Computational Statistics and Data Analysis 81 (2015) 1–9

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

# Simultaneous confidence bands for a percentile line in linear regression

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

[a] S3RI and School of Mathematics, University of Southampton, Southampton SO17 1BJ, UK
[b] Novartis Pharma AG, Basel, 4002, Switzerland
[c] Shanghai University of Finance and Economics, Shanghai, 200433, China

ARTICLE INFO

*Article history:*
Received 8 September 2013
Received in revised form 17 July 2014
Accepted 18 July 2014
Available online 23 July 2014

*Keywords:*
Confidence level
Linear regression
Percentile line
Simultaneous confidence band
Statistical simulation

ABSTRACT

Construction of simultaneous confidence bands for a percentile line in linear regression has been considered by several authors. But only conservative symmetric bands, which use critical constants over the whole covariate range $(-\infty, \infty)$, are available in the literature. New methods allow the construction of exact symmetric bands for a percentile line over a finite interval of the covariate $x$. The exact symmetric bands can be substantially narrower than the corresponding conservative symmetric bands. Several exact symmetric confidence bands are compared under the average band width criterion. Furthermore, new asymmetric confidence bands for a percentile line are proposed. They are uniformly and can be very substantially narrower than the corresponding exact symmetric bands. Therefore, asymmetric bands should always be used under the average band width criterion. The proposed methods are illustrated with a real example.



## 1. Introduction

The linear regression model $\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{e}$ with $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ is the basis for many different inference problems. Construction of simultaneous confidence bands for a percentile line, also known as a quantile line,

$$\boldsymbol{x}'\beta + z_\gamma\sigma, \tag{1}$$

where $z_\gamma$ is the $100\gamma$th percentile of the standard normal distribution, i.e., $\Phi(z_\gamma) = \int_{-\infty}^{z_\gamma} \phi(\omega)d\omega = \gamma$ with $\phi(\omega) = \exp(-\omega^2/2)/\sqrt{2\pi}$, has been considered by many researchers including Steinhorst and Bowden (1971), Turner and Bowden (1977, 1979) and Thomas and Thomas (1986). Recently, several articles have studied various applications of confidence bands for the mean regression line $\boldsymbol{x}'\beta$ which is a special case of the percentile line $\boldsymbol{x}'\beta + z_\gamma\sigma$ with $\gamma = 0.5$; see, for example, Spurrier (1999), Al-Saidy et al. (2003), Liu et al. (2004, 2007, 2009) and Piegorsch et al. (2005). A simultaneous confidence band can quantify the plausible range of the function of interest. It can also be used for statistical discrimination as considered by Easterling (1969). In this paper, the focus is on the percentile line $\boldsymbol{x}'\beta + z_\gamma\sigma$ for $\gamma \neq 0.5$. In some practical problems, for example, drug stability studies, the percentile function may be of more interest than the regression function $\boldsymbol{x}'\beta$. Drug stability studies are routinely carried out in the pharmaceutical industry in order to measure the degradation over time of an active pharmaceutical ingredient of a drug product. From the patients' point of view, it is expected that a large

* Corresponding author.
*E-mail addresses:* hanyang1122@gmail.com, y.han@soton.ac.uk (Y. Han), w.liu@soton.ac.uk (W. Liu).

http://dx.doi.org/10.1016/j.csda.2014.07.009
0167-9473/

proportion (e.g., $100(1-\gamma)\%$ with $\gamma = 0.05$) of dosage units (e.g., tablets, capsules, vials) should have the drug content level above a certain threshold, say, 98 (in percentage) before a specified expiry date, say, 2 years. It is thus of interest to estimate where the $100\gamma$th percentile line lies with a two-sided simultaneous confidence band. Comparing the lower part of the confidence band with the threshold 98, we can assess whether no more than 5% of all the dosage units have the drug content level below 98, for any given point in the time interval $(0, 2)$. Similarly, comparing the upper part of the confidence band with the threshold 98, we can evaluate whether more than 5% of all the dosage units have the drug content level below 98, for any time point in $(0, 2)$. More details of this example are given in Section 4. Extensive discussions on the usefulness of percentile points or percentile lines can be found in Harris and Boyd (1995), Gilchrist (2000), Koenker (2005) and Liu et al. (2013).

Without loss of generality, assume that all the covariates are mean-centered and the design matrix $\boldsymbol{X}$ is of full column rank and so $\boldsymbol{X}'\boldsymbol{X}$ is non-singular. The least squares estimator of $\beta$ is $\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \sim N_{p+1}(\beta, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$ where $p+1$ is the dimension of $\beta$. The usual unbiased estimator of $\sigma^2$ is $\hat{\sigma}^2 \sim \sigma^2\chi^2_\nu/\nu$ with $\nu = n-p-1$. It is well known that the random variables $\hat{\beta}$ and $\hat{\sigma}^2$ are independent. As in several published papers on this topic, we focus here on the simple linear regression model which has $p = 1$ covariate $x$. Our approach, however, can readily be generalized to polynomial regression and multiple linear regression where the covariates are assumed to have no functional relationships among them. More details are given in Section 2.

We first consider the two-sided symmetric confidence bands of the form

$$\boldsymbol{x}'\beta + z_\gamma\sigma \in \boldsymbol{x}'\hat{\beta} + z_\gamma\hat{\sigma}/\theta \pm c\hat{\sigma}\sqrt{\boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x} + (z_\gamma)^2\xi} \quad \text{for all } x \in (a, b), \tag{2}$$

where $\boldsymbol{x} = (1, x)'$, $(a, b)$ is a given covariate interval over which a confidence band is required, and the given constants $\theta \neq 0$ and $\xi$ are chosen to give different specific confidence bands. At a given $x$, the center of the band is $\boldsymbol{x}'\hat{\beta} + z_\gamma\hat{\sigma}/\theta$ while the width of the band is $2c\hat{\sigma}\sqrt{\boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x} + (z_\gamma)^2\xi}$. Hence the center of the band depends on $\theta$, while the width of the band depends on $\xi$. The central question is how to compute the critical constant $c$ to give the specified confidence level $1-\alpha$. All the published bands (e.g. Steinhorst and Bowden (1971), Turner and Bowden (1977, 1979)) and Thomas and Thomas (1986)) are of form (2) with a particular pair of $\theta$ and $\xi$ values. However, methods available in the literature so far are only for computing the conservative critical constants $c$ over the whole covariate range $(a, b) = (-\infty, \infty)$. As linear regression models often hold for only a finite covariate range, it is important to consider a finite interval $(a, b)$. A confidence band over a finite interval can be substantially narrower and so more efficient than a band over the whole range $(-\infty, \infty)$. Our methods allow the computation of the exact critical constant $c$ of the general form (2) for any given interval $(a, b)$, finite or infinite. We also compare the average band widths over an interval $(a, b)$ of several bands of form (2).

Various choices of $(\theta, \xi)$ in (2) have been studied in the literature in the hope of reducing the average width of a band. In this paper, we propose new asymmetric confidence bands which are uniformly narrower than the corresponding symmetric bands. Corresponding to each symmetric band of form (2), an asymmetric band will be constructed whose width is smaller than the width of the symmetric band at any $x \in (a, b)$ and can be substantially smaller especially when $\gamma$ is either close to zero or one.

It has to be emphasized that the construction of exact $1-\alpha$ simultaneous confidence bands (either symmetric or asymmetric) for the percentile line of the standard linear regression models is the focus of this paper. The standard linear regression model assumption (including normality) is crucial to the particular form (1) of the percentile line. As soon as one goes beyond the standard linear regression models, only approximate simultaneous confidence bands for a percentile line can be constructed. For example, one can use the large sample asymptotic normality of quantile regression (cf. Koenker, 2005) to construct only an approximate simultaneous band for a percentile curve.

The layout of the paper is as follows. Section 2 considers several symmetric bands of form (2) and their comparison under the average band width criterion. Section 3 shows how to construct an asymmetric band corresponding to any symmetric band and compares their average widths. Section 4 provides an illustrative example of application of confidence bands for a percentile line. Section 5 contains conclusions and discussions.

## 2. Two-sided symmetric confidence bands

Two-sided symmetric simultaneous confidence bands of form (2) can be divided into Type I and Type II according to whether $\xi = 0$ or $\xi \neq 0$, respectively. Table 1 presents the six two-sided symmetric simultaneous confidence bands (with the abbreviation name and origin), three of each type, which are studied in this section. It is noteworthy that all these bands are exact over a finite covariate interval $(a, b)$ and therefore new. The origins given in the table simply indicate the papers in which the corresponding conservative bands over the whole covariate range $x \in (-\infty, \infty)$ were investigated.

Note that when $\xi = 0$, in the general form of confidence bands (2) we have

$$\boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x} + (z_\gamma)^2\xi = \boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x} = \mathrm{Var}(\boldsymbol{x}'\hat{\beta})/\sigma^2.$$

Band *TBU* and band *UV* with $\theta = \sqrt{\frac{2}{\nu}}\Gamma(\frac{\nu+1}{2})/\Gamma(\frac{\nu}{2})$ satisfy $E(\hat{\sigma}/\theta) = \sigma$ and therefore $\boldsymbol{x}'\hat{\beta} + z_\gamma\hat{\sigma}/\theta$ is the uniformly minimum variance unbiased estimator (UMVUE) of $\boldsymbol{x}'\beta + z_\gamma\sigma$. Bands *V* and *UV* use the $\xi$ values chosen to satisfy $\mathrm{Var}(\boldsymbol{x}'\hat{\beta} + z_\gamma\hat{\sigma}/\theta) = \left(\boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x} + (z_\gamma)^2\xi\right)\sigma^2$ for $\theta = 1$ and $\sqrt{\frac{2}{\nu}}\Gamma(\frac{\nu+1}{2})/\Gamma(\frac{\nu}{2})$ respectively.