Statistical Methodology 16 (2014) 32–46

Contents lists available at ScienceDirect

Statistical Methodology

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

Efficient estimation in a regression model with missing responses

Scott D. Crawford *

University of Wyoming, 3332 UW Laramie, WY 82071, United States

ARTICLE INFO

Article history:
Received 23 May 2012
Received in revised form
21 June 2013
Accepted 1 July 2013

Keywords:
Efficiency
Missing at random
Full imputation

ABSTRACT

This article examines methods to efficiently estimate the mean response in a linear model with an unknown error distribution under the assumption that the responses are missing at random. We show how the asymptotic variance is affected by the estimator of the regression parameter, and by the imputation method. To estimate the regression parameter, the ordinary least squares is efficient only if the error distribution happens to be normal. If the errors are not normal, then we propose a one step improvement estimator or a maximum empirical likelihood estimator to efficiently estimate the parameter.

To investigate the imputation's impact on the estimation of the mean response, we compare the listwise deletion method and the propensity score method (which do not use imputation at all), and two imputation methods. We demonstrate that listwise deletion and the propensity score method are inefficient. Partial imputation, where only the missing responses are imputed, is compared to full imputation, where both missing and non-missing responses are imputed. Our results reveal that, in general, full imputation is better than partial imputation. However, when the regression parameter is estimated very poorly, the partial imputation will outperform full imputation. The efficient estimator for the mean response is the full imputation estimator that utilizes an efficient estimator of the parameter.

1. Introduction

This work examines methods to efficiently estimate the mean response in a semi-parametric model under the assumption that the responses are missing at random. A study by Elliot [4] illustrates the

* Tel.: +1 307 766 3341.
E-mail addresses: crawford@stat.tamu.edu, scrawfo8@uwyo.edu.

1572-3127/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.stamet.2013.07.001

complexity of such a problem. He investigated the link between specific minority groups (e.g., non-Mexican Hispanic Americans or Chinese Americans) and obesity in children. The response variable, weight of the child, was frequently missing and laws restricting personal information made it impossible to recover the missing data. Because the missing structure was correlated with other covariates in the model (e.g. height of the child and location), the results would have contained bias without imputation, i.e. without the estimation of the missing values.

The book by Little and Rubin [7] is well known for its explanation on the estimation of regression parameters under the assumption of data missing at random. Schick [15] explains how efficient estimators are formed for regression models when no distributional assumptions are made on the covariates. Müller et al. [9] propose the method of full imputation, which estimates all the responses, as an improvement over partial imputation, where only the missing responses are imputed. Müller [8] showed that the efficient estimation of the response requires an efficient estimation of the regression parameters.

We begin by investigating efficient estimation of the regression parameter when the error distribution is unknown. The ordinary least-squares method proves efficient when the error distribution happens to be normal. The complete case versions of the one step improvement estimator discussed in Forrester et al. [5] and the maximum empirical likelihood estimator discussed in Peng and Schick [12] are presented as efficient estimators regardless of the error distribution. Simulations disclose the mean square error of these estimators under various distributions.

To estimate the mean response with missing data we compare four common methods: the listwise deletion, propensity score method, partial imputation, and full imputation. The asymptotic variances for each method are derived and simulations display the MSE of the estimation of the mean response under various error distributions. We demonstrate how the MSE is affected by the method of imputation, and by the estimator of the regression parameter.

This research illustrates the imputation method's impact on estimation in regression models with missing data. Full imputation exhibits the least asymptotic variance when the parameter is estimated efficiently. With an inefficient estimate of the parameter we see that full imputation can contain more asymptotic variance than partial imputation. When the missing structure is not symmetric about the covariate, the listwise deletion methods will be biased. We find some non-regular errors where the OLS estimator for the regression parameter performs better than efficient estimators. The simulations reveal these estimator's variability for finite sample sizes.

The paper is organized into five sections. After the introduction in Section 1, Section 2 investigates the efficient estimation of the regression parameter. Section 3 shows the asymptotic variance for different estimation methods for the mean response with missing data. In Section 4 we compare the asymptotic variance of the partially imputed estimator to the fully imputed estimator. In Section 5 we show the asymptotic variances for all four imputation methods under various scenarios. Our conclusions are in Section 6.

2. Parameter estimation in linear regression

2.1. The model

We look at the linear regression model

$$Y = \vartheta^\top X + \varepsilon,$$

where ϑ is the vector of unknown regression coefficients, the covariate vector X and the error variable ε are independent with unknown distributions, and the error ε has mean zero, finite variance σ^2, and a density f with the finite Fisher information for location. The latter means that f is absolutely continuous and $\mathbb{I} = \int l^2(y)f(y)dy$ is finite, where $l = -f'/f$ is the score function for location. We allow the response Y to be missing. Then the observed variables are $(\delta, X, \delta Y)$, where δ is zero if the response Y is not observed and 1 if Y is observed. We assume that the responses are *missing at random*, which means

$$P(\delta = 1|X, Y) = P(\delta = 1|X) = E(\delta|X),$$