

Empirical Likelihood-Based Inference in Linear Errors-in-Covariables Models with Validation Data

J.N.K. Rao

Carleton University, School of Mathematics & Statistics

1125 Colonel By Drive

Ottawa, Canada

jrao@math.carleton.ca

Qihua Wang

Academy of Mathematics & System Science,

Chinese Academy of Science

Beijing 100080, P.R. China

qhwang@math4.amt.ac.cn

1. Introduction

Let Y be a response variable and X be a d -vector of explanatory variables. We consider the standard linear regression model $Y = X^T \mathbf{b} + e$, where \mathbf{b} is a d -vector of regression parameters, e is a random error, and given X , the errors e are i.i.d. Owen (1991) developed empirical likelihood (EL) confidence regions for \mathbf{b} when X is measured exactly. The main purpose of this paper is to extend Owen's work when the covariables are subject to measurement error. We assume a surrogate \tilde{X} is observed instead of X , but also assume the availability of independent validation data to relate X and \tilde{X} . In particular, we assume that an independent validation data $\left\{ \left(X_i, \tilde{X}_i \right)_{N+1}^{N+n} \right\}$

is available in addition to primary data $\left\{ \left(Y_i, \tilde{X}_i \right)_{i=1}^N \right\}$, where n is small relative to N .

We rewrite the model as $Y = u^T(\tilde{X})\mathbf{b} + \mathbf{h}$ with $u(\tilde{X}) = E(X|\tilde{X})$ and $\mathbf{h} = e + X^T \mathbf{b} - u^T(\tilde{X})\mathbf{b}$.

This motivates us to introduce the auxiliary random vector $Z_i(\mathbf{b}) = u(\tilde{X}_i) \left\{ Y_i - u^T(\tilde{X}_i)\mathbf{b} \right\}$ with $E Z_i(\mathbf{b}) = 0$, $i = 1, \dots, N$. Empirical log-likelihood evaluated at \mathbf{b} is then defined by

$$l_N(\mathbf{b}) = -2 \max_{\sum p_i Z_i(\mathbf{b}) = 0, \sum p_i = 1} \sum_{i=1}^N \log(N p_i), \quad (1)$$

where p_1, \dots, p_N are nonnegative numbers. If \mathbf{b} is the true parameter, then $l_N(\mathbf{b}) \xrightarrow{L} \mathbf{c}_d^2$ as $N \rightarrow \infty$,

but this result cannot be used to make inference on \mathbf{b} because $l_N(\mathbf{b})$ contains the unknown $u(\tilde{X})$.

We use the validation data to estimate $u(\tilde{x})$ by $\hat{u}_n(\tilde{x})$, using distribution free kernel density methods. Let $Z_{in}(\mathbf{b})$ denote $Z_i(\mathbf{b})$ with $u(\tilde{X}_i)$ replaced by $\hat{u}_n(\tilde{X}_i)$. We now obtain an estimated

empirical log-likelihood $\hat{l}(\mathbf{b})$ by replacing $Z_i(\mathbf{b})$ by $Z_{in}(\mathbf{b})$ in (1).

We prove that $\hat{l}(\mathbf{b})$ is distributed asymptotically as a weighted sum of independent standard \mathbf{c}_1^2 variables with unknown weights. By estimating the unknown weights consistently, we construct an estimated EL confidence region on \mathbf{b} . We also propose an adjusted empirical log-likelihood and prove that $\hat{l}_{ad}(\mathbf{b}) \xrightarrow{L} \mathbf{c}_d^2$. To avoid estimating the unknown weights or the adjustment factor, we propose a smoothed bootstrap EL to construct an asymptotically correct confidence region on \mathbf{b} . We conduct a simulation study to compare the proposed methods with a normal approximation based method in terms of coverage probability and average of the confidence intervals. The smoothed bootstrap EL and the adjusted EL performed better than the other methods.

REFERENCE

Owen, A. (1991). Empirical likelihood for linear models. *Annals of Statistics*, 19, 1725-1747.

RESUMÉ

Les modèles linéaires avec erreurs de mesure dans les covariables sont considérés, en supposant que les covariables validées indépendamment sont disponibles en plus de données sur la variable dépendante et des estimés approximatifs pour les covariables. Nous développons d'abord un estimé empirique du logarithme de la fonction de vraisemblance grâce aux données validées et nous montrons que sa distribution asymptotique est celle d'une somme pondérée de variables aléatoires, chacune de loi \mathbf{c}_1^2 . En utilisant des estimateurs convergents des poids, nous construisons des intervalles de confiance pour un paramètre de régression \mathbf{b} basés sur la méthode d'estimation de la fonction empirique. De plus, nous proposons d'ajuster l'estimé empirique du logarithme de la fonction de vraisemblance et montrons qu'elle suit asymptotiquement une loi \mathbf{c}^2 . Les deux méthodes mènent à des régions de confiance asymptotiquement correctes. Pour éviter d'estimer les poids inconnus ou les facteurs d'ajustement, nous proposons un estimé empirique du logarithme de la fonction de vraisemblance basé sur un bootstrap partiellement lisse de manière à construire des intervalles de confiance asymptotiquement correctes pour \mathbf{b} . Nous menons une étude de simulation afin de comparer les méthodes proposées à une méthode basée sur l'approximation normale, en terme de probabilité de couverture et longueur moyenne des intervalles de confiance (\mathbf{b} scalaire).