# 二值回归模型中自变量误差的处理*

郑 忠 国
（北京大学）

## §1 引 言

在回归问题中，我们所讨论的主要问题是自变量与应变量之间的关系问题．在某些实际问题中，因变量是二值的．例如在讨论冠心病的成因的问题中，因变量可以写成

$$Y=\begin{cases}1, & \text{受试者一年后发展成冠心病,}\\ 0, & \text{受试者一年后仍然健康.}\end{cases} \tag{1.1}$$

致病因子与发病率的关系成为当前讨论的中心问题。在〔1〕中，讨论了若干参数模型。设$X$是$R^p$中的一个向量，它代表可能的致病因子，发病率是$P\{Y=1\mid X\}$．设$G$是一个已知的一元分布，通常使用下述模型

$$P\{Y=1\mid X=x\}=G(\beta' x) \tag{1.2}$$

来表示发病率与致病因子的关系，其中$\beta$是未知的参数向量．这就是所谓二值回归模型．（1.2）中的$G$通常是下列函数之一，

$$\begin{aligned} &G(a)=(1+\exp(a))^{-1} && (logistic\ \text{回归}),\\ &G(a)=\Phi(a) && (probit\ \text{回归}),\end{aligned} \tag{1.3}$$

其中$\Phi(a)$是标准正态分布函数．在（1.3）式的$logistic$回归中，$G$不是一个分布函数，但这并不影响问题的本质．本文讨论$logistic$回归的一种比较重要的但是比较简单的情况，即

$$P\{Y=1\mid X=x\}=G(a_0+a_1 x) \tag{1.4}$$

其中$X$是一个数值随机变量．假定$X$具有已知的分布函数$F(x)$．又设$(X_1,Y_1),\cdots(X_n,Y_n)$的一组$i.i.d$观察值．本文中讨论变量具有误差的情况．当$X_i$在某区域以外时，其测量值往往是不准确的．在本文中我们假定$X_i$的观察值当超过指定的$A$值时，其观察值就不可靠了．为解决这种自变量的误差处理问题，我们采用$M$－估计的方法，求得参数$a=(a_0,a_1)'$的估计．

令$\psi_j(i,x,a)$，$i=0,1$，$j=1,2$，是一组函数．我们假定当自变量$x\geq A$时，$\psi_j(i,x,a)$与$x$无关．对于$\psi_i$，考虑下面的关于$a$的方程组

$$\begin{aligned}&\sum_{i=1}^{n}(\psi_j(Y_i,X_i,a)-\psi_j(1,X_i,a)G_a(X_i)-\psi_j(0,X_i,a)(1-G_a(X_i)))I_{(X_i<A)}\\ &+\sum_{i=1}^{n}(\psi_j(Y_i,X_i,a)-\psi_j(1,X_i,a)\int_{x\geq A}G_a(x)\mathrm{d}F(x)/\int_{x\geq A}\mathrm{d}F(x)\\ &-\psi_j(0,X_i,a)\int_{x\geq A}(1-G_a(x))\mathrm{d}F(x)/\int_{x\geq A}\mathrm{d}F(x))I_{(X_i\geq A)}=0,\quad j=1,2,\end{aligned} \tag{1.5}$$

*1983年5月5日收到．

上式中$G_a(x)\triangleq G(a_0+a_1x)$．方程组（1.5）的解$\hat{a}_n$称为$a$的$M$－估计。方程组（1.5）的解只与$\psi_j(1,x,a)-\psi_j(0,x,a)$有关．事实上，令

$$\psi_j(x,a)=\psi_j(1,x,a)-\psi_j(0,x,a),\quad j=1,2. \tag{1.6}$$

则方程组（1.5）等价于

$$\sum_{i=1}^{n}\psi_j(x_i,a)I_{Y_i=1}-G_a(x_i)I_{x_i\geq A}$$

$$+\sum_{i=1}^{n}\psi_j(x_i,a)(I_{Y_i=1}-\int_{x\geq A}G_a(x)\mathrm{d}F(x)/\int_{x\geq A}\mathrm{d}F(x))I_{(x_i\geq A)}=0,\ j=1,2, \tag{1.7}$$

在方程组（1.7）中，当$X_i\geq A$时，由于$\psi_j(X_i,a)=\text{const}$，即使$X_i$的值受到误差干扰，方程组（1.7）并不受干扰的影响．这说明（1.7）具有抗干扰作用．

在§2中，我们证明了（1.7）的相合解的渐近正态性，并且求得了最佳的$\psi$函数．在§3中，我们将利用一步牛顿法求出（1.7）式的近似解并且研究该近似解的优良性。

## §2 解的渐近正态性和最佳$\psi$函数

设$\mu_a$是$(-\infty,+\infty)$上的测度，由下式确定

$$\mathrm{d}\mu_a=G_a(x)(1-G_a(x))\mathrm{d}F \tag{2.1}$$

$$\text{记 } B=\begin{pmatrix}\int_{R^1}\psi_1(x,a)^2\mathrm{d}\mu_a(x) & \int_{R^1}\psi_1(x,a)\psi_2(x,a)\mathrm{d}\mu_a(x)\\ \int_{R^1}\psi_1(x,a)\psi_2(x_1,a)\mathrm{d}\mu_a(x) & \int_{R^1}\psi_2(x,a)^2\mathrm{d}\mu_a(x)\end{pmatrix}-\Delta_{11}\begin{bmatrix}d_1^2 & d_1d_2\\ d_2d_1 & d_2^2\end{bmatrix} \tag{2.2}$$

其中

$$\Delta_{11}=\int_{(A,+\infty)}G_a^2(x)\mathrm{d}F(x)-(\int_{(A,+\infty)}G_a(x)\mathrm{d}F(x))^2/\int_{(A,+\infty)}\mathrm{d}F(x) \tag{2.3}$$

$$d_i=\psi_i(A,a). \tag{2.4}$$

$$\text{记 } C=\begin{pmatrix}\int_{R^1}\psi_1(x,a)\mathrm{d}\mu_a(x) & \int_{R^1}\psi_1(x,a)x\mathrm{d}\mu_a(x)\\ \int_{R^1}\psi_2(x,a)\mathrm{d}\mu_a(x) & \int_{R^1}\psi_2(x,a)x\mathrm{d}\mu_a(x)\end{pmatrix} \tag{2.5}$$

**定理1** 设$\psi_i(x,a)$，$i=1,2$，满足

$$\lim_{a\to a^{(0)}}\frac{\partial\psi_i(x,a)}{\partial a_j}=\frac{\partial\psi_i(x,a)}{\partial a_j}\Big|_{a=a^{(0)}},\ j=0,1,\ a^{(0)}\in R^{(2)}, \tag{2.6}$$

并且上述极限是对$x\in R^1$一致收敛的．又设$\psi_i$使B，C的元素为有限，并且$|C|\neq 0$．则方方程组（1.7）的相合解$\hat{a}_n$满足

$$\sqrt{n}(\hat{a}_n-a)\to N(0,C^{-1}BC^{-1}). \tag{2.7}$$

**证明** 记

$$U_j(X,Y,a)=\psi_j(X,a)(I_{(Y=1)}-G_a(X))I_{(X<A)}+$$

$$+\psi_j(X,a)(I_{(Y=1)}-\int_{u\geq A}G_a(u)\mathrm{d}F(u)/\int_{u\geq A}\mathrm{d}F(u))I_{(X\geq A)} \tag{2.8}$$

利用这样的记号，方程（1.7）变成

$$\sum_{i=1}^{n}U_j(X_i,Y_i,a)=0,\quad j=1,2. \tag{2.9}$$

易证，当$a$为参数的真值时，$U=(U_1,U_2)'$的协方差阵为B．并且$EU_i=0$．利用中心极限定理可得

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}U(X_i,Y_i,a)\xrightarrow{\mathcal{L}}N(0,\mathrm{B})\,. \tag{2.10}$$

由于$\frac{\partial\psi_j(x,a)}{\partial a_k}$，$k=0,1$，对所有的$x$一致地是$a$的连续函数，$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial U_j(X_i,Y_i,a)}{\partial a_k}$对所有的$(X_i,Y_i)$，$i=1,\cdots,n$的值一致地是$a$的连续函数．这样，利用中值定理，可得

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}U(X_i,Y_i,a)=\frac{1}{\sqrt{n}}\sum(U(X_i,Y_i,a)-U(X_i,Y_i,\hat{a}_n))=-T_n\sqrt{n}(\hat{a}_n-a)$$

其中

$$T_n=\begin{pmatrix}\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial a_0}U_1(X_i,Y_i,a)\Big|_{a_n^*} & \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial a_1}U_1(X_i,Y_i,a)\Big|_{a_n^*}\\ \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial a_0}U_2(X_i,Y_i,a)\Big|_{a_n^{**}} & \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial a_1}U_2(X_i,Y_i,a)\Big|_{a_n^{**}}\end{pmatrix}\xrightarrow{P}\mathrm{C}\,.$$

在上式中$a_n^*$，$a_n^{**}$是在线段$(a,\hat{a}_n)$上的点．再利用Slutsky定理可得(2.7)．证毕．

现在我们需要求得最佳的$\psi$函数，亦就是求$\psi_1,\psi_2$，使得对应的$\mathrm{C}^{-1}\mathrm{BC}^{-1\prime}$达到极小．

**定理2** 令

$$I_1(x,a)=\begin{cases}1, & x\in(-\infty,A)\\ \int_{[A,+\infty)}\mathrm{d}\mu_a/(\int_{[A,+\infty)}\mathrm{d}\mu_a+\Delta_{11}), & x\in[A,+\infty)\end{cases}$$

$$I_2(x,a)=\begin{cases}x, & x\in(-\infty,A)\\ \int_{[A,+\infty)}x\mathrm{d}\mu_a(x)/(\int_{[A,+\infty)}\mathrm{d}\mu_a(x)+\Delta_{11}), & x\in[A,+\infty)\end{cases}$$

若以$I_1(x,a)$，$I_2(x,a)$作为$\psi$函数，则它们使定理1中的渐近正态协差阵达到极小．

**证明** 考虑定义在$(-\infty,A]$上的测度$v_a$．在区间$(-\infty,A)$上，$v_a$由下式定义

$$\mathrm{d}v_a\triangleq\mathrm{d}\mu_a=\frac{e^{a_0+a_1x}}{(1+e^{a_0+a_1x})^2}\mathrm{d}F(x)$$

在$A$点，$v_a(A)=\int_{[A,+\infty)}\mathrm{d}\mu_a+\Delta_{11}$．如此定义之后，矩阵B可以写成

$$\begin{pmatrix}\int\psi_1(x,a)^2\mathrm{d}v_a(x) & \int\psi_1(x,a)\psi_2(x,a)\mathrm{d}v_a(x)\\ \int\psi_2(x,a)\psi_1(x,a)\mathrm{d}v_a(x) & \int\psi_2(x,a)^2\mathrm{d}v_a(x)\end{pmatrix}$$

上式中积分是在$(-\infty,A]$上求的．同时，矩阵C可写成

$$\begin{pmatrix}\int\psi_1(x,a)I_1(x,a)\mathrm{d}v_a(x) & \int\psi_1(x,a)I_2(x,a)\mathrm{d}v_a(x)\\ \int\psi_2(x,a)I_1(x,a)\mathrm{d}v_a(x) & \int\psi_2(x,a)I_2(x,a)\mathrm{d}v_a(x)\end{pmatrix}$$

利用推广的Schwarz不等式，可得

$$\mathrm{C}^{-1}\mathrm{BC}^{-1\prime}\geqslant\begin{pmatrix}\int I_1^2(x,a)\mathrm{d}v_a(x) & \int I_1\cdot I_2\mathrm{d}v_a(x)\\ \int I_2\cdot I_1\mathrm{d}v_a(x) & \int I_2^2\mathrm{d}v_a(x)\end{pmatrix}$$

并且等号在$\psi_1=I_1$，$\psi_2=I_2$时达到．

## §3 用一步牛顿法求近似解

在(1.7)中令$\psi_j(x_i,a)=I_j(x_i,a)$，我们得到关于$a$的方程组

$$\sum_{i=1}^{n}I_{(X_i<A)}(I_{(Y_i=1)}-G_a(X_i))$$

$$+\frac{\sum_{i=1}^{n}\int_{[A,+\infty)}G_a(x)(1-G_a(x))\mathrm{d}F(x)(I_{\{Y_i=1\}}-\int_{[A,+\infty)}G_a(x)\mathrm{d}F(x)/\int_{[A,+\infty)}\mathrm{d}F(x))I_{\{X_i\geq A\}}}{\int_{[A,+\infty)}G_a(x)(1-G_a(x))\mathrm{d}F(x)+\Delta_{11}}$$

$$=0,$$

$$\sum_{i=1}^{n}X_iI_{\{X_i<A\}}(I_{\{Y_i=1\}}-G_a(X_i))$$

$$+\frac{\sum_{i=1}^{n}\int_{[A,+\infty)}xG_a(x)(1-G_a(x))\mathrm{d}F(x)(I_{\{Y_i=1\}}-\int_{[A,+\infty)}G_a(x)\mathrm{d}F(x)/\int_{[A,+\infty)}\mathrm{d}F(x))I_{\{X_i\geq A\}}}{\int_{[A,+\infty)}G_a(x)(1-G_a(x))\mathrm{d}F(x)+\Delta_{11}}$$

$$=0. \tag{3.1}$$

欲求得方程组（3.1）的解是很困难的．为此，我们需要求出（3.1）的近似解．首先，求解方程组

$$\sum_{i=1}^{n}I_{\{X_i<A\}}(I_{\{Y_i=1\}}-G_a(X_i))=0,\quad \sum_{i=1}^{n}X_iI_{\{X_i<A\}}(I_{\{Y_i=1\}}-G_a(X_i))=0. \tag{3.2}$$

$$\text{记 }\tilde{a}_n=\begin{cases}\begin{bmatrix}\tilde{a}_1\\ \tilde{a}_2\end{bmatrix},\text{若}\begin{bmatrix}\tilde{a}_1\\ \tilde{a}_2\end{bmatrix}\text{为（3.2）的唯一解；}\\ \begin{bmatrix}0\\ 0\end{bmatrix},\text{其它情况．}\end{cases} \tag{3.3}$$

**引理** 设$F$满足$\int|x|\mathrm{d}F<+\infty$，$F(-\infty,A)\in(0,1)$，并且$F$在$(-\infty,A)$上不集中于一点，则由（3.3）所确定的$\tilde{a}_n$是$a$的真值$a^{(0)}$的相合估计．

**证明** 基于引理中关于$F$的条件，不妨假设$X_1,\cdots,X_n$中至少有两个不同的点落在$(-\infty,A)$上．记

$$f_n(a)=\begin{pmatrix}f_{n1}(a)\\ f_{n2}(a)\end{pmatrix},\text{ 其中}$$

$$f_{n1}(a)=\frac{1}{n}\sum_{i=1}^{n}I_{\{X_i<A\}}/(1+e^{a_0+a_1X_i}),\quad f_{n2}(a)=\frac{1}{n}\sum_{i=1}^{n}X_iI_{\{X_i<A\}}/(1+e^{a_0+a_1X_i}).$$

容易验证，对固定的$X_i=x_i$，$i=1,\cdots,n$，$f_n(a)$是$\mathbf{R}^2$到$\mathbf{R}^2$内某开子集的一一连接续变换．并且对几乎所有的样本

$$f_n(a)\to f(a),\quad a\in\mathbf{R}^2,\quad (n\to+\infty). \tag{3.4}$$

并且对每一个$\mathbf{R}^2$的有界子集，上述的收敛是对$a$一致的．在（3.4）中的$f(a)$由下式确定

$$f(a)=\begin{pmatrix}f_1(a)\\ f_2(a)\end{pmatrix},\quad f_1(a)=\int_{x<A}\frac{dF(x)}{1+e^{a_0+a_1x}},\quad f_2(a)=\int_{x<A}\frac{x\mathrm{d}F(x)}{1+e^{a_0+a_1x}}.$$

$$\tilde{f}_n=\begin{pmatrix}\tilde{f}_{n1}\\ \tilde{f}_{n2}\end{pmatrix},\ \tilde{f}_{n1}=\frac{1}{n}\sum_{i=1}^{n}I_{\{X_i<A\}}I_{\{Y_i=1\}},\ \tilde{f}_{n2}=\frac{1}{n}\sum_{i=1}^{n}X_iI_{\{X_i<A\}}\cdot I_{\{Y_i=1\}}.$$

容易验证

$$\tilde{f}_n\to f(a^{(0)}),\ a.e.\ (n\to+\infty) \tag{3.5}$$

由（3.4），（3.5）可知，对几乎所有的样本，方程式（3.2）当$n$充分大时，存在唯一解．

又由于几乎所有的样本，$f_n(a)\to f(a)$ 在$a$的一个有界邻域内一致收敛，当$\tilde{f}_n\to f(a^{(0)})$时，方程$f_n(a)=\tilde{f}_n$之解必收敛于$a^{(0)}$．但上述方程即（3.2）．证毕．

由于$a_n$是相合的，再利用定理1，可得$\sqrt{n}(a_n-a^{(0)})=\ _p(1)$．

**定理3** 设$\tilde{a}_n$由（3.3）给出，$a_n$由下式解出

$$\frac{1}{n}T_n(\tilde{a}_n)\cdot\sqrt{n}(\hat{a}_n-\tilde{a}_n)=-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}U(X_i,Y_i,\tilde{a}_n) \tag{3.6}$$

其中$U=(U_1,U_2)'$由（2.8）所给出，$T_n(\tilde{a}_n)$由下式给出

$$T_n(a_n)=\begin{pmatrix}\sum_{i=1}^{n}\frac{\partial U_1}{\partial a_0} & \sum_{i=1}^{n}\frac{\partial U_1}{\partial a_1}\\ \sum_{i=1}^{n}\frac{\partial U_2}{\partial a_0} & \sum_{i=1}^{n}\frac{\partial U_2}{\partial a_1}\end{pmatrix}_{\tilde{a}_n}. \tag{3.7}$$

（在（2.8）的表达式中，$\psi_j$，$j=1,2$,用最佳函数$I_j$，$j=1,2$代替）．若$F(x)$满足$\int x^2\mathrm{d}F(x)<+\infty$和引理中所述的条件，则

$$\sqrt{n}(\hat{a}_n-a^{(0)})\to N(0,I(a^{(0)})^{-1}) \tag{3.8}$$

其中$a^{(0)}$是参数$a$的真值，

$$I(a^{(0)})=\begin{pmatrix}\int L_1^2(x,a^{(0)})\mathrm{d}v_{a^{(0)}}(x) & \int I_1\cdot I_2\mathrm{d}v_{a^{(0)}}(x)\\ \int I_2\cdot I_1\mathrm{d}v_{a^{(0)}}(x) & \int I_2^2\mathrm{d}v_{a^{(0)}}(x)\end{pmatrix} \tag{3.9}$$

**证明** 考虑函数$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}U(X_i,Y_i,a)$，其中$a$为真值$a^{(0)}$附近的任意一点．将它在$a^{(0)}$处展开，得

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}U(X_i,Y_i,a)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}U(X_i,Y_i,a^{(0)})=\frac{1}{\sqrt{n}}T_n(a^{(0)})(a-a^{(0)})+R_n(a) \tag{3.10}$$

上式中$R_n(a)$是一个向量．由于函数数$U_j(X,Y,a)$具有如下的特性：存在$a^{(0)}$的以$\varepsilon$为半径的球$S_\varepsilon(a^{(0)})$，使得

$$\sup_{a\in S_\varepsilon(a^{(0)})}\left|\frac{\partial^2}{\partial a_i\partial a_k}U_j(X,Y,a)\right|\le\varphi(X,Y),\quad i,k=0\text{或}1. \tag{3.11}$$

其中$\varphi(X,Y)$满足$E_{a^{(0)}}|\varphi(X,Y)|<+\infty$．由此可知，(3.10)中的余项可以写成

$$R_n(a)=|a-a^{(0)}|^2\cdot O_p(1),$$

上式中$O_p(1)$表示一个向量，其对应的分量全是$O_p(1)$．这样（3.6）式变成

$$\frac{1}{n}T_n(\tilde{a}_n)\cdot\sqrt{n}(\hat{a}_n-a^{(0)})=-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}U(X_i,Y_i,a^{(0)})$$

$$+\left(\frac{1}{n}T_n(\hat{a}_n)-\frac{1}{n}T_n(a^{(0)})\right)\left(\sqrt{n}(\tilde{a}_n-a^{(0)})\right)+|\tilde{a}_n-a^{(0)}|^2O_p(1).$$

利用（3.11），上式右边后两项为$o_p(1)$．同时由于

$$\frac{1}{n}T_n(\tilde{a}_n)=I(a^{(0)})(1+o_p(1)),$$

$$-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}U(X_i,Y_i,a^{(0)})\xrightarrow{\mathcal{L}}N(0,\ I(a^{(0)})),$$

利用Slutsky定理可知$\sqrt{n}(\hat{a}_n-a^{(0)})\to N(0,\ I(a^{(0)})^{-1})$．证毕．

## 参 考 文 献

[1] R. Carroll, C. Spiegelman, K. K. G. Lan, K. T. Bailey & R. D. Abbott, Errors-in-variables for binary regression models. Institute of Statistics mimeo series #1507, Uni. of North Carolina, Chapel Hill.

# Errors in variables in binary regression models (Summary)

*Zheng Zhongguo*

(Peking University)

### Abstract

In this paper the binary regression model, in which the response $Y$ is binary, *i.e.* $Y$ takes the value 0 or 1 only, is considered. The conditional probability $P\{Y=1 \mid X=x\}$ is assumed to be of the form $G(a_0+a_1x)$, where $x$ is the measurement of the factor, $a=(a_0,a_1)'$ is the vector of the unknown parameter being estimated and $G(a)=(1+e^a)^{-1}$. The measurement $x$ is assumed to be contaminated when the value of $x$ is out of certain limit. M-estimation is employed to estimate the unknown $a$ of the model. In this paper, the consistent M-estimation is proved to be asymptotically normal and the optimal M-equation is obtained. Finally, a computation method is introduced to solve the M-equation and to get the optimal solution.