ON A MODEL IN PROBIT ANALYSIS

BY MASAAKI SIBUYA

(Received Dec. 22, 1961)

1. The problem

Let (X, Y) be a bi-variate random variable. Assume that an observation on X is a quantal response datum, that is, we cannot observe its value directly but only whether or not it is larger than a constant s (say), which we have given in advance. Further, assume that another component Y is real-valued and may be measured only when X is larger than s. Our problem is to estimate, in such a situation, the parameters of the distribution of (X, Y). We give at first some examples which may occur in practice.

Example 1. An electric detonator consists of two parts, that is, the initiating explosive around an electric bridge and the larger charge of sensitive high explosive. Excited by electric current through the bridge, the initiating explosive begins to burst, and next the larger charge is fired by flying sparks. The excitation time X, the time necessary for the ignition, and the bursting time Y, the time interval from the beginning to the burst, are main characteristic values of detonator. We cannot measure the excitation time directly, but we see only whether or not the detonator bursts when the circuit is closed during a given time, and at the same time for the fired detonators we can observe the bursting time. In this case the component Y is observed when the component X is less than s.

Example 2. In bioassay we have a problem to study the relation between the response of animals to some poison and the weight (or some other characteristic) of some internal organ. In this case we can measure the weight of the internal organ only for animals which died by the poison of dose s.

Example 3. Electric engineers may concern about the relation between the dielectric breakdown strength of material X and its other characteristic after being examined by a breakdown test Y. We can test whether the material is destroyed or not at given high voltage s.

In example 2 and 3 the parameters of the distribution of (X, Y) may change according to the value s which we have given. Such a case is refered to in the last section.

A generalization of probit analysis in another direction, the study of the case where the response may take more than two "states" according to the values of more than two stimuli, are given by J. Aitchison and S. D. Silvey [2].

2. The model

Let the simultaneous distribution of (X, Y) be denoted by $F(x, y; \theta)$. The vector-valued parameter θ are, in many cases, composed of three parts;

$$\theta=(\alpha, \beta, \gamma) ,$$

where α and β are respectively the parameters for the marginal distribution function of X and Y, and γ is the parameter expressing the dependence between X and Y. That is, we may write as

$$G(x; \alpha)=F(x, \infty; \theta)$$

and

$$H(y; \beta)=F(\infty, y; \theta) .$$

We do not observe X directly but observe a random variable associated with X by the relation

$$\delta_s(X)=\begin{cases}1 & \text{if} \quad X\geqq s , \\ 0 & \text{if} \quad X<s ,\end{cases}$$

where s is a control variable. The probability that $\delta_s(X)$ takes the value 1 is equal to $p=1-G(s, \alpha)$. The component Y is observed only when $\delta_s(X)=1$.

The distribution function of Y under the condition that $X\geqq s$ is

$$H(y|s, \theta)=\{H(y; \beta)-F(s, y; \theta)\}/p ,$$

and the conditional probability density, if it exists, is

$$h(y|s, \theta)=\frac{\partial}{\partial y}H(y|s, \theta) .$$

We observe $n_1, n_2, \cdots, n_k$ trials at the levels $s_1, s_2, \cdots, s_k$ respectively.

Let $m_1, m_2, \cdots, m_k$ responses occur, at each level, and for these responses let the data $y_{11}, \cdots, y_{1m_1}; y_{21}, \cdots, y_{2m_2}; \cdots; y_{k1}, \cdots, y_{km_k}$ be obtained.

The probability density for these random variables is

$$\prod_{i=1}^{k}\binom{n_i}{m_i}p_i^{m_i}q_i^{n_i-m_i}\prod_{j=1}^{m_i}h(y_{ij}|s_i, \theta) ,$$

where

$$p_i=1-G(s_i, \alpha) ,$$

and its log-likelihood function is

$$L(\theta)=\sum_{i=1}^{k}\left\{c_i+(n_i-m_i)\log q_i+\sum_{j=1}^{m_i}\log l(y_{ij}|s_i, \theta)\right\} ,$$

where

$$l(y|s, \theta)=ph(y|s, \theta) .$$

From the assumption for θ, the maximum likelihood equations are

$$\sum_i \frac{n_i-m_i}{q_i}\frac{\partial q_i}{\partial\alpha}+\sum_{ij}\frac{1}{l(y_{ij})}\frac{\partial l(y_{ij})}{\partial\alpha}=0 ,$$

$$\sum_{ij}\frac{1}{l(y_{ij})}\frac{\partial l(y_{ij})}{\partial\beta}=0 ,$$

$$\sum_{ij}\frac{1}{l(y_{ij})}\frac{\partial l(y_{ij})}{\partial\gamma}=0 .$$

3. Bi-variate normal distribution

For the bi-variate normal distribution $N(\mu_x, \mu_y; \sigma_x^2, \sigma_y^2; \rho)$, with the probability density function $\phi(x, y; \theta)$, we have

$$l(y|s, \theta)=\frac{1}{\sigma_y}\phi(\eta)\left\{1-\Phi\left(\frac{\tau-\rho\eta}{\sqrt{1-\rho^2}}\right)\right\} ,$$

where

$$\tau=(s-\mu_x)/\sigma_x , \qquad \eta=(y-\mu_y)/\sigma_y ;$$

and

$$\phi(z)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right) , \qquad \Phi(z)=\int_{-\infty}^{z}\phi(t)dt ;$$

$$\theta=(\mu_x, \sigma_x^2; \mu_y, \sigma_y^2; \rho) .$$

Substituting the expressions

$$\left\{\begin{aligned}
&\frac{\partial l(y)}{\partial \mu_x}=\phi(s,y;\theta)\ ,\\
&\frac{\partial l(y)}{\partial \sigma_x}=\phi(s,y;\theta)\tau\ ,\\
&\frac{\partial l(y)}{\partial \mu_y}=l(y|s,\theta)\frac{\eta}{\sigma_y}-\phi(s,y;\theta)\frac{\rho\sigma_x}{\sigma_y}\ ,\\
&\frac{\partial l(y)}{\partial \sigma_y}=l(y|s,\theta)\frac{\eta^2-1}{\sigma_y}-\phi(s,y;\theta)\frac{\rho\sigma_x\eta}{\sigma_y}\ ,\\
&\frac{\partial l(y)}{\partial \rho}=\phi(s,y;\theta)\frac{\sigma_x(\eta-\rho\tau)}{1-\rho^2}\ ,
\end{aligned}\right.$$

into the maximum likelihood equations and rearranging them, we have the system of equations

$$\left\{\begin{aligned}
&\sqrt{1-\rho^2}\ \sum_i (n-m_i)N(\tau_i)=\sum_{i,j} M\Big(\frac{\tau_i-\rho\eta_{ij}}{\sqrt{1-\rho^2}}\Big)\ ,\\
&\sqrt{1-\rho^2}\ \sum_i (n_i-m_i)\tau_i N(\tau_i)=\sum_{i,j} \tau_i M\Big(\frac{\tau_i-\rho\eta_{ij}}{\sqrt{1-\rho^2}}\Big)\ ,\\
&\frac{\sqrt{1-\rho^2}}{\rho}\sum_{i,j}\eta_{ij}=\sum_{i,j} M\Big(\frac{\tau_i-\rho\eta_{ij}}{\sqrt{1-\rho^2}}\Big)\ ,\\
&\frac{\sqrt{1-\rho^2}}{\rho}\Big\{\sum_{i,j}\eta_{ij}^2-m\Big\}=\sum_{i,j}\eta_{ij} M\Big(\frac{\tau_i-\rho\eta_{ij}}{\sqrt{1-\rho^2}}\Big)\ ,\\
&\sum_{i,j}(\eta_{ij}-\rho\tau_i)M\Big(\frac{\tau_i-\rho\eta_{ij}}{\sqrt{1-\rho^2}}\Big)=0\ ,
\end{aligned}\right.$$

where

$$M(\xi)=\phi(\xi)/\{1-\Phi(\xi)\}\ ,\qquad N(\xi)=\phi(\xi)/\Phi(\xi)\ ,$$

$$\tau_i=(s_i-\mu_x)/\sigma_x\ ,\qquad \eta_{ij}=(y_{ij}-\mu_y)/\sigma_y\ ,\qquad m=\sum_{i=1}^k m_i\ .$$

To solve this system of equations, we have to take some complicated successive approximation process.

4. The moment method

For the bi-variate normal distribution, we have

$$E(Y|x)=\mu_y+\frac{\rho\sigma_y}{\sigma_x}(x-\mu_x)\ ,$$

$$V(Y|x)=(1-\rho^2)\sigma_y^2\ ,$$

and from these

$$E(Y|X\geqq s)=\mu_y+\rho\sigma_y M(\tau)\ ,$$

$$V(Y|X\geqq s)=\sigma_y^2-\rho^2\sigma_y^2 M(\tau)\{M(\tau)-\tau\}\ ,$$

where

$$M(\tau)=\phi(\tau)/\{1-\Phi(\tau)\}\ , \qquad \tau=(s-\mu_x)/\sigma_x\ .$$

It is easy to see that the conditional variance satisfies the inequality

$$(1-\rho^2)\sigma_y^2< V(y|X\geqq s)<\sigma_y^2\ .$$

If we disregard the supplemental variable Y, we can estimate μ_x and σ_x by the usual probit analysis technique (see e.g. [1]). We denote the estimates by $\hat\mu_x$ and $\hat\sigma_x$ and put $\hat\tau_i=(s_i-\hat\mu_x)/\hat\sigma_x$.

Computing the means $\bar y_i$ and the unbiased variance $s_{y_i}^2$ for the values at the i^{th} level $(i=1, 2, \cdots, k)$, and fitting straight lines to the points $(\bar y_i, M(\hat\tau_i))$ and $(s_{y_i}^2, M(\hat\tau_i)\{M(\hat\tau_i)-\hat\tau_i\})$, we can estimate the other parameters. As conditional variance of Y is large in comparison with conditional expectation, we have to take large sample to estimate ρ accurately.

TABLE 1

Data of one experiment

levels	s_i	-1.2	-0.8	-0.4	0	0.4	0.8	1.2	1.6
no. of trials	n_i	50	50	50	50	50	50	50	50
no. of obsv. x's	m_i	47	43	42	30	26	18	7	5
$\rho=.4$	y_i	$-.006$	.020	.180	.308	1.112	.566	.743	.510
	s_y^2	1.316	.739	.829	.653	.784	.758	.890	.278
$\rho=.7$	y_i	$-.016$	.060	.229	.494	.436	.878	1.112	.970
	s_y^2	1.092	.736	.713	.570	.668	.636	.626	.275
$\rho=.9$	y_i	$-.024$	.091	.248	.611	.690	1.068	1.328	1.288
	$s_{y_i}^2$	.902	.725	.535	.526	.484	.483	.330	.212

TABLE 2

Estimates of parameters

ρ No. of exp.	$\rho=.4$				$\rho=.7$				$\rho=.9$			
	I	II	III	IV	I	II	III	IV	I	II	III	IV
$\hat\mu_x$	$-.052$	$-.119$	$-.046$	.051	— Do —				— Do —			
$\hat\sigma_x$	.900	1.053	.959	1.026								
$\hat\mu_y$	$-.289$	$-.028$	$-.015$	$-.033$	$-.232$	$-.084$	$-.026$	.014	$-.150$	$-.127$	$-.034$	.056
$\hat\sigma_y^2$	.951	.748	1.241	.792	1.108	.780	1.114	.866	1.154	.839	1.020	.918
$\hat\rho_1$	.658	.267	.356	.456	.843	.668	.645	.743	.944	.939	.857	.918
$\hat\rho_2$	.071	0	.813	0	.761	0	.872	.295	.965	.668	.930	.868

$\hat\rho_1$ is based on $\widehat{\rho\sigma_y}$, and $\hat\rho_2$ on $\widehat{\rho^2\sigma_y^2}$. $\hat\rho_2=0$ means a negative value of $\widehat{\rho^2\sigma_y^2}$.

So far, we assumed that the parameters μ_y, σ_y, ρ were not affected by the values of s. Using the above moment method, however, we can estimate at least μ_y when it is a function of s, since $\rho\sigma_y$ appears in both regression equations.

For $N(0, 0, 1, 1, \rho)$, $\rho=0.4, 0.7, 0.9$, four experiments were made.

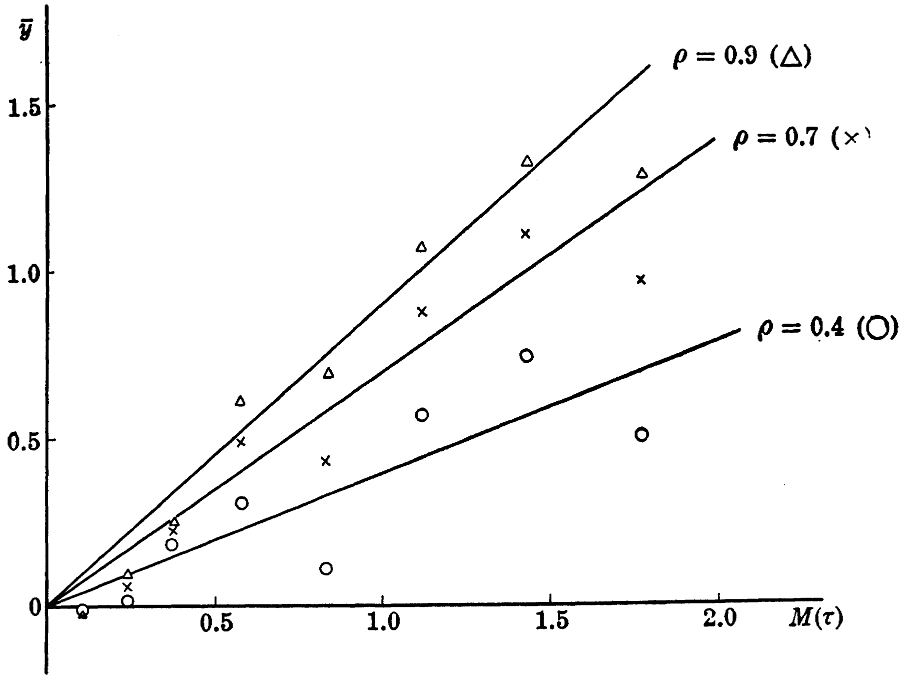

Fig. 1

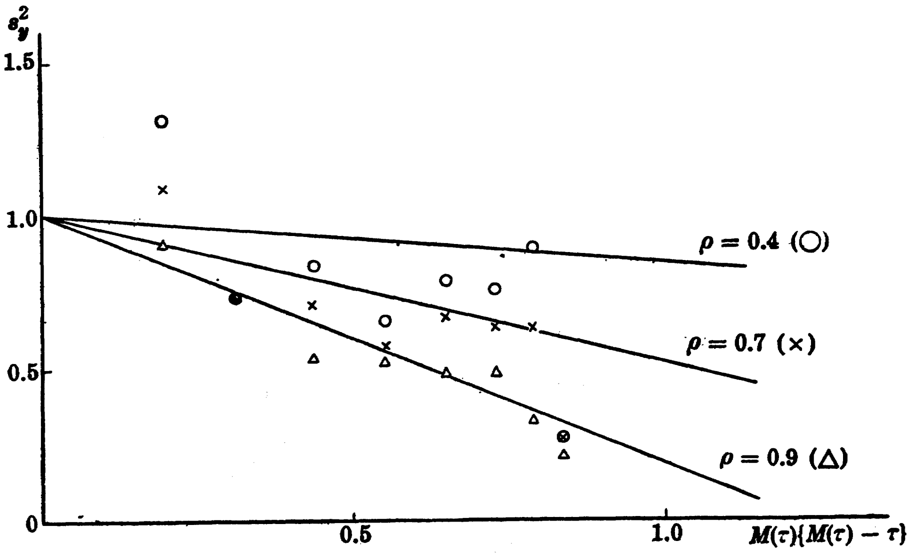

Fig. 2

Each of them consists of data such as in Table 1. For simplicity, the same values of x's and the same additional normal random variables were used to generate y's with different correlations. The results in Table 1 are shown also in Figure 1 and 2. Estimates of the population parameters are summarized in Table 2. Regression coefficients are estimated from a straight line fitted by the simple least square method, although some weighted least square may be preferable.

5. Acknowledgements

The author is grateful to Messrs. T. Okuno and T. Haga for useful discussions and to Miss. A. Maruyama for her aids in computation.

THE INSTITUTE OF STATISTICAL MATHEMATICS

REFERENCE

[1] E. J. Finney, *Probit analysis*, 2nd ed., 1952, Cambridge Univ. Press.

[2] J. Aitchison and S. D. Silvey, "The generalization of probit analysis to the case of multiple responses," *Biometrika*, Vol. 44 (1957), pp. 131-140.