The Influence Analysis of Parameter Estimation in Linear Errors-in-Variables Model

Abstract

This paper discusses properties of the parameter estimation of the Linear Errors-in-Variables Model $Y_i = x_i^T b + a + e_i, X_i = x_i + u_i (1 \leq i \leq n)$. The estimations of parameters a, b, σ^2 are derived by using the nearest neighbor-generalized least square method. By analyzing the properties of the unknown parameters a, b and σ^2, we show that the method can obtain good estimation when the data set is good, but the influence functions tell us that the method is not robust: even there is only one point in the data set that is polluted, the result would be unacceptable. Fortunately, we find that the influence functions also provide us with a diagnostic method, and we can obtain better estimations by deleting the outlier.

In the process of proving the main results, we first explore the properties of the explicit expressions of the parameters when X is one dimensional, and when we could not have explicit expressions of them in the case when X is more than one dimensional, we use another way—-Taylor's Expansion—-to gain the same results. Also, two results of Cui & Li (1998) are verified in this special case.

We prove the main results not only via theoretical method but also by simulation—-we developed data sets randomly for the model above and the properties of the data sets fit our results quite well. An example about the probability of the transformation from a suspected case to a real SARS case in Beijing is given to illustrate the application of our results.

Keywords: Errors-in-variables Model, Measurement Error, Local influence, Linear Regression.

1 Introduction

We consider the following linear Errors-in-Variables model(EV model):

$$\begin{cases} Y = x^\tau b + a + e \\ X = x + u \end{cases} \tag{1.1}$$

where X is a random vector in R^p, x and u are $p \times 1$ unobservable covariates and measurement error vectors respectively. b is a $p \times 1$ vector of unknown parameter. Y is a scaler response and e is the model error.

It is assumed that x and $(e, u^\tau)^\tau$ are independent. Let $\Sigma_x = Cov(x)$ and $\Sigma_u = Cov(u)$ be the covariance matrices of the covariates and the measurement error. In order to identify model (1.1), we assume Σ_x is a positive definite matrix(PDM) and $\Sigma_1 =: \Sigma_u / var(e)$ is a known $p \times p$ PDM. Without lose of generality (otherwise, transforming X to $\Sigma_1^{-1/2} X$), we may assume

$$E[(e, u^\tau)^\tau] = 0, Cov[(e, u^\tau)^\tau] = \sigma^2 I_{p+1},$$

which means e and u have the same dispersion parameter $\sigma^2 > 0$. This is the standard framework taken by Cui and Li[1]. Another way to identify model(1.1) is to assume that Σ_u is a known $p \times p$ PDM (Fuller[2]).

Model (1.1) is often encountered in the situations where the true values of a set of variables satisfy the following exact relationship

$$y = x^\tau b + a. \tag{1.2}$$

In these situations we often want to make inference of a and b through the values of y and x. However,what we often encounter is that y is unobservable or even both y and x are unobservable. If y is the only unobservable variable, the well-known linear model is introduced:

$$Y = x^\tau b + a + e. \tag{1.3}$$

In this case, we can use the least square method[3]. If both x and y in (1.2)are unobservable, it is natural and necessary to consider model(1.1), which is an EV model. EV models may be applied to many fields such as economics, biology, forestry and so on.

Many researchers have paid attention to EV models due to their simple form and wide application. Comprehensive reviews of the research and development of EV models can be found in Fuller[2] and the references therein. An important aspect of there research is to explore the whether the methods are robust. A rather comprehensive account of the approach based on influence functions was given in Frank R. Hampel, Elvezio M.Ronchetti,Peter J.Rousseeuw and Werner A.Stahel[11].

The objective of this paper is to discuss model(1.1) when the estimations of a, b, σ^2 are obtained by using the nearest neighbor-generalized least square method. It is shown that the method can obtain good estimations of a, b and σ^2 when the data set is good, but the method is not robust. The influence functions show this property and can help us detect outliers and get better estimations using this un-robust method.

The paper is organized as follows: we formulate the estimations and give the main results in section 2; methods to prove the main results are presented in section 3 (we use different methods to obtain the results in the one dimensional case and the multivariate case); section 4 provides simulation of all the results; an example about the probability of the transformation from a suspected case to a real SARS case in Beijing to illustrate the application of our results is given in section 5.

2 The construction of the estimations and main results

Suppose $\{(X_i = (X_{i1}, X_{i2}, \cdots, X_{ip})^\tau, Y_i) 1 \le i \le n\}$ is a sample of size n for model (1.1). The estimations of a, b, σ^2 are obtained through the following process.

We first give some notations:

$$\bar{X}_{(n)} = \frac{\sum_{i=1}^n X_i}{n}, \bar{Y}_{(n)} = \frac{\sum_{i=1}^n Y_i}{n}.$$

$$X_{(i,n)} = X_i - \bar{X}_{(n)}, Y_{(i,n)} = Y_i - \bar{Y}_{(n)}$$

$$X = (X_1, X_2, \cdots, X_n)^\tau, Y = (Y_1, Y_2, \cdots, Y_n)^\tau$$

$$\tilde{X} = (\tilde{X}_{(1,n)}, \tilde{X}_{(2,n)}, \cdots, \tilde{X}_{(n,n)})^\tau, \tilde{Y} = (\tilde{Y}_{(1,n)}, \tilde{Y}_{(2,n)}, \cdots, \tilde{Y}_{(n,n)})^\tau$$

$$\tilde{x}_{(i,n)} = x_i - \sum_{i=1}^n x_i, \tilde{x} = (\tilde{x}_{(1,n)}, \tilde{x}_{(2,n)}, \cdots, \tilde{x}_{(n,n)})^\tau$$

We estimate b first. Since x_i^τ's are unobservable, the least square method may be invalid. But we can obtain $\hat{b}_n$, the estimation of b, by using the generalized least square method, that is, defining $\hat{b}_n$ as one of the solutions of

$$\sum_{i=1}^n |\frac{\tilde{Y}_{(i,n)} - \tilde{X}^\tau_{(i,n)} b}{\sqrt{1 + \|b\|^2}}|^2 = min(b \in R). \tag{2.1}$$

It follows from (2.1) that $\hat{b_n}$ satisfies

$$(1+\|\hat{b}_n\|^2)(\tilde{X}^\tau \tilde{Y} - \tilde{X}^\tau \tilde{X} \hat{b_n}) + [\tilde{Y}^\tau \tilde{Y} - 2\tilde{Y}^\tau \tilde{X} \hat{b}_n + \hat{b}_n^\tau (\tilde{X}^\tau \tilde{X}) \hat{b_n}] \hat{b_n} = 0 \tag{2.2}$$

If $p = 1$,(2.2)becomes

$$(\sum_{i=1}^n \tilde{X}_{(i,n)} \tilde{Y}_{(i,n)}) \hat{b_n}^2 + (\sum_{i=1}^n \tilde{X}^2_{(i,n)} - \tilde{Y}^2_{(i,n)}) \hat{b_n} - \sum_{i=1}^n \tilde{X}_{(i,n)} \tilde{Y}_{(i,n)} = 0 \tag{2.3}$$

Remark 1. If $p = 1$, from (2.3), we obtain

$$\hat{b}_n(\tilde{X}_{(1,n)}, \tilde{Y}_{(1,n)}, \tilde{X}_{(2,n)}, \tilde{Y}_{(2,n)}, \cdots \tilde{X}_{(n,n)}, \tilde{Y}_{(n,n)}) \\ = \frac{\sum_{i=1}^{n}(\tilde{Y}^2_{(i,n)} - \tilde{X}^2_{(i,n)}) + \sqrt{(\sum_{i=1}^{n}(\tilde{X}^2_{(i,n)} - \tilde{Y}^2_{(i,n)}))^2 + 4(\sum_{i=1}^{n}\tilde{X}_{(i,n)}\tilde{Y}_{(i,n)})^2}}{2\sum_{i=1}^{n}\tilde{X}_{(i,n)}\tilde{Y}_{(i,n)}} \tag{2.4}$$

If $p \geq 2, \hat{b}_n$ has no explicit expression.

We define the estimation of a and σ^2 as

$$\hat{a}_n(X_1, Y_1, X_2, Y_2, \cdots X_n, Y_n) = \bar{Y}_{(n)} - \bar{X}^{\tau}_{(n)}\hat{b}_n \qquad (2.5),$$

$$\hat{\sigma}^2_n(X_1, Y_1, X_2, Y_2, \cdots X_n, Y_n) = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{(Y_i - X_i^{\tau}\hat{b}_n - \hat{a}_n)^2}{1 + \hat{b}_n^2} \tag{2.6}$$

respectively

Rewrite the n-th point (X_n, Y_n) as (x_*, y_*) and let

$$\tilde{x}_{(*,n)} = x_* - \bar{X}_{(n)}, \tilde{y}_{(*,n)} = y_* - \bar{Y}_{(n)},$$

we obtain

Theorem 1. When the former n-1 points were selected from the sample, and the n-th point (x_*, y_*) is a fixed point, then $\{\hat{a}_n\}_{n=1}^{\infty}$, $\{\hat{b}_n\}_{n=1}^{\infty}$ and $\{\hat{\sigma}^2_n\}_{n=1}^{\infty}$ are convergent.

Now suppose that

$$\lim_{n\to\infty} \hat{a}_n = a; \lim_{n\to\infty} \hat{b}_n = b; \lim_{n\to\infty} \hat{\sigma}^2_n = \sigma^2.$$

Write

$$\hat{b}_n(\tilde{X}_{(1,n)}, \tilde{Y}_{(1,n)}, \cdots \tilde{x}_{(*,n)}, \tilde{y}_{(*,n)}) = \hat{b}_n(x_*, y_*)$$

$$\hat{b}_{n-1}(\tilde{X}_{(1,n)}, \tilde{Y}_{(1,n)}, \cdots \tilde{X}_{(n-1,n-1)}, \tilde{Y}_{(n-1,n-1)}) = \hat{b}_{n-1}$$

$$\hat{a}_n(\tilde{X}_{(1,n)}, \tilde{Y}_{(1,n)}, \cdots \tilde{x}_{(*,n)}, \tilde{y}_{(*,n)}) = \hat{a}_n(x_*, y_*)$$

$$\hat{a}_{n-1}(\tilde{X}_{(1,n)}, \tilde{Y}_{(1,n)}, \cdots \tilde{X}_{(n-1,n-1)}, \tilde{Y}_{(n-1,n-1)}) = \hat{a}_{n-1}$$

$$\hat{\sigma}_n^2(\tilde{X}_{(1,n)}, \tilde{Y}_{(1,n)}, \cdots \tilde{x}_{(*,n)}, \tilde{y}_{(*,n)}) = \hat{\sigma}_n(x_*, y_*)^2$$

$$\hat{\sigma}_{n-1}^2(\tilde{X}_{(1,n)}, \tilde{Y}_{(1,n)}, \cdots \tilde{X}_{(n-1,n-1)}, \tilde{Y}_{(n-1,n-1)}) = \hat{\sigma}_{n-1}^2$$

Theorem 2.

$$\begin{aligned} &\lim_{n\to\infty} n[\hat{b}_n(x_*, y_*) - \hat{b}_{n-1}] \\ &= \tfrac{(1+||b||^2)[(x_*-Ex)^\tau(y_*-Ey)-(x_*-Ex)^\tau(x_*-Ex)b]}{(1+||b||^2)cov(\tilde{x})} \\ &+ \tfrac{[(y_*-Ey)^\tau(y_*-Ey)-2(y_*-Ey)^\tau(x_*-Ex)+b^\tau(x_*-Ex)^\tau(x_*-Ex)b]b}{(1+||b||^2)cov(\tilde{x})} \\ &=: f(x_*, y_*) \end{aligned} \tag{2.7}$$

$$E(f(X,Y)) = 0;$$

$$Cov(f(X,Y)) = \frac{1}{(cov(x))^2} \cdot Cov[(e_n - u_n b)(x_n + u_n) + \frac{(e_n - u_n b)^2 \cdot b}{1 + b^2}.$$

Remark 2. If $p = 1$,

$$f(x_*, y_*) = \frac{b((y_* - EY)^2 - (x_* - Ex)^2) - (b^2 - 1)(x_* - Ex)(y_* - EY)}{(1 + b^2) \cdot cov(x)} \tag{2.8}$$

Theorem 3.

$$\begin{aligned} &\lim_{n\to\infty} n[\hat{a}_n(x_*, y_*) - \hat{a}_{n-1}] \\ &= (y_* - EY) - (x_* - EX)^\tau b - (Ex)^\tau \cdot f(x_*, y_*) \\ &=: g(x_*, y_*) \end{aligned} \tag{2.9}$$

$$E(g(X,Y)) = 0$$

Theorem 4.

$$\begin{aligned} &\lim_{n\to\infty} n[\hat{\sigma}_n^2(x_*, y_*) - \hat{\sigma}_{n-1}] \\ &= \tfrac{((y_*-Ey)-(x_*-Ex)^\tau b)^2}{1+||b||^2} - \sigma^2 \\ &=: h(x_*, y_*) \end{aligned} \tag{2.10}$$

$$E(h(X,Y)) = 0$$

$$Cov(h(X,Y)) = Cov(\frac{(e_n - u_n^{\tau} b)^2}{1 + \|b\|^2})$$

Theorem 5. If p=1, the graphs of $f(x_*, y_*)$ and $h(x_*, y_*)$ are **hyperbolic paraboloid** (see Figure 1 below) and **parabolic cylinder** (see Figure 2 below) respectively.

Figure1

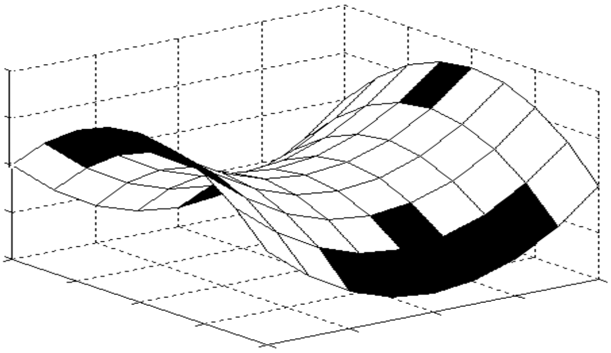
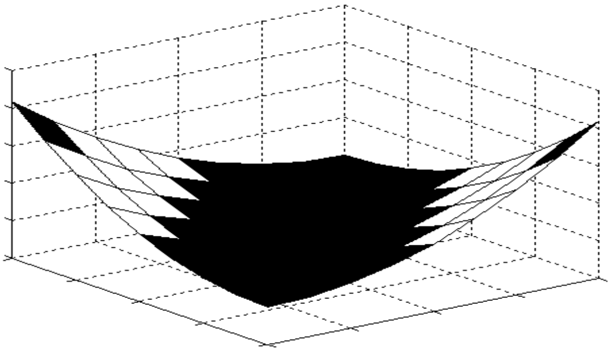


Figure2



Remark 3. We can see the influence of one point(x_*, y_*) to parameters by exploring the heights of $f(x_*, y_*) and h(x_*, y_*)$ in the graphs. Additionally, by deleting some point that is most influential, we may obtain better estimation of the parameters.
Property 1. The value of the influence function $f(x_*, y_*)$ can reach both $+\infty$ and $-\infty$. It equals zero when the n-th point(x_*, y_*) when the n-th point(x_*, y_*) satisfies $(y_* - EY) - (x_* - Ex)^\tau b$
Property 2. The graph of the influence function $h(x_*, y_*)$ is symmetric with respect to the line $y = x^\tau b$. Its value ranges from $-\sigma^2$ to $+\infty$: it reaches its minimal value when the the n-th point(x_*, y_*) satisfies $(y_* - EY) - (x_* - Ex)^\tau b$ and the value would be close to zero when (x_*, y_*) comes from model(1.1).

3 Proofs of main results

3.1 Proof of Theorem 2 when $p = 1$

$$n[\hat{b}_n(x_*, y_*) - \hat{b}_{n-1}] = \frac{A - B + C - D}{E}$$

where

$$A = [\sum_{i=1}^{n}((Y_i - \bar{Y}_{(n)})^2 - (X_i - \bar{X}_{(n)})^2)] \sum_{i=1}^{n-1}(X_i - \bar{X}_{(n-1)})(Y_i - \bar{Y}_{(n-1)});$$

$$B = \sum_{i=1}^{n-1}((Y_i - \bar{Y}_{(n-1)})^2 - (X_i - \bar{X}_{n-1})^2)] \sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_{(n)});$$

$$C = \sum_{i=1}^{n-1}(X_i - \bar{X}_{(n-1)})(Y_i - \bar{Y}_{(n-1)}) \cdot \sqrt{C\prime},$$

$$C' = [\sum_{i=1}^{n}((Y_i - \bar{Y}_{(n)})^2 - (X_i - \bar{X}_{(n)})^2)]^2 + 4(\sum_{i=1}^{n}(X_i - \bar{X}_{(n)})(Y_i - \bar{Y}_{(n)}))^2;$$

$$D = \sum_{i=1}^{n}(X_i - \bar{X}_{(n)})(Y_i - \bar{Y}_{(n)}) \cdot \sqrt{D'},$$

$$\begin{aligned} D' &= [\textstyle\sum_{i=1}^{n-1}((Y_i - \bar{Y}_{(n-1)})^2 - (X_i - \bar{X}_{(n-1)})^2)]^2 \\ &+4(\textstyle\sum_{i=1}^{n-1}(X_i - \bar{X}_{(n-1)})(Y_i - \bar{Y}_{(n-1)}))^2; \end{aligned}$$

and

$$E = 2\sum_{i=1}^{n}(X_i - \bar{X}_{(n)})(Y_i - \bar{Y}_{(n)}) \sum_{i=1}^{n-1}(X_i - \bar{X}_{(n-1)})(Y_i - \bar{Y}_{(n-1)}).$$

Note that

$$\bar{X}_{(n)} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}_{(n-1)} + \frac{x_* - \bar{X}_{(n-1)}}{n}, \qquad (3.1)$$

$$\bar{Y}_n = \frac{\sum_{i=1}^{n} Y_i}{n} = \bar{Y}_{(n-1)} + \frac{y_* - \bar{Y}_{(n-1)}}{n}, \qquad (3.2)$$

$$\begin{aligned} &\textstyle\sum_{i=1}^{n}(X_i - \bar{X}_{(n)})(Y_i - \bar{Y}_{(n)}) \\ &= \textstyle\sum_{i=1}^{n}(X_i - \bar{X}_{(n-1)} - \frac{x_* - \bar{X}_{(n-1)}}{n})(Y_i - \bar{Y}_{(n-1)} - \frac{y_* - \bar{Y}_{(n-1)}}{n}) \\ &+ \tfrac{n-1}{n}(x_* - \bar{X}_{(n-1)})\tfrac{n-1}{n}(y_* - \bar{Y}_{(n-1)}) \\ &= \textstyle\sum_{i=1}^{n}(X_i - \bar{X}_{(n-1)})(Y_i - \bar{Y}_{(n-1)}) + \frac{(n-1)(x_* - \bar{X}_{(n-1)})(y_* - \bar{Y}_{(n-1)})}{n} \\ &- \frac{\sum_{i=1}^{n}(Y_i - \bar{Y}_{(n-1)}(x_* - \bar{X}_{(n-1)}) + \sum_{i=1}^{n}(X_i - \bar{X}_{(n-1)})(y_* - \bar{Y}_{(n-1)})}{n}, \end{aligned} \qquad (3.3)$$

$$\begin{aligned} &\textstyle\sum((Y_i - \bar{Y}_{(n)})^2 - (X_i - \bar{X}_{(n)})^2) \\ &= \textstyle\sum_{i=1}^{n-1}[(Y_i - \bar{Y}_{(n-1)} - \frac{y_* - \bar{Y}_{(n-1)}}{n})^2 - (X_i - \bar{X}_{(n-1)} - \frac{x_* - \bar{X}_{(n-1)}}{n})^2] \\ &+((n-1)\tfrac{y_* - \bar{Y}_{(n-1)}}{n})^2 - ((n-1)\tfrac{x_* - \bar{X}_{(n-1)}}{n})^2 \\ &= \textstyle\sum_{i=1}^{n}[(Y_i - \bar{Y}_{(n-1)})^2 - (X_i - \bar{X}_{(n-1)})^2] \\ &+ \frac{(n-1)[(y_* - \bar{Y}_{(n-1)})^2 - (x_* - \bar{X}_{(n-1)})^2]}{n} \\ &- \frac{2[\sum_{i=1}^{n}[(Y_i - \bar{Y}_{(n-1)})(y_* - \bar{Y}_{(n-1)}) - (X_i - \bar{X}_{(n-1)})(x_* - \bar{X}_{(n-1)})]}{n}, \end{aligned} \qquad (3.4)$$

we obtain

$$\begin{aligned}
&A-B\\
&=\{\tfrac{(n-1)[(y_*-\bar{Y}_{(n-1)})^2-(x_*-\bar{X}_{(n-1)})^2]}{n}\\
&-\tfrac{2[\sum_{i=1}^{n-1}((Y_i-\bar{Y}_{(n-1)})(y_*-\bar{Y}_{(n-1)})-(X_i-\bar{X}_{(n-1)})(x_*-\bar{X}_{(n-1)}))]}{n}\}\\
&\cdot\textstyle\sum_{i=1}^{n-1}(X_i-\bar{X}_{(n-1)})(Y_i-\bar{Y}_{(n-1)})\\
&-\{\textstyle\sum_{i=1}^{n-1}[(Y_i-\bar{Y}_{(n-1)})^2-(X_i-\bar{X}_{(n-1)})^2]\}\\
&\cdot[\tfrac{(n-1)(x_*-\bar{X}_{(n-1)})(y_*-\bar{Y}_{(n-1)})}{n}\\
&-\tfrac{\sum_{i=1}^{n-1}(((Y_i-\bar{Y}_{(n-1)})(x_*-\bar{X}_{(n-1)})+(X_i-\bar{X}_{(n-1)})(y_*-\bar{Y}_{(n-1)}))}{n}].
\end{aligned}$$

It is obvious that

$$E(\tilde{X})=E(\tilde{Y})=0.$$

We know

$$\begin{aligned}
&\lim_{n\to\infty}\tfrac{A-B}{n}\\
&=E(\tilde{X}\tilde{Y})\cdot((y_*-EY)^2-(x_*-EX)^2)\\
&-E(\tilde{Y}^2-\tilde{X}^2)\cdot(x_*-EX)(y_*-EY)
\end{aligned}$$

$$\lim_{n\to\infty}\frac{E}{n^2}=2(E(\tilde{X}\tilde{Y}))^2$$

So

$$\begin{aligned}
&\lim_{n\to\infty}\tfrac{n(A-B)}{E}\\
&=\tfrac{E(\tilde{X}\tilde{Y})\cdot((y_*-EY)^2-(x_*-EX)^2)-E(\tilde{Y}^2-\tilde{X}^2)\cdot(x_*-EX)(y_*-EY)}{2(E(\tilde{X}\tilde{Y}))^2}
\end{aligned}$$

Similarly, we can calculate out that

$$C=\sum_{i=1}^{n-1}\tilde{X}_{(i,n-1)}\tilde{Y}_{(i,n-1)}\sqrt{C'}$$

where

$$\begin{aligned}
&C'=[\textstyle\sum_{i=1}^{n-1}(\tilde{Y}^2_{(i,n-1)}-\tilde{X}^2_{(i,n-1)})+\tfrac{n-1}{n}(\tilde{y}^2_{(*,n-1)}-\tilde{x}^2_{(*,n-1)})\\
&-\tfrac{2\sum_{i=1}^{n-1}(\tilde{Y}_{(i,n-1)}\tilde{y}_{(*,n-1)}-\tilde{X}_{(i,n-1)}\tilde{x}_{(*,n-1)})}{n}]^2\\
&+4(\textstyle\sum_{i=1}^{n-1}\tilde{X}_{(i,n-1)}\tilde{Y}_{(i,n-1)}+\tfrac{n-1}{n}\tilde{x}_{(*,n-1)}\tilde{y}_{(*,n-1)}\\
&-\tfrac{\sum_{i=1}^{n-1}\tilde{Y}_{(i,n-1)}\tilde{x}_{(*,n-1)}}{n}-\tfrac{\sum_{i=1}^{n-1}\tilde{X}_{(i,n-1)}\tilde{y}_{(*,n-1)}}{n})^2
\end{aligned}$$

$$\begin{array}{l}
D=(\sum_{i=1}^{n-1}\tilde{X}_{(i,n-1)}\tilde{Y}_{(i,n-1)}+\frac{(n-1)}{n}\tilde{x}_{(*,n-1)}\tilde{y}_{(*,n-1)} \\
-\frac{1}{n}\sum_{i=1}^{n-1}\tilde{Y}_{(i,n-1)}\tilde{x}_{(*,n-1)}-\frac{1}{n}\sum_{i=1}^{n-1}\tilde{X}_{(i,n-1)}\tilde{y}_{(*,n-1)}) \\
\cdot\sqrt{(\sum_{i=1}^{n-1}(\tilde{Y}^2_{(i,n-1)}-\tilde{X}^2_{(i,n-1)}))^2+4(\sum_{i=1}^{n-1}\tilde{X}_{(i,n-1)}\tilde{Y}_{(i,n-1)})^2}
\end{array}$$

$$\begin{array}{l}
C^2-D^2 \\
=(\sum_{i=1}^{n-1}\tilde{X}_{(i,n-1)}\tilde{Y}_{(i,n-1)})^2\cdot\{\frac{(n-1)(\tilde{y}^2_{(*,n-1)}-\tilde{x}^2_{(*,n-1)})}{n} \\
+(\frac{2\sum_{i=1}^{n-1}(\tilde{Y}_{(i,n-1)}\tilde{y}_{(*,n-1)}-\tilde{X}_{(i,n-1)}\tilde{x}_{(*,n-1)})}{n})^2 \\
+2\sum_{i=1}^{n-1}(\tilde{Y}^2_{(i,n-1)}-\tilde{X}^2_{(i,n-1)})\cdot\frac{(n-1)(\tilde{y}^2_{(*,n-1)}-\tilde{x}^2_{(*,n-1)})}{n} \\
-\frac{4}{n}\sum_{i=1}^{n-1}(\tilde{Y}_{(i,n-1)}\tilde{y}_{(*,n-1)}-\tilde{X}_{(i,n-1)}\tilde{x}_{(*,n-1)})\cdot\sum_{i=1}^{n-1}(\tilde{Y}^2_{(i,n-1)}-\tilde{X}^2_{(i,n-1)}) \\
-\frac{4(n-1)}{n^2}\sum_{i=1}^{n-1}(\tilde{Y}_{(i,n-1)}\tilde{y}_{(*,n-1)}-\tilde{X}_{(i,n-1)}\tilde{x}_{(*,n-1)})\cdot(\tilde{y}^2_{(*,n-1)}-\tilde{x}^2_{(*,n-1)})\} \\
-(\sum_{i=1}^{n-1}(\tilde{Y}^2_{(i,n-1)}-\tilde{X}^2_{(i,n-1)}))^2\cdot[(\frac{n-1}{n}\tilde{x}_{(*,n-1)}\tilde{y}_{(*,n-1)})^2 \\
+(\frac{\sum_{i=1}^{n-1}\tilde{X}_{(i,n-1)}\tilde{y}_{(*,n-1)}}{n})^2+(\frac{\sum_{i=1}^{n-1}\tilde{Y}_{(i,n-1)}\tilde{x}_{(*,n-1)}}{n})^2 \\
+\frac{2(n-1)}{n}\tilde{x}_{(*,n-1)}\tilde{y}_{(*,n-1)}\sum_{i=1}^{n-1}\tilde{X}_{(i,n-1)}\tilde{Y}_{(i,n-1)} \\
-\frac{2}{n}\sum_{i=1}^{n-1}\tilde{Y}_{(i,n-1)}\tilde{x}_{(*,n-1)}\sum_{i=1}^{n-1}\tilde{X}_{(i,n-1)}\tilde{Y}_{(i,n-1)} \\
-\frac{2}{n}\sum_{i=1}^{n-1}\tilde{X}_{(i,n-1)}\tilde{y}_{(*,n-1)}\sum_{i=1}^{n-1}\tilde{X}_{(i,n-1)}\tilde{Y}_{(i,n-1)} \\
-\frac{2(n-1)}{n^2}\tilde{x}_{(*,n-1)}\tilde{y}_{(*,n-1)}\sum_{i=1}^{n-1}\tilde{Y}_{(i,n-1)}\tilde{x}_{(*,n-1)} \\
-\frac{2(n-1)}{n^2}\tilde{x}_{(*,n-1)}\tilde{y}_{(*,n-1)}\sum_{i=1}^{n-1}\tilde{X}_{(i,n-1)}\tilde{y}_{(*,n-1)} \\
+\frac{2}{n^2}\sum_{i=1}^{n-1}\tilde{Y}_{(i,n-1)}\tilde{x}_{(*,n-1)}\sum_{i=1}^{n-1}\tilde{X}_{(i,n-1)}\tilde{y}_{(*,n-1)}]
\end{array}$$

$$\begin{array}{l}
\lim_{n\to\infty}\frac{C^2-D^2}{n^3} \\
=E(\tilde{X}\tilde{Y})^2\cdot 2E(\tilde{Y}^2-\tilde{X}^2)\cdot((y_*-EY)^2-(x_*-EX)^2) \\
-(E(\tilde{Y}^2-\tilde{X}^2))^2\cdot E(\tilde{X}\tilde{Y})\cdot 2(x_*-EX)(y_*-EY)
\end{array}$$

$$\lim_{n\to\infty}\frac{(C+D)\cdot E}{n^4}=4(E(\tilde{X}\tilde{Y}))^3\cdot\sqrt{(E(\tilde{Y}^2-\tilde{X}^2))^2+4(E(\tilde{X}\tilde{Y}))^2}$$

So we get

$$\begin{array}{l}
\lim_{n\to\infty}\frac{n(C-D)}{E} \\
=\lim_{n\to\infty}\frac{n(C^2-D^2)}{(C+D)\cdot E} \\
=\frac{E(\tilde{X}\tilde{Y})\cdot E(\tilde{Y}^2-\tilde{X}^2)\cdot((y_*-EY)^2-(x_*-EX)^2)-(E(\tilde{Y}^2-\tilde{X}^2))^2\cdot(x_*-EX)(y_*-EY)}{2(E(\tilde{X}\tilde{Y}))^2\cdot\sqrt{(E(\tilde{Y}^2-\tilde{X}^2))^2+4(E(\tilde{X}\tilde{Y}))^2}}
\end{array}$$

thus,

$$\begin{aligned}
&\lim_{n\to\infty} n[\hat{b}_n(x_*, y_*) - \hat{b}_{n-1}] \\
&= \tfrac{(E(Y^2-X^2)+\sqrt{(E(Y^2-X^2))^2+4(E(XY))^2]})}{2(E(XY))^2\cdot\sqrt{(E(Y^2-X^2))^2+4(E(XY))^2}} \\
&\cdot[E(XY)\cdot((y_* - EY)^2 - (x_* - EX)^2) \\
&-E(Y^2 - X^2)\cdot(x_* - EX)(y_* - EY)] \\
&=: f(x_*, y_*)
\end{aligned}$$

Because

$$\begin{aligned}
&EX^2 = Ex^2 + \sigma^2; EY^2 = E(xb)^2 + \sigma^2; \\
&E(XY) = E((x+u)\cdot(xb+e)) = E(x^2b + xub + xe + eu) = bEx^2; \\
&E(Y^2 - X^2) = (b^2 - 1)Ex^2,
\end{aligned}$$

we obtain

$$\begin{aligned}
&f(x_*, y_*) \\
&= \tfrac{b((y_*-Ey)^2-(x_*-Ex)^2)-(b^2-1)(x_*-Ex)(y_*-Ey)}{(1+b^2)\cdot cov(x)}.
\end{aligned} \tag{3.5}$$

It is easily verified that

$$E(f(X,Y)) = 0 \tag{3.6}$$

and

$$\begin{aligned}
Cov(f(X,Y)) &= \tfrac{Cov[b((bx_n+e_n)^2-(x_n+u_n)^2)-(b^2-1)(x_n+u_n)(bx_n+e_n)]}{(1+b^2)^2(cov(x))^2} \\
&= \tfrac{1}{cov(x)}\cdot Cov[(e_n - u_nb)(x_n + u_n) + \tfrac{(e_n-u_nb)^2\cdot b}{1+b^2}]
\end{aligned} \tag{3.7}$$

which is the same as the result of Professor Cui[2].

3.2 Proof of Theorem 2 when $p > 1$

If $p > 1$, $\hat{b}_n$ has no explicit expression, and we can't calculate $n(\hat{b}_n - \hat{b}_{n-1})$ directly. We tried to reach the result via another way: using Taylor's expansion.

Define vector function

$$
\begin{aligned}
&F_{(n)}(b)\\
&= (F_1(b), F_2(b) \cdots F_p(b))^\tau\\
&= (1 + \|b\|^2)(\tfrac{1}{n}\tilde{X}^\tau \tilde{Y} - \tfrac{1}{n}\tilde{X}^\tau \tilde{X} b) + [\tfrac{1}{n}\tilde{Y}^\tau \tilde{Y} - \tfrac{2}{n}\tilde{Y}^\tau \tilde{X} b + b^\tau(\tfrac{1}{n}\tilde{X}^\tau \tilde{X})b]b.
\end{aligned}
$$

From (2.2) we know

$$F_{(n)}(\hat{b}_n) = 0$$

$$F_{(n-1)}(\hat{b}_{n-1}) = 0$$

On the other hand, using Taylor's expansion, we have

$$F_{(n)}(\hat{b}_n) = F_{(n)}(\hat{b}_{n-1}) + C_n(\hat{b}_n - \hat{b}_{n-1}) \tag{3.8}$$

where

$$
C_n = \begin{pmatrix}
\frac{\partial F_{(n),1}}{\partial b_1}, \frac{\partial F_{(n),1}}{\partial b_2}, \cdots \frac{\partial F_{(n),1}}{\partial b_p}\Big|_{b=\hat{b}_{n-1}+\xi_1(\hat{b}_n-\hat{b}_{n-1})}\\
\frac{\partial F_{(n),2}}{\partial b_1}, \frac{\partial F_{(n),2}}{\partial b_2}, \cdots \frac{\partial F_{(n),2}}{\partial b_p}\Big|_{b=\hat{b}_{n-1}+\xi_2(\hat{b}_n-\hat{b}_{n-1})}\\
\cdots, \cdots, \cdots\\
\frac{\partial F_{(n),p}}{\partial b_1}, \frac{\partial F_{(n),p}}{\partial b_2}, \cdots \frac{\partial F_{(n),p}}{\partial b_p}\Big|_{b=\hat{b}_{n-1}+\xi_p(\hat{b}_n-\hat{b}_{n-1})}
\end{pmatrix}
$$

for some $\xi = (\xi_1, \xi_2, \cdots, \xi_p)^\tau \in [0,1]^p$.

Note that

$$
\begin{aligned}
&F_{(n)}(\hat{b}_{n-1})\\
&= (1 + \|\hat{b}_{n-1}\|^2)(\tfrac{1}{n}\tilde{X}^\tau_{(n)}\tilde{Y}_{(n)} - \tfrac{1}{n}\tilde{X}^\tau_{(n)}\tilde{X}_{(n)}\hat{b}_{n-1})\\
&+[\tfrac{1}{n}\tilde{Y}^\tau_{(n)}\tilde{Y}_{(n)} - \tfrac{2}{n}\tilde{Y}^\tau_{(n)}\tilde{X}_{(n)}\hat{b}_{n-1} + \hat{b}^\tau_{n-1}(\tfrac{1}{n}\tilde{X}^\tau_{(n)}\tilde{X}_{(n)})\hat{b}_{n-1}]\hat{b}_{n-1}\\
&= (1 + \|\hat{b}_{n-1}\|^2)\{\tfrac{n-1}{n}[\tfrac{1}{n-1}(\tilde{X}^\tau_{(n-1)}\tilde{Y}_{(n-1)} + \tilde{x}^\tau_*\tilde{y}_*)\\
&-\tfrac{1}{n-1}(\tilde{X}^\tau_{(n-1)}\tilde{X}_{(n-1)} + \tilde{x}^\tau_*\tilde{x}_*)\hat{b}_{n-1}]\\
&+[\tfrac{1}{n-1}(\tilde{Y}^\tau_{(n-1)}\tilde{Y}_{(n-1)} + \tilde{y}^\tau_*\tilde{y}_*) - \tfrac{2}{n-1}(\tilde{Y}^\tau_{(n-1)}\tilde{X}_{(n-1)} + \tilde{y}^\tau_*\tilde{x}_*)\hat{b}_{n-1}]\}\hat{b}_{n-1}\\
&= \tfrac{n-1}{n}F_{(n-1)}(\hat{b}_{n-1}) + (1 + \|\hat{b}_{n-1}\|^2)(\tfrac{1}{n}\tilde{x}^\tau_*\tilde{y}_* - \tfrac{1}{n}\tilde{x}^\tau_*\tilde{x}_*\hat{b}_{n-1})\\
&+[\tfrac{1}{n}\tilde{y}^\tau_*\tilde{y}_* - \tfrac{2}{n}\tilde{y}^\tau_*\tilde{x}_*\hat{b}_{n-1} + \hat{b}^\tau_{n-1}(\tfrac{1}{n}\tilde{x}^\tau_*\tilde{x}_*)\hat{b}_{n-1}]\hat{b}_{n-1},
\end{aligned}
$$

so (3.8) becomes

$$\begin{aligned}&0 = 0 + (1 + \|\hat{b}_{n-1}\|^2)(\tfrac{1}{n}\tilde{x}_*^\tau \tilde{y}_* - \tfrac{1}{n}\tilde{x}_*^\tau \tilde{x}_* \hat{b}_{n-1})\\ &+[\tfrac{1}{n}\tilde{y}_*^\tau \tilde{y}_* - \tfrac{2}{n}\tilde{y}_*^\tau \tilde{x}_* \hat{b}_{n-1} + \hat{b}_{n-1}^\tau(\tfrac{1}{n}\tilde{x}_*^\tau \tilde{x}_*)\hat{b}_{n-1}]\hat{b}_{n-1} + C_n(\hat{b}_n - \hat{b}_{n-1})\end{aligned}$$

which means that

$$\begin{aligned}&n[\hat{b}_n(x_*, y_*) - \hat{b}_{n-1}]\\ &= C_n^{-1} \cdot \{-(1 + \|\hat{b}_{n-1}\|^2)(\tfrac{1}{n}\tilde{x}_*^\tau \tilde{y}_* - \tfrac{1}{n}\tilde{x}_*^\tau \tilde{x}_* \hat{b}_{n-1})\\ &-[\tfrac{1}{n}\tilde{y}_*^\tau \tilde{y}_* - \tfrac{2}{n}\tilde{y}_*^\tau \tilde{x}_* \hat{b}_{n-1} + \hat{b}_{n-1}^\tau(\tfrac{1}{n}\tilde{x}_*^\tau \tilde{x}_*)\hat{b}_{n-1}]\hat{b}_{n-1}\}\end{aligned} \tag{3.9}$$

Next, let's explore the properties of C_n:
by direct calculation, we get

$$\begin{aligned}&\tfrac{\partial F(n)}{\partial b}\\ &= -(1 + \|b\|^2)\tfrac{1}{n}\tilde{X}^\tau \tilde{X} + 2a(\tfrac{1}{n}\tilde{X}^\tau \tilde{Y} - \tfrac{1}{n}\tilde{X}^\tau \tilde{X} b)^\tau\\ &+(\tfrac{1}{n}\tilde{Y}^\tau \tilde{Y} - \tfrac{2}{n}\tilde{Y}^\tau \tilde{X} b + b^\tau(\tfrac{1}{n}\tilde{X}^\tau \tilde{X})b)I_p + (-\tfrac{2}{n}\tilde{X}^\tau \tilde{Y} + \tfrac{2}{n}\tilde{X}^\tau \tilde{X} b)b^\tau\end{aligned}$$

From (3.9) we see that $\{\hat{b}_n\}_{n=1}^\infty$ is convergent, so let's suppose $\hat{b}_n \to b$ as $n \to \infty$. We have the following results:
as $n \to \infty$,

$$\frac{1}{n}\tilde{X}^\tau \tilde{X} \to Cov(\tilde{x}) + \sigma^2 I_p$$

$$\frac{1}{n}\tilde{X}^\tau \tilde{Y} \to Cov(\tilde{x}) \cdot b$$

$$\frac{1}{n}\tilde{Y}^\tau \tilde{Y} \to b^\tau Cov(\tilde{x}) b + \sigma^2.$$

So we know

$$C_n \to -(1 + \|b\|^2) \cdot Cov(\tilde{x}) \qquad (when\ n \to \infty).$$

Then we can conclude

$$\begin{aligned}&\lim_{n\to\infty} n[\hat{b}_n(x_*, y_*) - \hat{b}_{n-1}]\\ &= \tfrac{(1+||b||^2)[(x_*-Ex)^\tau(y_*-EY)-(x_*-Ex)^\tau(x_*-Ex)b]}{(1+||b||^2)cov(\tilde{x})}\\ &+\tfrac{[(y_*-EY)^\tau(y_*-EY)-2(y_*-EY)^\tau(x_*-Ex)+b^\tau(x_*-Ex)^\tau(x_*-Ex)b]b}{(1+||b||^2)cov(\tilde{x})}.\end{aligned}$$

It is the same as what we obtained in the case of $p = 1$.

3.3 Proof of Property 1

When $p = 1$,

$$\begin{aligned} f(x_*, y_*) &= \tfrac{b((y_* - Ey)^2 - (x_* - Ex)^2) - (b^2 - 1)(x_* - Ex)(y_* - EY)}{(1+b^2)\cdot cov(x)} \\ &= \tfrac{b}{(1+b^2)\cdot cov(x)}[(y_* - \tfrac{b^2-1}{2b}x_*)^2 - ((\tfrac{b^2+1}{2b}x_*)^2], \end{aligned}$$

thus one easily sees that the graph of $f(x_*, y_*)$ is a hyperbolic paraboloid. It is unbounded: its value can reach both $+\infty$ and $-\infty$ (see graph1).

Also, the value equals zero when $((x_* - Ex), (y_* - EY))$ satisfies

$$(y - \frac{b^2 - 1}{2b}x) = \pm((\frac{b^2 + 1}{2b})x. \tag{3.10}$$

(3.4) is equivalent to $y = bx$ or $y = -\frac{1}{b}x$. That is to say, if and only if the point $((x_* - Ex), (y_* - EY))$ is on the line $y = bx$ or on the line $y = -\frac{1}{b}x$, the added point dose not affect the estimation $\hat{b}_n$. But when it goes far away from these two lines, the effect to the estimation would become big.

By the similar way, we calculate out the results when $p > 1$.

3.4 Proof of Theorem 3

$$\begin{aligned} & n(\hat{a}_n(x_*, y_*) - \hat{a}_{n-1}) \\ & = n(\bar{Y}_{(n)} - \bar{X}^{\tau}_{(n)}\hat{b}_n - \bar{Y}_{(n-1)} - \bar{X}^{\tau}_{(n-1)}\hat{b}_{n-1}) \\ & = n(\bar{Y}_{(n-1)} + \tfrac{y_* - \bar{Y}_{(n-1)}}{n} - (\bar{X}^{\tau}_{(n-1)} + \tfrac{x^{\tau}_* - \bar{X}^{\tau}_{(n-1)}}{n})\hat{b}_n - \bar{Y}_{(n-1)} - \bar{X}^{\tau}_{(n-1)}\hat{b}_{n-1}) \\ & = n(\tfrac{y_* - \bar{Y}_{(n-1)}}{n} - \tfrac{x^{\tau}_* - \bar{X}^{\tau}_{(n-1)}}{n}\hat{b}_n - \bar{X}^{\tau}_{(n-1)}(\hat{b}_n - \hat{b}_{n-1})) \\ & = (y_* - \bar{Y}_{(n-1)}) - (x^{\tau}_* - \bar{X}^{\tau}_{(n-1)})\hat{b}_n - \bar{X}^{\tau}_{(n-1)}n(\hat{b}_n - \hat{b}_{n-1}) \\ & \to (y_* - EY) - (x_* - EX)^{\tau}b - (Ex)^{\tau} \cdot f(x_*, y_*) \\ & := g(x_*, y_*) \quad (\ as\ n\ \to\ \infty) \end{aligned} \tag{3.11}$$

It is obvious that $E(g(X, Y)) = 0$.

3.5 Proof of Theorem 4

Using the same technique as in the section 3.1, noticing the formula (3.1),(3.2),(3.3)and (3.4), we obtain

$$\begin{aligned} & n[\hat{\sigma}_n^2(x_*, y_*) - \hat{\sigma}_{n-1}^2] \\ & = n[\tfrac{1}{n} \cdot \tfrac{\sum_{i=1}^{n-1}(\tilde{Y}_{(i,n)} - \tilde{X}_{(i,n)}\hat{b}_n(x_*,y_*))^2 + (y_* - x_*\hat{b}_n(x_*,y_*))^2}{1+\hat{b}_n^2(x_*,y_*)} \\ & -\tfrac{1}{n-1} \cdot \tfrac{\sum_{i=1}^{n-1}(\tilde{Y}_{(i,n)} - \tilde{X}_{(i,n)}\hat{b}_{n-1})^2}{1+\hat{b}_{n-1}^2}] \\ & = O(x_*, y_*) + P(x_*, y_*) + Q(x_*, y_*) \end{aligned} \tag{3.12}$$

where

$$\begin{aligned} O(x_*, y_*) \quad & = (\hat{b}_n(x_*, y_*) - \hat{b}_{n-1}) \\ & \cdot [\tfrac{\sum_{i=1}^{n-1} Y_i^2(-(\hat{b}_n(x_*,y_*) + b_{n-1})) - \sum_{i=1}^{n-1} 2X_iY_i(1 - \hat{b}_n(x_*,y_*)\hat{b}_{n-1})}{(1+\hat{b}_n(x_*,y_*)^2)(1+\hat{b}_{n-1}^2)} \\ & + \tfrac{\sum_{i=1}^{n-1} X_i^2(\hat{b}_n(x_*,y_*) + \hat{b}_{n-1})}{(1+\hat{b}_n(x_*,y_*)^2)(1+\hat{b}_{n-1}^2)}]; \end{aligned}$$

$$P(x_*, y_*) = \frac{-[\sum_{i=1}^{n-1} Y_i^2 - \hat{b}_{n-1} \sum_{i=1}^{n-1} 2X_iY_i + \hat{b}_{n-1}^2 \sum_{i=1}^{n-1} X_i^2]}{(n-1)(1+\hat{b}_{n-1}^2)};$$

$$\begin{aligned} Q(x_*, y_*) \quad & = \tfrac{\sum_{i=1}^{n-1} -2\tilde{Y}_{(i,n)}\frac{\tilde{y}_{(*,n-1)}}{n} + \sum_{i=1}^{n-1} 2\tilde{Y}_{(i,n)}\frac{\tilde{x}_{(*,n-1)}}{n}\hat{b}_n(x_*,y_*)}{(1+\hat{b}_n(x_*,y_*)^2)} \\ & + \tfrac{\sum_{i=1}^{n-1} 2\tilde{X}_{(i,n)}\frac{\tilde{y}_{(*,n-1)}}{n}\hat{b}_n(x_*,y_*) - \sum_{i=1}^{n-1} 2\tilde{X}_{(i,n)}\frac{\tilde{x}_{(*,n-1)}}{n}\hat{b}_n^2(x_*,y_*)}{(1+\hat{b}_n(x_*,y_*)^2)} \\ & + \tfrac{\frac{n-1}{n}(\tilde{y}_{(*,n-1)} - \tilde{x}_{(*,n-1)}\hat{b}_n(x_*,y_*))^2}{(1+\hat{b}_n(x_*,y_*)^2)}. \end{aligned}$$

From (3.12) we see that$\{\hat{\sigma}_n^2\}$ is convergent.
Suppose $\hat{\sigma}_n^2 \to \sigma^2 (n \to +\infty)$. We can see that

$$\begin{aligned} & O(x_*, y_*) \\ & \to \tfrac{f(x_*,y_*)}{(1+b^2)^2} \cdot [(-2b)(b^2Ex^2 + \sigma^2) - (1-b^2)2bEx^2 + 2b(Ex^2 + \sigma^2)] \\ & = 0 \ (n \to +\infty) \end{aligned}$$

$$\begin{aligned} P(x_*, y_*) &= \frac{-[\sum_{i=1}^{n-1}(\tilde{Y}_{(i,n)} - \hat{b}_{n-1}\tilde{X}_{(i,n)})^2]}{(n-1)(1+\hat{b}_{n-1}^2)} \\ &= -\hat{\sigma}_{n-1}^2 \to -\sigma^2(n \to +\infty) \end{aligned}$$

$$Q(x_*, y_*) = \to \frac{((y_* - EY) - (x_* - Ex)b)^2}{(1 + \|b\|^2)}(n \to +\infty)$$

Thus we conclude that

$$n(\sigma_n^2 - \sigma^2) \to \frac{[(y_* - EY) - (x_* - EX)b]^2}{(1 + b^2)} - \sigma^2 := h(x_*, y_*)(n \to +\infty). \tag{3.13}$$

We can calculate out by the same method that in case $p > 1$,

$$\begin{aligned} &\lim_{n\to\infty} n[\hat{\sigma}_n^2(x_*, y_*) - \hat{\sigma}_{n-1}^2] \\ &= \frac{((y_* - EY) - (x_* - Ex)^\tau b)^2}{(1+\|b\|^2)} - \sigma^2 \end{aligned} ,$$

which is the same as that in case $p = 1$.

We see that $E(h(X, Y)) = 0$

and

$$\begin{aligned} &Cov(h(X, Y)) \\ &= E(h(X, Y))^2) \\ &= E[(\frac{(e_n - u_n b)^2}{(1+b^2)})^2 - 2\sigma^2 \frac{(e_n - u_n b)}{(1+b^2)} + \sigma^4] \\ &= E(\frac{(e_n - u_n b)^2}{(1+b^2)})^2 \\ &= Cov(\frac{(e_n - u_n b)^2}{(1+b^2)}) \end{aligned}$$

which is also the same as the result of Professor Cui[1]

3.6 Proof of Property 2

When $p = 1$, from (3.12) we see that the graph of $h(x_*, y_*)$ is a parabolic cylinder: it is symmetric with respect to the line $y - EY = b(x - Ex)$. When (x_*, y_*) lies on the line, $h(x_*, y_*)$ reaches its minimum value $-\sigma^2$; when (x_*, y_*) satisfies $\frac{[(y_*-EY)-(x_*-EX)b]^2}{(1+b^2)} = \sigma^2$, $h(x_*, y_*) = 0$. It is to say, when (x_*, y_*) comes from model(1.1), $h(x_*, y_*)$ would be close to zero. While if the point (x_*, y_*) goes far away from its symmetry axis $y - EY = b(x - Ex)$, the value of $h(x_*, y_*)$ may reach ∞.

Similarly, we can extend the results to the case $p > 1$.

3.7 Proof of Theorem 5

The proof of theorem 5 can be found in the proof of property 1 and the proof of property 2.

3.8 Proof of Theorem 1

(3.9) tells us that $\{\hat{b}_n\}_{n=1}^{\infty}$ is convergent ; (3.11) tells us that $\{\hat{a}_n\}_{n=1}^{\infty}$ is convergent ; (3.12) tells us that $\{\hat{\sigma}_n^2\}_{n=1}^{\infty}$ is convergent. These complete the proof of Theorem 1.

4 SIMULATION:

Without lost of generality, we only consider the case
when $p = 1$.

4.1 About $\hat{b}_n$

We developed 200 data sets by computer random simulation (See Program 1). Each of the data sets contains 100 points of (X,Y),

$$\begin{cases} Y = x + e \\ X = x + u \end{cases}$$

where $x \sim N(0, 1), e \sim N(0, 0.25), u \sim N(0, 0.25)$. By formula (2.4), we can figure out $\hat{b}_{100}$ from the j-th data set, so we have

200 $\hat{b}'_{100}s$ (see attached table 1).
Their average is

$$\tilde{E}(b) = \frac{\Sigma_{j=1}^{200} \hat{b}_{100,j}}{200} = 0.9991,$$

quite close to the real value "1". And the variance is

$$\tilde{D}(b) = \frac{\Sigma_{j=1}^{200} \hat{b}_{100,j}^2}{200} - (\frac{\Sigma_{j=1}^{200} \hat{b}_{100,j}}{200})^2 = 0.0060,$$

which is very small. These tell us that formula (2.4) can obtain quite good estimation.

4.2 About $\hat{a}_n$

By formula (2.5) we obtain $\hat{a}_{100,j}(j = 1 \cdots 200)$ (see attached table 2). Their average is

$$\tilde{E}(a) = \frac{\Sigma_{j=1}^{200} \hat{a}_{100,j}}{200} = -3.9267e - 005,$$

quite close to the real value "0". And the variance is

$$\tilde{D}(b) = \frac{\Sigma_{j=1}^{200} \hat{b}_{100,j}^2}{200} - (\frac{\Sigma_{j=1}^{200} \hat{b}_{100,j}}{200})^2 = 0.0038,$$

which is very small. These tell us that formula (2.5) can obtain quite good estimation.

4.3 About $\hat{\sigma}_n^2$

By formula (2.6) we obtain $\hat{\sigma}_{100,j}^2(j = 1 \cdots 200)$ (see attached table 3). Their average is

$$\tilde{E}(\hat{\sigma}_{100}^2) = \frac{\Sigma_{j=1}^{200} \hat{\sigma}_{100,j}^2}{200} = 0.2495,$$

quite close to the real value “0.25”. And the variance is

$$\tilde{D}(\sigma^2) = \frac{\Sigma_{j=1}^{200}(\hat{\sigma}^2_{100,j})^2}{200} - (\frac{\Sigma_{j=1}^{200}\hat{\sigma}^2_{100,j}}{200})^2 = 0.0011,$$

also very small. These tell us that formula (2.6) can obtain quite good estimation.

4.4 About Theorem 1 and Theorem 2

Let’s choose randomly one data set from the sets developed in section (4.1)(see the data set in table 4). It is a good data set and
by (2.4) we get $\hat{b}_{100} = 1.0203$;
by (2.5) we get $\hat{a}_{100} = 0.0096$;
by (2.6) we get $\hat{\sigma}^2_{100} = 0.2596$.
It is a quite good estimation as we see that errors are small and the estimated line fits the real line quite well:

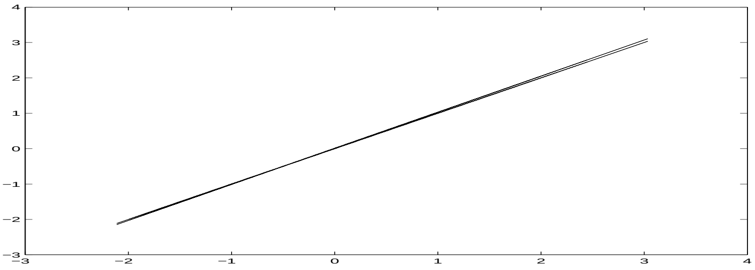
$$|b - \hat{b}_{100}| = 2.03\% \qquad (4.1)$$

$$|a - \hat{a}_{100}| = 0.96\% \qquad (4.2)$$

$$|\sigma - \hat{\sigma}_{100}| = 3.84\% \qquad (4.3)$$

We compare the real line and the estimated line in graph3:

Figure3



But if one of its points is polluted, the result may be not that good. For example, randomly choose one of the $Y_i's$ and let it be enlarged by 10 times, then $\hat{b}_{100}$ becomes 1.9136 , $\hat{a}_{100}$ becomes -0.2073, and $\hat{\sigma}^2_{100}$ becomes 0.7311. The error becomes

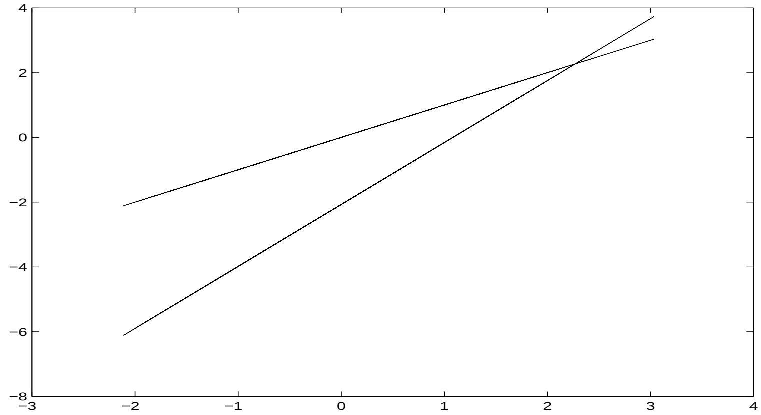
$$|b - \hat{b}_{100}| = 91.36\%$$

$$|a - \hat{a}_{100}| = 20.73\%$$

$$|\sigma - \hat{\sigma}_{100}| = 192.44\%$$

and the graph becomes

Figure4



All these tell us that the result would be unacceptable. But as we know, there is only one point which affects so much to the estimation of the parameters.

Let's recall Theorem 2 and Theorem 4. For the functions $f(x_*, y_*)$ and $h(x_*, y_*)$ have good properties as property (1) and property (2), maybe they can help us detect the bad point.

We try to find all the points' heights in graph(1) to find which point affects the estimation most and see what will happen if we delete it.

First, let us check whether the limited value can be used to estimate the effect of one point in a 100 point data set.

Without loss of generality, we check the $f(X_{100}, Y_{100})$ and $h(X_{100}, Y_{100})$in the the data set1.
By (2.4) we obtain $\hat{b}_{100} = 1.0203, \hat{b}_{99} = 1.0281$;
By (2.6) we obtain $\hat{\sigma}^2_{100} = 0.2596, \hat{\sigma}^2_{99} = 0.2602$.
Thus

$$\hat{b}_{100} - \hat{b}_{99} = -0.0078;$$

$$\hat{\sigma}^2_{100} - \hat{\sigma}^2_{99} = -0.0006$$

Via formula (2.8), we know the height of (X_{100}, Y_{100}) in graph1 is -0.7447, divided by 100, the weight of (X_{100}, Y_{100}), is -0.0074, quite close to the practical value -0.0078.
Via formula (2.10), we know the height of (X_{100}, Y_{100}) in graph2 is -0.0594, divided by 100, the weight of (X_{100}, Y_{100}), is -0.000594, also quite close to the practical value -0.0006.
These demonstrate that the Theorem 2 and theorem 4 can be used here.

Second, let the approximate value $\overline{X}_{(99)} = \frac{\sum_{i \neq j}^{100} X_i}{99}$ replace the E(x)in (2.5), and let the approximate value $\overline{X^2}_{(99)} - (\overline{X}_{(99)})^2 = \frac{\sum_{i \neq j}^{100} X_i^2}{99} - (\frac{\sum_{i \neq j}^{100} X_i}{99})^2$ replace the $Cov(x)$ in (2.8) and (2.10). We can figure out the value of $f(X_j, Y_j)$ and $h(X_j, Y_j)$ (see table 4 to find the data, see program2 to find the program) , then we can plot the graph of $f(X_j, Y_j)$ and $g(X_j, Y_j)$:

Figure5

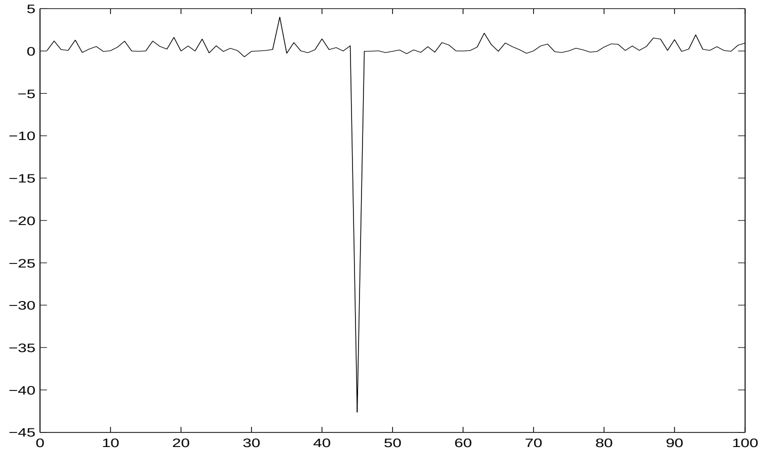
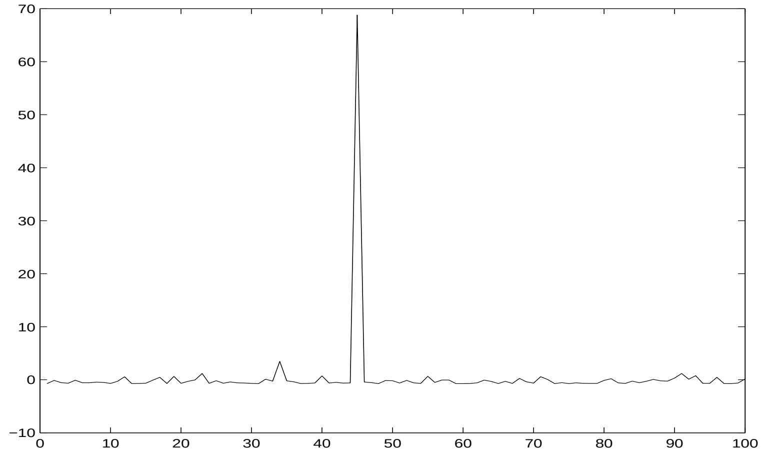


Figure6



From the graphs above, we see clearly that the 45th point is an outlier: it affects both the estimation of b and σ^2 a lot. Next, let us delete this point and find new estimations for the parameters. We use the data set $\{(X_i, Y_i) : i \neq 45\}$ to find $\hat{b}_{99}$, $\hat{a}_{99}$ and $\hat{\sigma}^2_{99}$:

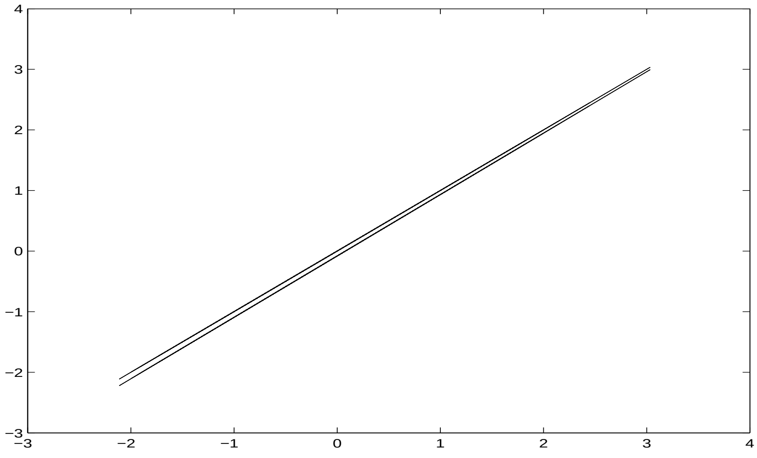
$$\hat{b}_{99} = 1.0134$$
$$\hat{a}_{99} = -0.0811$$
$$\hat{\sigma}^2_{99} = 0.2543$$

Again, a quite good estimation was derived. we see that errors are small and the estimated line fits the real line quite well:

$$|b - \hat{b}_{100}| = 1.34\%$$
$$|a - \hat{a}_{100}| = 8.11\%$$
$$|\sigma - \hat{\sigma}_{100}| = 1.72\%$$

We compare the real line and the estimated line in the graph7:

Figure7



Till now, we have already seen how Theorem 2 and Theorem 4 can help us when there is an outlier. One may want to ask what will happen if all the points were good.

Let us also give the graphes of $f(X_j, Y_j)$ and $h(X_j, Y_j)$ of the unpolluted data set(data set 1)(see table5 to find the value of f and h):

Figure8

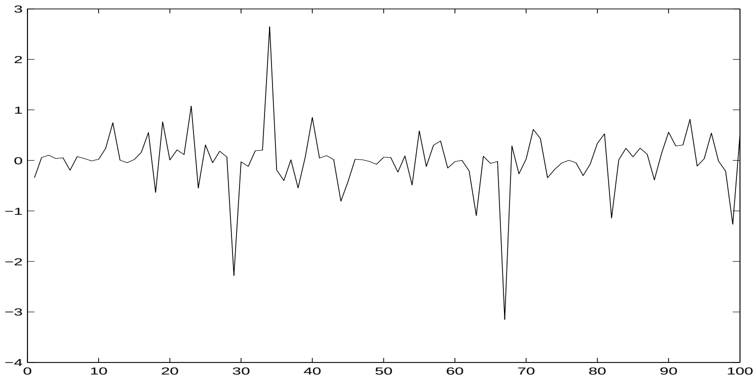
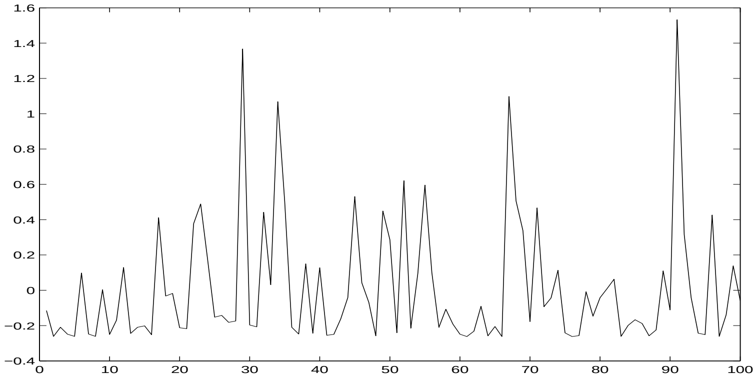


Figure9



From the graphs we see all the points are distributed uniformly and all are close to zero. Also, direct computation yields the average of $f(X_j, Y_j)$ is -1.3600e-004 and the variance is 0.3653; the average of $h(X_j, Y_j)$ is 0.0080 and the variance is 0.1259.

The numbers demonstrate that all the points in the data set1 are good enough to be used to estimate the parameters, and in fact, as we have already seen above((4.1),(4.2), (4.3)and graph3), we surely have gotten a good estimation using this data set without deleting any point.

From all those we listed above, we see that when we use this method to estimate the parameters in model(1.1), it may not "the more the better". We should first explore whether all the data in the data set are "good". When one point itself affects the estimation a lot, maybe it is an outlier, and perhaps we would obtain a good result after deleting it.

4.5 Conclusion of the simulation:

From the simulation, we see the method is good concerning the model when the data set is good. The theorems can be used to analyze practical data. The influence functions provide us with a diagnostic method and we can obtain better estimations after deleting the outlier which influences the most.

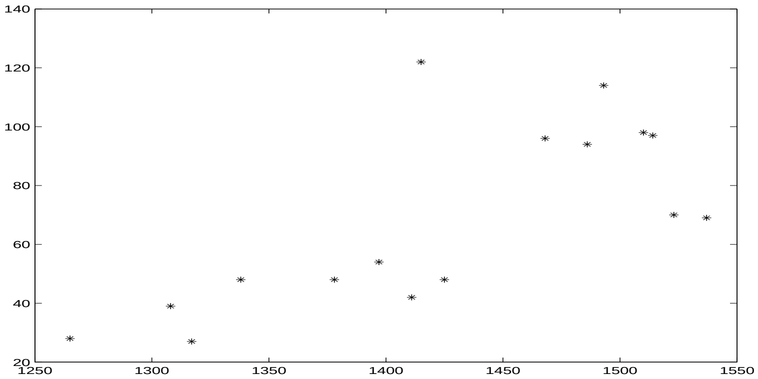
5 Application

SARS, an atypical pneumonia of unknown aetiology, was recognized at the end of February 2003. In late March,2003, it came to Beijing. Unfortunately, more and more people are infected by it and many people are suspected to be infected.

The cumulative SARS cases are increasing everyday, what's more, there are still a large number of people that are suspected to be infected. so people living in Beijing feel more and more anxious about the illness. But maybe after we analyze the probability of the transformation from a suspected case to a real SARS case, we will feel some what released.

We consider the numbers of those after May 1. The data are as follows: (coming from the Ministry of Health P.R.China):
New SARS Cases:
[122 96 114 69 98 70 97 94 48 54 42 48 48 39 27 28];
cumulative Doubted Cases:
[1415 1468 1493 1537 1510 1523 1514 1486 1425 1397 1411
1378 1338 1308 1317 1265].

Figure10



The cumulative suspected cases and the new SARS cases in Beijing can be described by Model (1.1). As it is a special practical problem, we may not follow the step above. First, we should explore the "a". Since it describes the number of new SARS cases when the suspected case is 0, we can suppose $a = 0$ by our experience. After we fix the parameter a, we can compute the estimations as follows:

$$\hat{b} = 0.0501$$

$$\hat{\sigma^2} = 714.2123$$

That is to say, the probability of the transformation from a suspected case to a real SARS case is 5.01%.

It is a big number: many of us don't hope it happen. We try to know weather all the data support this result or weather there's some outlier that made the result not true:

let us plot out the graph of $f(X_i, Y_i)$ and $h(X_i, Y_i)$:

Figure12

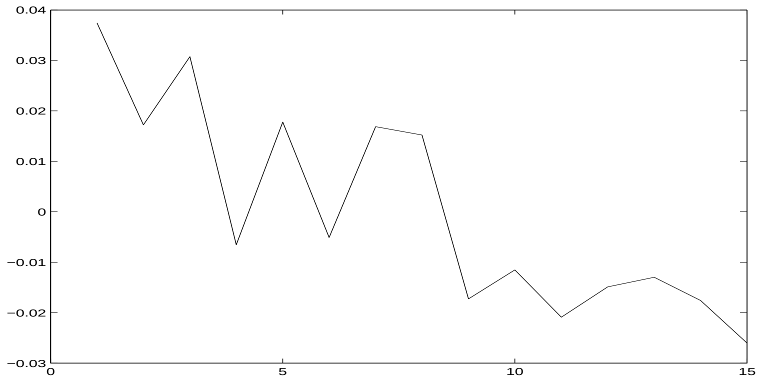
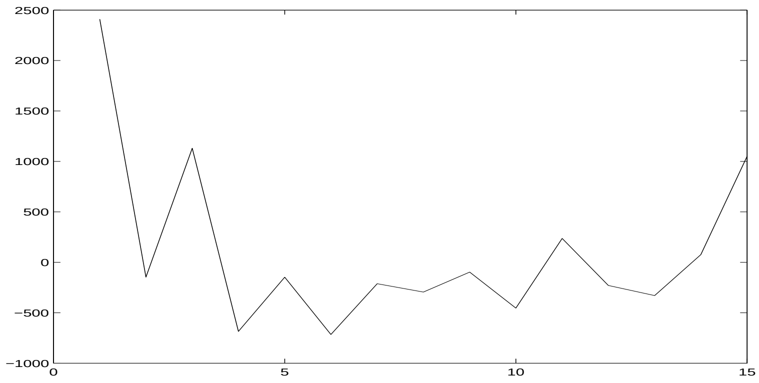


Figure13



We use one sentence from WHO to explain why there may be an outlier : "As SARS is a diagnosis of exclusion, the status of a reported case may change over time. This means that previously reported cases may be discarded after further investigation and follow-up." From the graphs, we find that the 1-st point is an outlier. We delete it and compute the estimations again, we obtain:

$$\hat{b} = 0.0476$$

$$\hat{\sigma^2} = 566.4116$$

This result may be closer to the real rule of this disease. It tell us that the rate of the transformation from a suspected case to a real SARS case is even smaller than it seems, and in fact, we could see from the data set that the rate is not big especially lately, for it is still changing, we should update our model frequently and we can fully understand the nature of it finally.

Thus, we could conclude that as a person living in Beijing, we should not worry too much about the seemingly large number of the suspected cases, the rate of the transition of them to a real SARS case is very small. So, just do not feel too worried, but of course you should always take good care of yourself!

Acknowledgement The author is grateful to Professor Hengjian Cui who enlightened her to write down this paper.

References

[1] Hengjian CUI & Rongcai LI, On Parameter Estimation for Semi-Linear Errors-in-Variables model. *J.Multivarate Anal.,*1-64, 1998. page 1-24

[2] W.A.Fuller, “Measurement Error Models”, *John Wiley&Sons* New York, 1987.

[3] Shouzheng Tang& Yong Li , “The statistical basis for Bio-Mathematical Model”, *Science Press*, 2002.

[4] Hengjian Cui& Songxi Chen, Empirical Likelihood Confidence Region for Parameter in the Errors-in-Variables Models, *J. Multivariate Anal.*,84 (1), 2003.page 101-115.

[5] David C. Hoaglin, Frederick Mosteller & John W. Tukey, “Understanding Robust and Exploratory Data Analysis”, *China Statistical Press*, 1998.

[6] Kaitai Fang & Jianlun Xu, “Statistical Distribution”, *Science Press*, 1987.

[7] Xuan Lu, “Elementary Mathematical Statistics”, *Tsinghua University Press*, 1998.

[8] Zongshu Wei, “Probability and Mathematical Statistics”, *Higher Education Press*, 1983.

[9] Xiaoting Chang & Kaitai Fang, “Multivariate Statistical Analysis”, *Science Press*, 1999.

[10] Xuping Zhong, Guoming Meng, Haibin Wang & Bochenng Wei, influence Analysis on Nonlinear Measurement Error Models, *Chinese Journal of Applied Probability and Statistics* Vol. 19 No.1 Feb. 2003. page 31–39.

[11] Frank R. Hampel, Elvezio M.Ronchetti,Peter J.Rousseeuw and Werner A.Stahel, “Robust Statistics—-The Approach Based On Influence Functions”, *John Wiley&Sons* New York, 1986.

线性测量误差模型的影响分析

摘要

本文讨论了线性测量误差模型的参数估计, 对由正交回归方法得到的参数估计a, b, σ^2 做出了影响分析,给出了影响函数及其图形(b 的影响函数的图形为一个双曲抛物面, σ^2的影响函数图形为抛物柱面), 让大家可以直观地看到用这种方法进行参数估计的性质: 在数据集较好时(例如,样本都取自要分析的线性模型), 所得的参数估计值将较准确, 但这种方法不稳健, 即即使只有一个污染点也可能很大的影响估计值, 导致荒谬的结果. 我们同时也注意到影响函数及其图形可作为诊断及筛选数据的依据. 在对所得出的每一结论作了随机模拟检测后, 我们看到了这一估计方法的性质在真实数据集中的体现,也看到了我们的影响函数及其图形在诊断筛选数据中确实能起到很大的作用.

在证明本文主要结论的过程中, 我们先对一元有显式表达式的情况作了分析, 由显式表达式直接计算得出一系列结论, 在多元没有显式表达式的情况下, 我们换了另一种方法—-从泰勒展开式入手进行分析—-最后得出了同样的结论. 并用我们的两种方法验证了$Cui\&Li$(1998)的结果.

除了理论推导,本文还应用了随机模拟来验证结果, 我们把我们的理论在随机产生的数据集上试验, 所得的结果都非常好的符合我们的结论. 在文章的最后, 我们给出了一个具体应用实力, 分析近日北京SARS 由疑似转为临床确诊的转变率, 对本文结论的应用作了示范, 并同时帮助北京民众更好的认识SARS.

关键词：测量误差, 正交回归, 线性模型, 局部影响.