, Shanghai, 200433, China
[e] Department of Statistics, The Chinese University of Hong Kong, Hong Kong

ARTICLE INFO

Article history:
Received 9 April 2015
Received in revised form 20 July 2015
Accepted 21 July 2015
Available online 29 July 2015

Keywords:
Confidence level
Linear regression
Polynomial regression
Quantile line
Simultaneous confidence band
Simultaneous tolerance intervals
Statistical simulation

ABSTRACT

Statistical calibration using linear regression is a useful statistical tool having many applications. Calibration for infinitely many future y-values requires the construction of simultaneous tolerance intervals (STI's). As calibration often involves only two variables x and y and polynomial regression is probably the most frequently used model for relating y with x, construction of STI's for polynomial regression plays a key role in statistical calibration for infinitely many future y-values. The only exact STI's published in the statistical literature are provided by Mee et al. (1991) and Odeh and Mee (1990). But they are for a multiple linear regression model, in which the covariates are assumed to have no functional relationships. When applied to polynomial regression, the resultant STI's are conservative. In this paper, one-sided exact STI's have been constructed for a polynomial regression model over any given interval. The available computer program allows the exact methods developed in this paper to be implemented easily. Real examples are given for illustration.

1. Introduction

Statistical calibration using linear regression has a rich history going back to Eisenhart (1939). The problem involves a quantity of interest x which is expensive or difficult to measure, a surrogate quantity y which is cheaper or easy to measure, and the assumption that y and x are related by a linear regression model. For example, x is the true concentration of radon, $^{222}R_n$, while y is the concentration reading on an alpha track detector (ATD), at a place, or x is the true alcohol level in blood stream while y is the reading on a breathalyzer, of a driver. In order to use an observed y to infer the corresponding but unobserved x, a calibration experiment is carried out to measure y_{0i} corresponding to a known x_{0i} for $i = 1, \ldots, n$. A regression model of y on x is then fitted by using the training data $\mathcal{E} = \{(x_{0i}, y_{0i}),\ i = 1, \ldots, n\}$ and used to infer the x-values corresponding to infinitely many y-values to be observed in future. The inference for the x-value corresponding to one single future y-value is considered by Eisenhart (1939), Brown (1982) and Smith and Corbett (1987) among others, and the relevant literature is reviewed in Osborne (1991) and Brown (1993).

This paper focuses on inference for infinitely many future y-values. Specifically, a confidence set $C(y_x)$ for the unknown x corresponding to each observed future y_x is constructed and the infinite sequence of confidence sets $C(y_x)$ corresponding

* Corresponding author.
E-mail address: w.liu@maths.soton.ac.uk (W. Liu).

http://dx.doi.org/10.1016/j.jspi.2015.07.005
0378-3758/

to an infinite sequence of observed future y_x-values has the property: *with confidence level* γ, *with respect to the randomness in the training data* $\mathcal{E}$, *that the proportion of confidence sets* $C(y_x)$ *containing the corresponding true* x*-values is at least* β, *where* $0 < \gamma, \beta < 1$ *are pre-specified constants*. This *property* can be expressed as

$$P_{\mathcal{E}}\left\{\liminf_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N} I_{\{x_i\in C(y_{x_i})\}} \geq \beta\right\} \geq \gamma \tag{1}$$

where I_A denotes the indicator function of the set A and hence $\frac{1}{N}\sum_{i=1}^{N} I_{\{x_i\in C(y_{x_i})\}}$ is the proportion of the N confidence sets that contain the true x-values. It is argued by Lieberman et al. (1967), Scheffé (1973), Aitchison (1982), Mee et al. (1991), Mee and Eberhardt (1996), Mathew and Zha (1997), Mathew et al. (1998) and Krishnamoorthy and Mathew (2009, Chapter 3) among others that this *property* is highly desirable in many applications, and overwhelming majority publications on infinite many calibrations aim to guarantee this *property*. Other properties that may have useful applications are discussed in Mee and Eberhardt (1996).

One standard way to construct the confidence sets $C(y_x)$ having this *property* is to use the (β, γ)-simultaneous tolerance intervals (STI's). Assume *a priori* that the unknown x-values corresponding to all the future y_x's are in a given interval $[a, b]$. For example, the true blood alcohol level of any driver cannot be lower than $a = 0$ or higher than some upper threshold b. The (β, γ)-STI's $[L(x; \mathcal{E}),\ U(x; \mathcal{E})]$ over the interval $x \in [a, b]$ satisfy

$$P_{\mathcal{E}}\left\{P_{y_x}\left\{L(x; \mathcal{E}) < y_x < U(x; \mathcal{E})|\mathcal{E}, x\right\} \geq \beta \text{ for all } x \in [a, b]\right\} \geq \gamma \tag{2}$$

where y_x denotes a future y-value corresponding to x and is independent of the training data $\mathcal{E}$, the probability P_{y_x} is with respect to y_x and conditional on $\mathcal{E}$, and the probability $P_{\mathcal{E}}$ is with respect to $\mathcal{E}$. Then for each future y_x the confidence set $C(y_x)$ for the corresponding x is defined as

$$C(y_x) = \left\{x \in [a, b] : L(x; \mathcal{E}) \leq y_x \leq U(x; \mathcal{E})\right\}. \tag{3}$$

It is shown in Scheffé (1973, Appendix B) that these confidence sets $C(y_x)$ have the *property* in (1).

Numerical results in Mee and Eberhardt (1996) and Lee (1999) lead to the conjecture that the property in (1) is guaranteed by using the pointwise tolerance intervals (PTI's) instead of the STI's in (3). We have constructed counter examples to show that the property in (1) does not hold in general if the STI's are replaced by the PTI's in the construction of $C(y_x)$ in (3). The counter examples are not given here to save space but available from the authors. Hence the STI's are central to the construction of $C(y_x)$'s in order to guarantee the property in (1).

Construction of (β, γ)-STI's is considered first by Lieberman and Miller (1963) for simultaneous predictions, and there are three construction methods available in the literature. The first is the construction of *central* (β, γ)-STI's by Lieberman and Miller (1963), Lieberman et al. (1967) and Scheffé (1973). Note that the central (β, γ)-STI's are two-sided and even the exact central (β, γ)-STI's are conservative as (β, γ)-STI's. The aforementioned papers only provide conservative central (β, γ)-STI's. The second is the probability set methodby Wilson (1967) and Limam and Thomas (1988). Similar to the confidence set construction method of Rao (1973, p. 473), this method hinges on a γ level confidence set for the unknown parameters of the regression model and the resultant (β, γ)-STI's are also conservative and two-sided. The third is an exact method by Mee et al. (1991) for two-sided (β, γ)-STI's and Odeh and Mee (1990) for one-sided (β, γ)-STI's. Since the first two methods are conservative while Mee et al.'s (1991) method is exact, the two-sided (β, γ)-STI's of Mee et al. (1991) are usually narrower and so better than the conservative (β, γ)-STI's, as demonstrated numerically in Mee et al. (1991, Section 3).

In this paper, we focus on polynomial regression. A calibration problem often involves only two quantities y and x (or their suitable transformations), and a polynomial regression model is a simple yet probably the most frequently used model to relate two quantities. Exact one-sided (β, γ)-STI's will be constructed for a polynomial model of any order $p - 1$ over any given covariate interval $x \in [a, b]$. While the construction method of this paper is also applicable to other regression models, such as the fractional polynomials (cf. Royston and Altman, 1994), the key step of maximizing $K(\boldsymbol{x})$ in the expression (9) below may require a different optimization method depending on the specific form of the regression model considered.

The upper (β, γ)-STI's have $L(x; \mathcal{E}) = -\infty$ in (2), and the lower (β, γ)-STI's have $U(x; \mathcal{E}) = \infty$ in (2). The confidence set $C(y_x)$ corresponding to the upper STI's often takes the form of a lower confidence limit, which is most relevant for the example of blood alcohol level since the police want to catch those drivers whose blood alcohol levels are above the legal limit by using the lower confidence limits (cf. Krishnamoorthy et al., 2001). The confidence set $C(y_x)$ corresponding to the lower STI's often takes the form of an upper confidence limit, which is most relevant for the example of ATD since the company wants to monitor that the radon concentrations are not above the safety threshold set by government agency by using the upper confidence limits.

Note that the exact (β, γ)-STI's of Mee et al. (1991) and Odeh and Mee (1990) are for a multiple linear regression model, in which the covariates are assumed to have no functional relationships, over a special covariate region only. These STI's become conservative when applied to a polynomial regression model of order two (i.e. quadratic regression) or above. Even for the simple linear regression (i.e. polynomial regression of order one), these STI's are only over a covariate set that is symmetric about $\bar{x}$, the mean of the observed covariate values in $\mathcal{E}$. See Section 2.2 for more details.

The layout of this paper is as follows. Section 2 deals with the construction of exact one-sided (β, γ)-STI's for a polynomial regression model over a given covariate interval. It also shows why the exact one-sided (β, γ)-STI's for a multiple linear