ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 228, 2009

Andrzej Mantaj[*], *Wiesław Wagner*[**]

MODELS OF PROBABILITY FOR RANDOM VARIABLES OF BERNOULLI DISTRIBUTION

Abstract. In the paper, on the selected distributions of probability there were given models allowing for estimation of structural parameters by the generalized least squares method, by applying various types of function of the independent variable in models. It allowed a uniform perspective of the method of estimation of structural parameters and forecasting the value of parameter p, presented in the final part of the article.

Key words: Bernoulli distribution, probabilistic model, probit model, logit model.

I. INTRODUCTION

In economic research we meet situations of estimation of parameters of distribution of qualitative random variables. It particularly concerns random variables of Bernoulli distribution. The parameter of this distribution is probability p of occurrence of number of successes in a finite series of independent experiences. Shaping of the value of this parameter can depend on one or many established causes, i.e. on one or many independent variables. Taking into consideration the linear dependence in which particular components are functions of one considered independent variable, we obtain models allowing estimation of structural parameters of a model and forecasting the value of parameter p.

In the paper, on the selected distributions of probability there were given models allowing for estimation of structural parameters by the generalized least squares method, by applying various types of function of the independent variable in models. It allowed a uniform perspective of the method of estimation of structural parameters and forecasting the value of parameter p, presented in the final part of the article.

The basis of undertaken considerations were, among other things, the works of W. Ostasiewicz (1999), B. Guzik and W. Jurek (2000), M. Lipiec-Zajchowska (2003), and G.S. Maddala (2006).

[*] Ph.D., University of Information Technology and Management in Rzeszów.

[**] Professor, University of Information Technology and Management in Rzeszów.

II. GENERAL ASSUMPTIONS

Let there be a finite general population $\mathbf{P}$ having n statistical units, of which every unit takes two possible values 0 or 1 corresponding to occurrence of random events A' (failure) or A (success). The events A, A' create the complete set of random events. In population $\mathbf{P}$ there is determined random variable ξ expressing the number of successes (the appearances of A event in the series of n independent experiences) of Bernoulli distribution $\xi \sim B(n,p)$ such as $P(A) = p$ and $P(A') = 1 - p = q$. The very first moments of the distribution for ξ are: expected value $E(\xi) = np$ and variance $D^2(\xi) = npq$.

For ξ we determine the new random variable $\varepsilon = \frac{\xi}{n}$ expressing the fraction of occurrence of successes in population $\mathbf{P}$ with distribution: $E(\varepsilon) = p$ and $D^2(\varepsilon) = \frac{pq}{n}$.

Parameter p is unknown, it can be treated as a function h of established casual (interpreted) variable x. It is assumed that the function h expresses itself by a linear combination of the set $m + 1$ of structural parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_m)'$ and the function $g_j(x)$, $j = 0, 1, 2, \ldots, m$ of variable x, i.e. its form is

$$h(x, \boldsymbol{\beta}) = h(x; \beta_0, \beta_1, \ldots, \beta_m) = \sum_{j=0}^{m} \beta_j g_j(x) = \boldsymbol{\beta}' \mathbf{g}(x), \tag{1}$$

where $\mathbf{g}(x) = (g_0(x), g_1(x), \ldots, g_m(x))'$, but $g_0(x) \equiv 1$. Functions g are selected arbitrarily depending on the formulated research problem (e.g. $g_1(x) = x$, $g_2(x) = x^2$).

For general population $\mathbf{P}$ we assume the dependence $p = f[h(x, \boldsymbol{\beta})]$, i.e. parameter p is treated as a dependent variable expressed by a function f, the one which would take values of the range $(0,1)$ and would be determined at the set $\Theta = (x, \boldsymbol{\beta}) \subset R^{m+2}$.

The problem to be solved is reduced to the estimation of the vector of structural parameters $\boldsymbol{\beta}$ on the basis of available observations of interpreted variable x.

It is assumed that general population $\mathbf{P}$ is divisible to k subpopulations $\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_k$, which create the complete set. It leads to allocating to particular subpopulations the sizes $n_1, n_2, \ldots, n_k$, in such a way that $n = \sum_{i=1}^{k} n_i$. For subpopulation $\mathbf{P}_i$ we determine the random variable ξ_i of Bernoulli distribution $B(n_i, p_i)$, where p_i is the probability of occurrence of the event (success) A in this subpopulation, and the fraction of successes is expressed by the random variable $\varepsilon_i = \frac{\xi_i}{n_i}$ with moments:

$$E(\varepsilon_i) = p_i \text{ and } D^2(\varepsilon_i) = \frac{p_i q_i}{n_i}, \tag{2}$$

where $q_i = 1 - p_i$.

III. PROBABILISTIC MODEL

We assume that in subpopulation $\mathbf{P}_i$ there was determined the unbiased estimator $\hat{p}_i$ of parameter p_i on the basis of n_i element random sample. For this estimator being a random variable of Bernoulli distribution we assume that its moments of distribution are identical to the random variable ε_i given in chapter 2. Since $\hat{p}_i$ is the observed value p_i, therefore there occurs the dependence

$$\hat{p}_i = p_i + \eta_i \text{ or } \hat{p}_i = f[h(x_i, \boldsymbol{\beta})] + \eta_i, \tag{3}$$

where

$$h(x_i, \boldsymbol{\beta}) = \sum_{j=0}^{k} \beta_j g_j(x_i) = \boldsymbol{\beta}' \mathbf{g}(x_i) \tag{4}$$

and $\mathbf{g}(x_i) = (g_0(x_i), g_1(x_i), \ldots, g_k(x_i))'$. In the model (3) there occurs random variable η_i expressing the measure of error which is made when assuming for the estimation of parameter p_i its value from the sample $\hat{p}_i$. For this variable we have:

(a) expected value

$$E(\hat{p}_i) = E(p_i + \eta_i) = p_i + E(\eta_i), \text{ i.e. } p_i = p_i + E(\eta_i), \text{ and thus } E(\eta_i) = 0, \tag{5}$$

(b) variance

$$D^2(\hat{p}_i) = D^2(p_i + \eta_i) = D^2(\eta_i), \text{ i.e. } D^2(\eta_i) = \frac{p_i q_i}{n_i}, \tag{6}$$

in accordance with the term proposed in (2).

Moreover, for random variables η_i there is assumed non-correlation, i.e. $Cov(\eta_i, \eta_{i'}) = 0,\ i, i' = 1,2,...,k; i \neq i'$.

Determining the estimations of vector $\boldsymbol{\beta}$ of structural parameters from the model (3) is possible at adequately selected functions *f*. In the simplest situation, when this function is linear in the form $f(u) = u$, then the problem of estimation of vector $\boldsymbol{\beta}$ is solved directly from the linear probabilistic model (LPM).

Assuming that in each subpopulation the values x_i of the interpreted variable *x* are known and the form of the function *h* is retained, there is established (k ×(m+1))-dimensional observation matrix

$$\mathbf{X} = \begin{bmatrix} g_0(x_1) & g_1(x_1) & \dots & g_m(x_1) \\ g_0(x_2) & g_1(x_2) & \dots & g_m(x_2) \\ \dots & \dots & \dots & \dots \\ g_0(x_k) & g_1(x_k) & \dots & g_m(x_k) \end{bmatrix} = [\mathbf{g}_0 \quad \mathbf{g}_1 \quad \dots \quad \mathbf{g}_m],$$

where $\mathbf{g}_j = (g_j(x_1), g_j(x_2), ..., g_m(x_k))', j = 1,2, ..., m$, but $g_0(x_i) \equiv 1,\ i = 1,2,...,k$.

Introducing vectorial designations:

$$\hat{\mathbf{p}} = \begin{bmatrix} \hat{p}_1 \\ \hat{p}_2 \\ \dots \\ \hat{p}_k \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_m \end{bmatrix}, \quad \boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \dots \\ \eta_k \end{bmatrix},$$

we get LPM

$$\hat{\mathbf{p}} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\eta}, \tag{7}$$

for which: $E(\boldsymbol{\eta}) = \mathbf{0}$ i $D(\boldsymbol{\eta}) = \boldsymbol{\Omega} = diag(D^2(\eta_1), D^2(\eta_2),\ ...,\ D^2(\eta_m))$, i.e. variance-covariance matrix is a diagonal matrix with elements given in (6). Determining from LPM the vector $\boldsymbol{\beta}$ is done by the generalized least squares method, obtaining $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\hat{\mathbf{p}}$, replacing unknown diagonal elements in matrix $\boldsymbol{\Omega}$ by substitution $p_i = \hat{p}_i$. Finally the estimated unknown probabilities p_i take the form

$$\widetilde{p}_i = \sum_{j=0}^{m} \hat{\beta}_j g_j(x_i). \tag{8}$$

Similar solution $\widetilde{p}_i$ directly depends on set values x_i and assumed linear combination (4).

The considerations which we have presented so far can be referred to the case of the distribution function $F_J(u) = u$ of random variable of uniform distribution at the range $(0, 1)$, i.e. the function f in (3) is replaced by the distribution function F_J.

If the examined uniform distribution is considered at the range (a, b), then we replace the function f in (3) with the distribution function $F_J(u; a, b) = \dfrac{u-a}{b-a}$. We replace the model (3) with its another form $\hat{p}_i = \dfrac{\boldsymbol{\beta}'\mathbf{g}(x_i) - a}{b-a} + \eta_i$, which after transformations turns into

$$\hat{p}_i^* = \boldsymbol{\beta}'\mathbf{g}(x_i) + \eta_i^*,\ i = 1,2, ..., m \tag{9}$$

where $\hat{p}_i^* = (b-a)\hat{p}_i + a$ and $\eta_i^* = (b-a)\eta_i$. For the random variable η_i^* the moments of distribution are: $E(\eta_i^*) = 0$ and $D^2(\eta_i^*) = \dfrac{p_i q_i}{n_i}(b-a)^2$. After application of adequate designations we obtain the linear model of the form analogical to (7), from which we estimate the vector of structural parameters $\boldsymbol{\beta}$ and the forecasted values (8).

IV. PROBIT MODEL

Now we replace the function in (3) with the distribution function F_N of standardized normal distribution (distribution $N(0,1)$), i.e. the probability p_i is equal to the value of the distribution function for the quantile $z_i = \boldsymbol{\beta}'\mathbf{g}(x_i)$. In the considered problem the quantiles z_i at set p_i are called probits, which are expressed by

$$p_i = F_N(z_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_i} e^{-t^2/2} dt . \qquad (10)$$

A function inverse to F_N is F_N^{-1}, i.e. $F_N^{-1}(p_i) = z_i = \boldsymbol{\beta}'\mathbf{g}(x_i)$, and a derivative of the given distribution function is density $f_N(z_i) = \frac{1}{\sqrt{2\pi}} e^{-z_i^2/2}$.

For further considerations we will use the expansion of the function $f(x+a)$ into Taylor's series in the neighbourhood of x (see e.g. Mizerski 1999) $f(x+a) = \sum_{j=0}^{n} \frac{f^{(j)}(x)}{j!} \cdot a^j + r$, where $f^{(j)}$ denotes the derivative of the function f, and r is the expansion of the remainder.

We use (3), substituting it to the inverse function F_N^{-1} and expanding it into Taylor's series to linear term we obtain

$$F_N^{-1}(\hat{p}_i) = F_N^{-1}(p_i + \eta_i) = F_N^{-1}(p_i) + (F_N^{-1}(p_i))' \cdot \eta_i = \boldsymbol{\beta}'\mathbf{g}(x_i) + \eta_i f_N^{-1}(p_i),$$

where f_N^{-1} is the function inverse to density f_N. It takes the form

$$f_N^{-1}(p_i) = \sqrt{-\ln(2\pi \cdot p_i^2)}, \text{ at } p_i \in \left(0, \frac{1}{\sqrt{2\pi}}\right).$$

Figure 1 illustrates the graph of the given function.

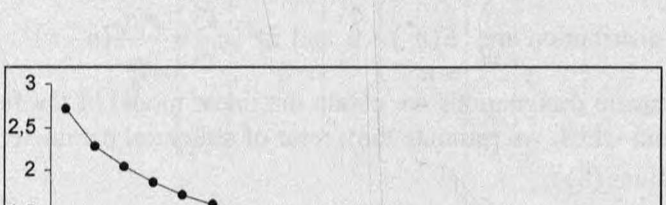

Fig. 1. Graph of the function inverse to density of distribution $N(0, 1)$
Source: Own elaboration

Now the linear model for estimation of probabilities p_i takes the form

$$p_i^* = \boldsymbol{\beta}'\mathbf{g}(x_i) + \eta_i^*, \tag{11}$$

where $p_i^* = F_N^{-1}(\hat{p}_i)$ and $\eta_i^* = \eta_i \cdot \sqrt{-\ln(2\pi \cdot p_i^2)}$. Moments of distribution of the random component in (11) are: $E(\eta_i^*) = 0$ and $D^2(\eta_i^*) = (-\ln(2\pi \cdot p_i^2) \cdot \frac{p_i q_i}{n_i}$ for i = 1,2, ..., k. Further procedure to the model (11) is analogical to models (7) and (9).

Probits in (10) can be considered for more general normal distribution $N(\mu,\sigma)$ at set values of parameters μ,σ. Then in the model (11) we should make a modification into:

$$p_i^* = F^{-1}(\hat{p}_i;\mu,\sigma) = \mu + \sigma \cdot F_N^{-1}(\hat{p}_i) \text{ and } \eta_i^* = \eta_i \cdot \{\mu + \sigma\sqrt{-\ln(2\pi\sigma^2 \cdot p_i^2)},$$

at $p_i \in \left(0, \frac{1}{\sigma\sqrt{2\pi}}\right)$.

V. LOGIT MODEL

At present as the function f in (3) we take the distribution function of standardized logistic distribution $p_i = F_L(z_i) = \frac{1}{1+e^{-z_i}}$, where $z_i = \boldsymbol{\beta}'\mathbf{g}(x_i)$. We obtain the inverse function F_L^{-1} from the series of transformations: $\frac{1}{p_i} = 1 + e^{-z_i}$; $\frac{1-p_i}{p_i} = e^{-z_i}$; $\ln\left(\frac{p_i}{1-p_i}\right) = z_i = F_L^{-1}(p_i)$. Figure 2 illustrates the graph of the function.

We expand the function $F_L^{-1}(p_i)$ in point $\hat{p}_i = p_i + \eta_i$ in neighbourhood p_i using the expansion of Taylor's series, which leads to

$$F_L^{-1}(\hat{p}_i) = F_L^{-1}(p_i + \eta_i) = F_L^{-1}(p_i) + (F_L^{-1}(p_i))' \cdot \eta_i =$$

$$= z_i + \eta_i \cdot \left[\ln\left(\frac{p_i}{1-p_i}\right)\right]' = z_i + \frac{\eta_i}{p_i(1-p_i)}.$$

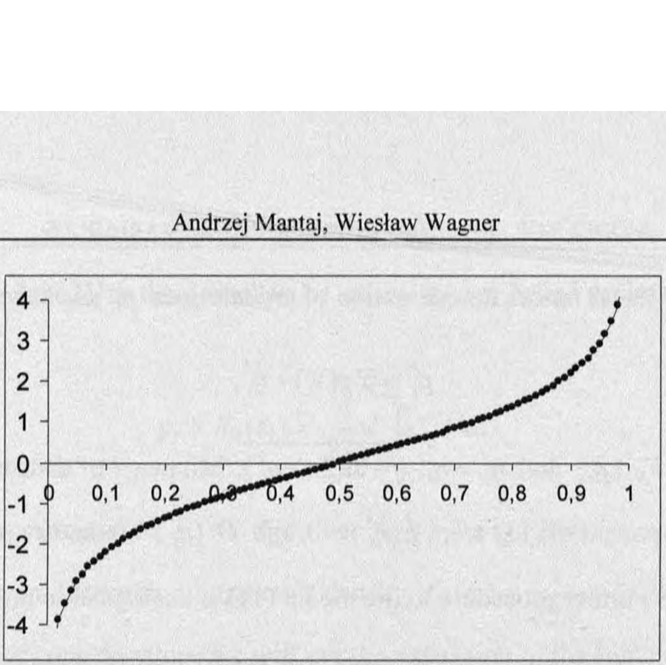

Fig. 2. Graph of the function inverse to the distribution function of the logistic distribution

Source: Own elaboration

Finally, for estimation of probabilities p_i we apply the model (11), where $p_i^* = \ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right)$ and $\eta_i^* = \frac{\eta_i}{p_i(1-p_i)}$, whereas $E(\eta_i^*) = 0$ and $D^2(\eta_i^*) = \frac{1}{n_i \cdot p_i q_i}$.

VI. GENERALIZED APPROACH

The cases of determining the estimations for probabilities p_i, given in chapters 3 and 4, can be generalized to the case of any continuous distribution function, at assumptions on random variables η_i as in chapter 3, in the following way:

(a) Z – continuous random variable,

(b) $p = F_Z(z;\boldsymbol{\theta})$ – the distribution function of random variable Z, at set vector of constants $\boldsymbol{\theta}$, where $z = \boldsymbol{\beta}'\mathbf{g}(x)$,

(c) $F_Z^{-1}(z;\boldsymbol{\theta})$ the function inverse to the distribution function F_Z,

(d) $F_Z^{-1}(\hat{p}_i;\boldsymbol{\theta}) = \boldsymbol{\beta}'\mathbf{g}(x_i) + \eta_i \cdot f_Z^{-1}(p_i;\boldsymbol{\theta})$ – the linear model of estimation of vector $\boldsymbol{\beta}$.

The quantities occurring on the left side of the equation in (d) are called *Z*-its.

Example. Exponential distribution

(a) W – random variable of exponential distribution,

(b) $p_i = F_W(z_i;\lambda) = 1 - e^{-z_i/\lambda}$, where $z_i = \boldsymbol{\beta}'\mathbf{g}(x_i)$ i $\boldsymbol{\theta} = (\lambda)$,

(c) $F_W^{-1}(p_i;\lambda) = -\lambda \ln(1-p_i)$,

(d) $F_W^{-1}(\hat{p}_i;\lambda) = \boldsymbol{\beta}'\mathbf{g}(x_i) + \eta_i^*$, where $\eta_i^* = \eta_i[-\lambda \cdot \ln(\lambda \cdot p_i)]$,

(e) $E(\eta_i^*) = 0,\ D^2(\eta_i^*) = \lambda^2[\ln(\lambda \cdot p_i)]^2 \cdot \dfrac{p_i q_i}{n_i}$.

VII. SUMMARY

In the paper there has been presented methods of estimation of structural parameters in models of probability of selected random variables. Within these the inverse distribution functions are used, which at set empirical values of parameter p of random variable of Bernoulli distribution allow determining the values of dependent variable. In consequence it leads to adequate linear models which enable determining the searched structural parameters by the generalized least squares method. Such models were built for the following distributions: uniform, normal, logistic and exponential, but in the case of uniform and exponential distributions there were also considered their parameters. This procedure allowed formulating the general procedure of estimation of parameters of considered models.

REFERENCES

Guzik B., Jurek W., (20003), *Podstawowe metody ekonometrii* (Basic Methods of Econometrics), Akademia Ekonomiczna w Poznaniu.

Lipiec-Zajchowska M. (red.), (2003), *Wspomaganie procesów decyzyjnych* (Enhancement of the Decision-Making Processes), Tom II. Ekonometria, C.H. Beck, Warszawa.

Mizerski W. (red.), (1999), *Tablice matematyczne (Mathematical Tables)*, Adamantan, Warszawa.

Maddala G.S., (2006), *Ekonometria* (Econometrics), PWN, Warszawa.

Ostasiewicz W. (red.), (1999), *Statystyczne metody analizy danych* (Statistical methods in data analysis), Akademia Ekonomiczna we Wrocławiu.

Andrzej Mantaj, Wiesław Wagner

MODELE PRAWDOPODOBIEŃSTWA DLA ZMIENNYCH LOSOWYCH O ROZKŁADZIE BERNOULLI'EGO

W pracy przedstawiono metodę szacowania parametrów strukturalnych modelach prawdopodobieństwa wybranych zmiennych losowych. W metodzie wykorzystuje się ich dystrybuanty odwrotne, które przy zadanych wartościach empirycznych parametru p zmiennej losowej o rozkładzie Bernoulli'ego pozwalają wyznaczyć wartości zmiennej

zależnej. Prowadzi to w konsekwencji do odpowiednich modeli liniowych, które umożliwiają uogólnioną metodą najmniejszych kwadratów wyznaczyć poszukiwane parametry strukturalne. Modele takie zostały zbudowane dla rozkładów: jednostajnego, normalnego, logistycznego i wykładniczego, przy czym w przypadku rozkładów jednostajnego i wykładniczego uwzględniono także ich parametry. Postępowanie to pozwoliło na sformułowanie ogólnej procedury estymacji parametrów rozważnych modeli.