行政院國家科學委員會專題研究計畫成果報告

Estimation in log-linear model with measurement errors and without extra information

計畫編號: NSC 90-2118-M-032-008
執行期限: 90年8月01日至91年7月31號
主持人: 黃逸輝 淡江大學數學系
計畫參與人員: 林宜河 淡江大學數學系

中文摘要

在分析變數含有誤差的資料時, 一般會假設有一些額外的訊息已知, 方能使模式是可辨認或是可進一步分析的, 然而在某些情形中, 例如沒有重複測量或沒有其他的相關變數可觀測時, 額外訊息就難以取得, 此時能分析此資料的統計方法並不多, 而必須訴求於蒐集其他的資料或資訊. 在本篇報告中我們提出一個估計方法 適用於對數線性模式, 並能在有測量誤差且無額外訊息的情形下進行估計, 此估計方式並不需假設 未知自變數的分佈. 本文提供了一個可分析的函數型測量誤差模式, 這在有關測量誤差模式的文獻中 是少見的.

Abstract

In analysis of an errors-in-variables model, it is conventional to assume some extra information are available in order to make the model identifiable or the analysis easier. However, it can happen that such information are hard to access when there are no any replications or instrumental variable are available. In such cases, littie statistical methods are applicable, and one is told to colloect more information for further analysis. In this report, we develop an estimation method for a log-linear model with measurement errors and without any extra information. It doesn't require any distribution assumption on the mismeasured covariate. It is also applicable when the mismeasured covariate is unknown and fixed constant. In a short, a functional measurement error model can be analyzed without extra information, which is uncommon in analyses of errors-in-variables models.

Key words: ERRORS-IN-VARIABLES; MEASUREMENT ERROR; LOG-LINEAR; IDENTIFICATION

1 Motivation and Object

When measurement errors are present in a regression analysis, it is well known that the naive estimators are usually inconsistent and hence not satisfactory. In order to make a better inference, one usually needs some extra information like additional variables are observed or some parameters' values are known. However, knowing this variance may not always plausible in practices, especally when there are no replicate measurements or other instrumental variables can be observed.

In this report, we analyzed the log-linear model with normal measurement errors. The anslysis needs little assumption and hence is more widely applicable than the analysis provided by Nakamura (1990), which requires the knowledge of error's variance.

2 Estimation without extra information

Consider a simple log-linear model with two variables Y and X, and assume the following relation holds

$$Y_i \mid X_i \overset{i.i.d.}{\sim} \text{Poisson}(\lambda_i),$$

$$\lambda_i = exp(\beta_0 + \beta_1 X_i),\ i = 1, \cdots, n.$$

If the X_i can be observed, the equations derived from setting score functions to zeros are

$$\sum_1^n Y_i = \sum_1^n e^{(\beta_0+\beta_1 X_i)},$$

$$\sum_1^n Y_i X_i = \sum_1^n e^{(\beta_0+\beta_1 X_i)} X_i.$$

In stead of observing the true covariate X_i, we assume that only surrogate W_i of X_i are observed, where $W_i = X_i + \delta_i$, and $\delta_i \sim N(0, \sigma_\delta^2)$. In practices, the X_i may be random variables or unknown constants. We assume the later, i.e. we have a functional case. The case when X_i are random variables can be included when inference is conditional on the X_i.

When measurement errors are present, the X_i are unknown like other parameters, and the likelihood approach is not feasible. Nakamura (1990) had proposed the estimating equations

$$\frac{1}{n}\sum_{i=1}^n Y_i - \frac{1}{n} e^{-\frac{1}{2}\sigma_\delta^2\beta_1^2} \sum_{i=1}^n e^{\beta_0+\beta_1 W_i} = 0$$

$$\frac{1}{n}\sum_{i=1}^n Y_i W_i - \frac{1}{n} e^{-\frac{1}{2}\sigma_\delta^2\beta_1^2} \sum_{i=1}^n (W_i - \beta_1\sigma_\delta^2) e^{\beta_0+\beta_1 W_i} = 0,$$

to solve for estimates of β_0 and β_1 when σ_δ^2 is known. Under normality assumption of δ, these equations are 0-unbiased and the resultant estimators are consistent.

However, when σ_δ^2 is unknown, the above two equations are not enough for determining the estimates of β_0 and β_1. As a natural extension from these two equations, it can be shown that

$$\frac{1}{n}\sum_{i=1}^n Y_i W_i^2 - \frac{1}{n} e^{-\frac{1}{2}\sigma_\delta^2\beta_1^2} \sum_{i=1}^n (W_i - \beta_1\sigma_\delta^2)^2 e^{\beta_0+\beta_1 W_i}$$

is also a 0-unbiased statistic. Thus we have three equations for three unkown β_0, β_1 and σ_δ^2. In summary, we propose solving

$$\frac{1}{n}\sum_{i=1}^n Y_i - \frac{1}{n} e^{-\frac{1}{2}\sigma_\delta^2\beta_1^2} \sum_{i=1}^n e^{\beta_0+\beta_1 W_i} = 0 \quad (2.1)$$

$$\frac{1}{n}\sum_{i=1}^n Y_i W_i - \frac{1}{n} e^{-\frac{1}{2}\sigma_\delta^2\beta_1^2} \sum_{i=1}^n (W_i - \beta_1\sigma_\delta^2) e^{\beta_0+\beta_1 W_i} = 0 \quad (2.2)$$

$$\frac{1}{n}\sum_{i=1}^n Y_i W_i^2 - \frac{1}{n} e^{-\frac{1}{2}\sigma_\delta^2\beta_1^2} \sum_{i=1}^n (W_i - \beta_1\sigma_\delta^2)^2 e^{\beta_0+\beta_1 W_i} = 0, \quad (2.3)$$

for estimates of β_0, β_1 and σ_δ^2.

In order to show these estimates are consistent and asymptotic normal distributed. Three conditions are imposed to the unobserved X_i.

i). Let $M_n(s) = \frac{1}{n}\sum_{i=1}^n e^{sX_i}$. We assume there is a function $M(s)$ such that $M_n(s) \to M(s)$, $\forall s \in R$, and $\frac{d^k M_n(s)}{ds^k} (= \frac{1}{n}\sum X_i^k e^{sX_i}) \to \frac{d^k M(s)}{ds^k}$ for $k = 1, \ldots, 6$, as $n \to \infty$.

ii). $lim_{n\to\infty} \sum_{i=1}^n \frac{e^{\beta_1 X_i}}{i^2} < \infty$,
$lim_{n\to\infty} \sum_{i=1}^n \frac{X_i^2 e^{\beta_1 X_i}}{i^2} < \infty$,
$lim_{n\to\infty} \sum_{i=1}^n \frac{X_i^4 e^{\beta_1 X_i}}{i^2} < \infty$.

iii). $3M(\beta_1)M'(\beta_1)M''(\beta_1) - 2[M'(\beta_1)]^3 - [M(\beta_1)]^2 M^{(3)}(\beta_1) \neq 0$.

With these conditions, the folowing theorem, which is our main result, is shown to be valid.

Theorem. Suppose that the the conditions i), ii) and iii) are hold, then with probability converges to 1, there is a solution $\hat{\theta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}_\delta^2)'$ of (2.1), (2.2) and (2.3) such that $\hat{\theta} \overset{a.s.}{\to} \theta$, where $\theta = (\beta_0, \beta_1, \sigma_\delta^2)'$.

It is also true that $\sqrt{n}(\hat{\theta} - \theta) \overset{d.}{\to} N(0, H)$, where $H = A^{-1}(\theta)B(\theta)A^{-t}(\theta)$, and

$$A(\theta) = n^{-1}\sum_{i=1}^n \mathrm{E}\left[\frac{\partial}{\partial\theta^t}\phi_i(\theta)\right]$$

$$B(\theta) = n^{-1}\sum_{i=1}^n \mathrm{Cov}\left\{\phi_i(\theta)\right\}, \text{ where}$$

$$\phi_i(\theta) = \begin{pmatrix} \phi_{1i}(\theta) \\ \phi_{2i}(\theta) \\ \phi_{3i}(\theta) \end{pmatrix}$$

$$= \begin{pmatrix} Y_i - e^{\beta_0+\beta_1 W_i - \frac{1}{2}\sigma_\delta^2\beta_1^2} \\ Y_i W_i - (W_i - \beta_1\sigma_\delta^2)e^{\beta_0+\beta_1 W_i - \frac{1}{2}\sigma_\delta^2\beta_1^2} \\ Y_i W_i^2 - (W_i - \beta_1\sigma_\delta^2)^2 e^{\beta_0+\beta_1 W_i - \frac{1}{2}\sigma_\delta^2\beta_1^2} \end{pmatrix}$$

3 Conclusion and Discussion

We have developed a method that can estimate the regression parameters in log-linear model, with measurement error and without any extra information. It is applicable when the unobserved covariates are random variables or fixed constants. The method can extend to the case when the regression function is a multiple one easily, as long as the measurement errors are assumed to be normal distributed. According to some simulation not shown here, the proposed estimates works fine in finite samples.

Some furture work can be motivated from this report. It is uncommon that a functional measurement error model can be analysed without any extra information. In some other problems, the model can be unidentifiable, and the estimation may needs extra information and/or the distribution assumption on the unobserved covariates. Why the log-linear is different from them ? Is there any other model that can be analysed without extra information ? If we can find the answers, then may be we can develop some new estimation method that needs less assumption than the conventional ones, and hence the new estimation methods will be more widely applicable.

References

[1] Nakamura, T. (1990). Corrected Scores Function for Errors-in-Variables Models: Methodology and Application to Generalized Linear Models.*Biometrika*, 77, 127-137.