行政院國家科學委員會專題研究計畫　成果報告

測量誤差模式在無額外訊息下的估計方法研究

計畫類別：個別型計畫
計畫編號：NSC93-2118-M-032-006-
執行期間：93 年 08 月 01 日至 94 年 07 月 31 日
執行單位：淡江大學數學系

計畫主持人：黃逸輝

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 94 年 10 月 31 日

Estimation in generalized linear model with measurement error and without extra information

Y. H. HUANG

Department of Mathematics, Tamkang University, Taipei, Taiwan

email: huang@math.tku.edu.tw

Summary: Extra information besides the regression model is required by traditional estimation methods for analyzing a measurement error problem. However, by over-parameterization, we are able to construct an additional estimation equation to release extra information requirement. We demonstrate that it is possible to estimate regression parameter consistently without any extra information in some regression model. Furthermore, if approximately consistent estimators are tolerable, than the estimation without extra information is applicable for the quasi-likelihood/variance model which is more general than generalized linear model.

Key words: extra information; over-parameterization; measurement error, generalized linear model; quasi-likelihood/variance model

1 Introduction

Measurement error is a common problem for regression analyses in practice. It is well known that ignore the measurement error will cause biases in parameters' estimations, make the model unidentifiable (Carroll, Ruppert and Stefanski, 1995) or distort the true coverage probability of a confidence interval (Gleser and Hwang, 1987). To mitigate these effects, there are many analyses that account for the measurement error. However, most methods that account for the measurement error usually require extra information about the measurement error's model, say, variance of the error, reliability data or validation data. Thus, it seems that one has no choice but the naive estimator if there is no extra information. But, in this paper, we provide another choice in addition to the naive estimations. The method developed here is called the over-parameterization that can generate an addition estimating function. This method can be easily applied on a quasi-likelihood/variance model and provide consistent or approximately consistent estimators even when there is no extra information.

Section 2 illustrates the basic idea of over-parameterization and the model's notations that we used. Section 3 applies the approach to the linear and probit models, and section 4 demonstrate how to approximate the score function when no corrected score can be found. A brief conclusion is given at the final section.

2 The model and the idea of over-parameterization

Consider the quasi-likelihood/variance model with response variable Y and linear predictor $\beta_0 + \beta_1 X$, and

$$E(Y \mid X) = f_m(\beta_0 + \beta_1 X),\ Var(Y \mid X) = \delta f_v(\beta_0 + \beta_1 X, \theta),$$

where (β_0, β_1) is our primary interest and θ being the nuisance parameter. Conventionally, when (Y_i, X_i) are available one can use the quasi-score function

$$S(Y, X : \beta_0, \beta_1) = \sum_{i=1}^{n} \frac{1}{f_v(\beta_0 + \beta_1 X_i, \theta)} [Y_i - f_m(\beta_0 + \beta_1 X_i)] f_m'(\beta_0 + \beta_1 X_i) \begin{pmatrix} 1 \\ X_i \end{pmatrix}$$

as the estimating function. However, if there are measurement error in measuring X, one can't observe X but its surrogate W. We assume that $W_i = X_i + \delta U_i$ where δU_i is the error in measuring X_i with $\mathrm{E}(U_i) = 0$ and $Var(U_i) = 1$. If δ is known or there are replicate measurements W_{ij} for X_i, than there are many methods are applicable. For example the conditional score approach by Stefanski and Carroll (1987), quasi-likelihood/variance approach by Stefanski and Carroll (1990) or the corrected score by Nakamura (1990). The corrected score approach is related to our estimating method. The corrected score looks for substitutes $S^*(Y, W : \beta_0, \beta_1, \delta^2)$ of $S(Y, X : \beta_0, \beta_1)$ which has conditional expectation $E(S^*(Y, W : \beta_0, \beta_1, \delta^2) \mid Y, X) = S(Y, X : \beta_0, \beta_1)$ and thus $E(E[S^*(Y, W : \beta_0, \beta_1, \delta^2) \mid Y, X]) = ES(Y, X : \beta_0, \beta_1) = (0, 0)'$.

When there are no extra information, usually $S^*(Y, W : \beta_0, \beta_1, \delta^2)$ is not enough to determine estimates of $\beta' s$. Furthermore, the model can be unidentifiable under some restriction. However, as long as the model is not identifiable, it is possible to estimate the $\beta' s$ and δ^2 without extra any information. To do this, we introduce the method called over-parameterization.

The most convenient way to over-parameterize a regression model is to extend the linear predictor to a higher order one, say, extend $E(Y_i \mid X_i) = \beta_0 + \beta_1 X_i$ to $E(Y_i \mid X_i) = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$ with $\beta_2 = 0$. Thus we are lead to the extended quasi-score estimation function

$$S_E(Y, X : \beta_0, \beta_1, \beta_2) = \sum_{i=1}^{n} \frac{1}{f_v(\beta_0 + \beta_1 X_i, \theta)} [Y_i - f_m(\beta_0 + \beta_1 X_i)] f_m'(\beta_0 + \beta_1 X_i) \begin{pmatrix} 1 \\ X_i \\ X_i^2 \end{pmatrix}.$$

When X_i is measured with error and δ^2 is known, than it is possible to find a corrected score $S_E^*(Y, W : \beta_0, \beta_1, \beta_2, \delta^2)$ for $S_E(Y, X : \beta_0, \beta_1, \beta_2)$. By the definition of corrected score it requires that $S_E^*(Y, W : \beta_0, \beta_1, \beta_2, \delta^2) = S_E(Y, X : \beta_0, \beta_1, \beta_2)$ for all values of $\beta_0, \beta_1, \beta_2$ and when δ is fixed at the true value. Thus the S_E^* can be used to determine a consistent estimates of β_0, β_1 and β_2 when δ^2 is known. However, since the value of β_2 is known but δ^2 is unknown, we can treat the S_E^* as a function with three unknowns β_0, β_1 and δ^2. Then we can use $S_E^*(\cdot)$ to solve for estimates of β_0, β_1 and δ.

3 Extended Corrected Score in Linear and Probit model

It is known that the linear and Probit models are unidentifiable when X_i are normal distributed if δ^2 is unknown. Nevertheless, for other distribution of $X_i's$, it is possible to find a nonredundant corrected score $S_E^*(\cdot)$ to estimate β_0, β_1 and δ^2 consistently.

3.1 The Linear model

For a simple linear regression with measurement error, we have

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad W_i = X_i + \delta U_i,$$

and the quasi-score function is

$$S(Y, X : \beta_0, \beta_1) = \sum_{i=1}^{n} [Y_i - \beta_0 - \beta_1 X_i] \begin{pmatrix} 1 \\ X_i \end{pmatrix}.$$

The extended quasi-score for the extended model $E(Y_i \mid X_i) = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$ is

$$S_E(Y, X : \beta_0, \beta_1, \beta_2) \;=\; \sum_{i=1}^{n} [Y_i - \beta_0 - \beta_1 X_i - \beta_2 X_i^2] \begin{pmatrix} 1 \\ X_i \\ X_i^2 \end{pmatrix}$$

$$= \begin{pmatrix} Y_i - \beta_0 - \beta_1 X_i - \beta_2 X_i^2 \\ Y_i X_i - \beta_0 X_i - \beta_1 X_i^2 - \beta_2 X_i^3 \\ Y_i X_i^2 - \beta_0 X_i^2 - \beta_1 X_i^3 - \beta_2 X_i^4 \end{pmatrix}.$$

To derive a corrected score function, we note that the following hold when U_i are i.i.d. normal distributed.

$$E(W_i \mid X_i) = X_i, \quad E(W_i^2 - \delta^2 \mid X_i) = X_i^2, \quad E(W_i^3 - 3W_i\delta^2 \mid X_i) = X_i^3,$$
$$E(Y_i W_i \mid Y_i, X_i) = Y_i X_i, \quad \text{and} \quad E(Y_i(W_i^2 - \delta^2) \mid Y_i, X_i) = Y_i X_i^2.$$

Replace terms that involve $X_i, X_i^2, X_i^3, Y_i X_i$ and $Y_i X_i^2$ by their corresponding unbiased "estimates", we have a corrected score $S_E^*(\cdot)$

$$S_E^*(Y, W : \beta_0, \beta_1, \beta_2, \delta^2) = \sum_{i=1}^{n} \begin{bmatrix} Y_i - \beta_0 - \beta_1 W_i - \beta_2(W_i^2 - \delta^2) \\ Y_i W_i - \beta_0 W_i - \beta_1(W_i^2 - \delta^2) - \beta_2(W_i^3 - 3W_i\delta^2) \\ Y_i(W_i^2 - \delta^2) - \beta_0(W_i^2 - \delta^2) - \beta_1(W_i^3 - 3W_i\delta^2) - \beta_2 X_i^4 \end{bmatrix}.$$

Apparently $ES_E^* = (0, 0, 0)'$ holds for true β_0, β_1 and δ^2 when $\beta_2 = 0$. Set $\beta_2 = 0$, the estimation function becomes

$$S_E^*(Y, W : \beta_0, \beta_1, 0, \delta^2) = \sum_{i=1}^{n} \begin{bmatrix} Y_i - \beta_0 - \beta_1 W_i \\ Y_i W_i - \beta_0 W_i - \beta_1(W_i^2 - \delta^2) \\ Y_i(W_i^2 - \delta^2) - \beta_0(W_i^2 - \delta^2) - \beta_1(W_i^3 - 3W_i\delta^2) \end{bmatrix},$$

which is an unbiased estimating function. A sufficient condition for $S_E^*(Y, W : \beta_0, \beta_1, 0, \delta^2) = (0, 0, 0)'$ can determine a consistent estimator of $(\beta_0, \beta_1, \delta^2)$ is given as following.

Theorem 1. Let $X_i's, i = 1, \cdots, n$ be i.i.d. distributed random variables, then a sufficient condition for the equation $S_E^*(Y, W : \beta_0, \beta_1, 0, \delta^2) = (0, 0, 0)'$ can determine a consistent estimator is that $-3EX_i^2 EX_i + EX_i^3 + 2(EX_i)^3$ does not equal 0.

proof: Let $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\delta}^2$ be the solution of $S_E^*(Y, W : \beta_0, \beta_1, 0, \delta^2) = (0, 0, 0)'$, then a straightforward calculation shows that

$$\hat{\beta}_1 = \frac{-2\overline{YW}\ \overline{W} + 2(\overline{W})^2\overline{Y} + \overline{YW^2} - \overline{W^2}\ \overline{Y}}{-3\overline{W^2}\ \overline{W} + \overline{W^3} + 2(\overline{W})^3},$$

$$\begin{aligned}\hat{\beta}_0 &= \overline{Y} - \hat{\beta}_1\overline{W}, \quad \text{and} \\ \hat{\sigma}^2 &= \frac{\hat{\beta}_1\overline{W^2} + \hat{\beta}_0\overline{W} - \overline{YW}}{\hat{\beta}_1}\end{aligned}$$

where $\overline{YW} = \sum_1^n Y_i W_i/n$ and others are defined similarly. By law of large numbers, it is obviously that the denominator of $\hat{\beta}_1$ converges to $-3EX_i^2EX_i + EX_i^3 + 2(EX_i)^3$, and the numerator of $\hat{\beta}_1$ converges to $\beta_1[-3EX_i^2EX_i + EX_i^3 + 2(EX_i)^3]$. Hence when $-3EX_i^2EX_i + EX_i^3 + 2(EX_i)^3$ is not zero, $\hat{\beta}_1$ will be consistent. □

Note that the condition in the theorem holds whenever $EX_i^3 \neq 0$ and $EX_i = 0$.

3.2 The probit model

Consider a probit model with binary response Y_i and $E(Y_i \mid X_i) = \Phi(\beta_0 + \beta_1 X_i)$. When X_i can't be observed but observe the surrogate $W_i = X_i + \delta_i U_i$, then (Y_i, W_i) will still be a probit model if (X_i, U_i) are i.i.d. normal distributed, and $E(Y_i \mid W_i) = \Phi(\beta_0' + \beta_1' W_i)$ where $(\beta_0', \beta_1') = (\beta_0, \beta_1)/\sqrt{1 + \beta_1^2\delta^2}$. Thus the porbit model is not identifiable as long as the covariate X_i is normal distributed. However, when X_i is not normal distributed, the parameter (β_0, β_1) and δ^2 may be estimated consistently through the over-parameterization.

Extend the original probit model to a higher order one, that is $E(Y_i \mid X_i) = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$ with $\beta_2 = 0$. The score when X_i is observable is

$$S(Y, X : \beta_0, \beta_1, \beta_2, \delta^2) = \sum_{i=1}^n [Y_i - \Phi(\beta_0 + \beta_1 X_i + \beta_2 X_i^2)] \begin{pmatrix} 1 \\ X_i \\ X_i^2 \end{pmatrix}.$$

When only W_i are observable, we have to invoke the corrected score approach. To find unbiased estimating functions of the terms in $S(Y, X : \beta_0, \beta_1, \beta_2, \delta^2)$, we need the following lemma (Wei, 2005).

Lemma 2. Let $W_i = X_i + \delta U_i$ and U_i are i.i.d. normal distributed. Define $(\beta_0^*, \beta_1^*) = (\beta_0, \beta_1)/\sqrt{1 - \delta^2\beta_1^2}$, then the following equations hold.

(a) $E[\Phi(\beta_0^* + \beta_1^* W_i)|X_i] = \Phi(\beta_0 + \beta_1 X_i)$

(b) $E[W_i\Phi(\beta_0^* + \beta_1^* W_i) - \frac{\delta}{\sqrt{2\pi}}\beta_1^* e^{\frac{-(\beta_0^*+\beta_1^* W_i)^2}{2}}|X_i] = X_i\Phi(\beta_0 + \beta_1 X_i)$

(c) $E[(W_i^2 - \delta^2)\Phi(\beta_0^* + \beta_1^* W_i) - \frac{\sqrt{2}}{\sqrt{\pi}}\delta e^{\frac{-(\beta_0^*+\beta_1^* W_i)^2}{2}}(\beta_1^{*2}\beta_0^*\delta^2 + \beta_1^* W_i$
$+ \beta_1^{*3} W_i\delta^2 - \frac{1}{2}\delta^3\beta_1^{*2}\beta_0^* - \frac{1}{2}\delta^3\beta_1^{*3} W_i)|X_i] = X_i^2\Phi(\beta_0 + \beta_1 X_i)$

With this lemma, it is easy to show that the extended corrected function

$$S_E^*(Y, W : \beta_0, \beta_1, \delta^2) \equiv \sum_{i=1}^{n} \begin{bmatrix} Y_i - \Phi(\beta_0^* + \beta_1^* W_i) \\ W_i Y_i - W_i\Phi(\beta_0^* + \beta_1^* W_i) + \frac{\delta}{\sqrt{2\pi}}\beta_1^* e^{\frac{-(\beta_0^*+\beta_1^* W_i)^2}{2}} \\ Y_i(W_i^2 - \delta^2) - D(\beta_0^*, \beta_1^*, \delta^2) \end{bmatrix}$$

has mean $(0, 0, 0)'$ when conditioned on $X_i's$. When X is not symmetrically distributed, a simulation result not shown here did exhibit the consistent property of the estimators derived from $S_E^*(Y, W : \beta_0, \beta_1, \delta^2) = (0, 0, 0)'$.

4 Approximate extended corrected score in general models

It may happens that for the terms in the original score function, no unbiased "estimates" can be found. For example, in a logistic regression model, there does not exist unbiased estimates for the logistic function $H(\beta_0 + \beta_1 X_i)$, where $H(s) = \frac{1}{1+exp(-s)}$ (Nakamura, 1992). In such situation, we suggest to use an approximation based on small measurement error assumption. The basic idea of our approach on approximating any $X's$ function is illustrated as following.

Let $F(\beta, X_i)$ be the function that we like to estimate by using function of W_i, where $F(\cdot, \cdot)$ is a known function and β is unknown parameter. By Taylor expansion, one has

$$F(\beta, W_i) \quad = \quad F(\beta, X_i) + F_2(\beta, X_i)(W_i - X_i)$$

$$+ \quad \frac{1}{2!}F_{22}(\beta, X_i)(W_i - X_i)^2 + \frac{1}{3!}F_{222}(\beta, X_i^*)(W_i - X_i)^3,$$

where F_2 means the $1st$ partial derivative of F with respect to its 2nd argument, and F_{22} is the 2nd partial derivative. Since $W_i = X_i + \delta U_i$ and U_i are i.i.d. r.v.'s with mean 0 and variance 1, by the law of large numbers and central limit theorem, it follows that

$$\frac{1}{n}\sum_{i=1}^{n} F(\beta, X_i) = \frac{1}{n}\sum_{i=1}^{n} F(\beta, W_i) - \frac{\delta^2}{2n}\sum_{i=1}^{n} F_{i22}(\beta, X_i) + O_p(\delta n^{-1/2}) + O(\delta^3). \tag{4.1}$$

Hence, $\frac{1}{n}\sum_{i=1}^{n} F(\beta, X_i)$ can be approximated by $\frac{1}{n}\sum_{i=1}^{n} F(\beta, W_i) - \frac{\delta^2}{2n}\sum_{i=1}^{n} F_{i22}(\beta, X_i)$ which has error of order δ^3.

As an illustration of this approximation, we consider the logistic regression. When $X_i's$ are observable, the score function is

$$S(Y, X : \beta_0, \beta_1) = \sum_{1}^{n} [Y_i - H(\beta_0 + \beta_1 X_i)] \begin{pmatrix} 1 \\ X_i \end{pmatrix},$$

and the extended score function is

$$S_E(Y, X : \beta_0, \beta_1, \beta_2) = \sum_{1}^{n} [Y_i - H(\beta_0 + \beta_1 X_i + \beta_2 X_i^2)] \begin{pmatrix} 1 \\ X_i \\ X_i^2 \end{pmatrix}.$$

To find an approximate extended corrected score, we need to approximate terms that appear in the S_E function, say, $\frac{1}{n}\sum_{i=1}^{n} X_i^k H(\beta_0 + \beta_1 X_i), k = 0, 1, 2$. Apply the approximation (4.1) to these functions, with some straightforward computations, one can derived an approximate extended corrected score

$$S_E^*(Y, W : \beta_0, \beta_1, \beta_2, \delta^2) = \begin{pmatrix} \sum_1^n Y_i - \sum_1^n [F(\beta_0, \beta_1, \beta_2, W_i) - \frac{\delta^2}{2}\sum_1^n F_{i44}(\beta_0, \beta_1, \beta_2, W_i)] \\ \sum_1^n W_i Y_i - \sum_1^n [G(\beta_0, \beta_1, \beta_2, W_i) - \frac{\delta^2}{2}\sum_1^n G_{i44}(\beta_0, \beta_1, \beta_2, W_i)] \\ \sum_1^n (W_i - \delta^2) Y_i - \sum_1^n [Z(\beta_0, \beta_1, \beta_2, W_i) - \frac{\delta^2}{2}\sum_1^n Z_{i44}(\beta_0, \beta_1, \beta_2, W_i)] \end{pmatrix},$$

where $F(\beta_0, \beta_1, \beta_2, W_i) = H(\beta_0 + \beta_1 W_i + \beta_2 W_i^2)$, $G(\beta_0, \beta_1, \beta_2, W_i) = W_i F(\beta_0, \beta_1, \beta_2, W_i)$, $Z(\beta_0, \beta_1, \beta_2, W_i) = W_i G(\beta_0, \beta_1, \beta_2, W_i)$, and F_{i4}, F_{i44} are the 1st and 2nd partial derivative of F_i with respect to W_i.

Since we know that $\beta_2 = 0$, a simplification shows that the estimates derived from setting $S_E^*(Y, W : \beta_0, \beta_1, 0, \delta^2) = (0, 0, 0)'$ is equivalent to the equation

$$\sum_{i=1}^{n} Y_i - [\sum_{i=1}^{n} F_i - \frac{\delta^2}{2} \sum_{i=1}^{n} F_{i44}] = 0$$
$$\sum_{i=1}^{n} W_i Y_i - [\sum_{i=1}^{n} W_i F_i - \frac{\delta^2}{2} \sum_{i=1}^{n} (2F_{i4}F_i + W_i F_{i44} F_i)] = 0$$
$$\sum_{i=1}^{n} W_i^2 Y_i - [\sum_{i=1}^{n} W_i^2 F_i - \frac{\delta^2}{2} \sum_{i=1}^{n} (4W_i F_{i4} F_i + (W_i^2 + \delta^2) F_{i44} F_i)] = 0,$$

where $F_i = F(\beta_0, \beta_1, 0, W_i)$.

A small simulation was conducted to see if the approximately extended corrected score did yield nearly consistent estimates. We draw $X_i's$ from a standradized χ_1^2 and let $\beta_0 = 1$, $\beta_1 = -1$ and $\delta = 0.333$ so that the reliability is 0.9. With the sample size being 1,000, the results show that the bias in estimating (1, -1, 0.333) is (0.0013, -0.0186, 0.0222) and the mean square errors is (0.00567,0.01799, 0.02058). These values reveal that estimation withut extra information is possible.

5 conclusion

In a measurement error model, the model may be unidentifiable if there is no extra information. The identifiability depends on the distribution of the true covariate $X_i's$ which can be told by examing the distribution of its surrogate $W_i's$. When $X_i's(W_i's)$ is not symmetric distributed, the over-parameterization seems to work fine and did provide consistent estimates in linear and probit regression models. Nevertheless, the (sufficient) condition for when the over-parameterization approach is applicable to the probit model still remain unknown and needs further investigation.

For more general models, it may be difficult to find any corrected score for the original score, not mention the extended corrected score. To overcome this difficulty, we use the small

error approximation and derived approximately consistent estimates. Such approximation seems works well in the logistic model.

When using over-parameterization, a question may be raised. What terms should we choose as the extended terms. In this paper, the X_i^2 is chosen because it is convenient and easy to compute, but not for any efficiency consideration. How to choose an extended term to use in the extended estimating function is worth further investigation. We should pursue this problem in the near future.

References

Carroll, R.J., Ruppert, D., Stefanski, L.A. (1995). *Measurement Errors in Nonlinear Models.* Chapman & Hall, London.

Carroll, R. J. and Stefanski, L. A. (1990). Approximate quasi-likelihood estimation in models with surrogate predictors. *Journal of the American Statistical Association* **85**

Gleser, L. J. and Hwang, J. T. (1987). The nonexistence of 100(1-α)% confiddence sets of finite expected diameter in errors-in-variables and related models. The Annals of Statistic **15,** 1351–1362.

Nakamura, T. (1990). Corrected score function for errors-in-variables models: Methodology and application to generalized linear models. *Biometrika* **77,** 127–137.

Stefanski, L.A., Carroll, R.J. (1987). Conditional scores and optimal scores for generalized linear measurement-error models. *Biometrika* **74,** 703-716.

Wei, Y. F. (2005). Estiamting methods in linear and probit models with measurement errors and without extra information. *Master thesis, Tamkang university.*