PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1140 — 2006
Praktyka statystyki

Dariusz Biskup

Akademia Ekonomiczna we Wrocławiu

WYBÓR MODELU ORAZ ZMIENNYCH DO MODELU W UJĘCIU BAYESOWSKIM

1. Wstęp

Wybór, a więc identyfikacja postaci analitycznej modelu regresji, jest jednym z istotniejszych problemów analizy danych eksperymentalnych. W ujęciu bayesowskim odbywa się on przy użyciu kryterium prawdopodobieństwa *a posteriori* prawdziwości modelu. Prawdziwość modelu rozumiana jest jako poprawność identyfikacji mechanizmu generującego dane będące przedmiotem analizy. Mimo że model, według którego generowane są dane, jest zdeterminowany (nie jest losowy), w ujęciu bayesowskim określa się rozkład prawdopodobieństwa na zbiorze rozpatrywanych modeli, wyrażając w ten sposób niepewność badacza względem postaci modelu. Takie podejście pozwala również na uwzględnienie w analizie subiektywnej wiedzy badacza poprzez określenie rozkładu *a priori* na przestrzeni modeli. Oznacza to, że w ujęciu bayesowskim jako zmienne losowe mogą być traktowane nie tylko parametry modelu regresji, ale również jego postać analityczna.

Praktyczne obliczanie prawdopodobieństw modeli w ujęciu bayesowskim wymusza zwykle zastosowanie metod numerycznych – głównie algorytmów Monte Carlo (np. algorytm Gibbsa).

Obszerne omówienie zagadnień związanych z wyborem modelu regresji oraz zmiennych do modelu można znaleźć na przykład w [Bernardo 1995; Congdon 2001; Koop 2003].

2. Wybór modelu regresji

Rozważmy r konkurujących modeli regresji i niech model j $(j = 1,..,r)$ będzie zdefiniowany równaniem

$$Y(\mathbf{x}) = g_j\left(\mathbf{x}, \theta^{(j)}\right) + \varepsilon_j(\mathbf{x}), \tag{1}$$

gdzie: $Y(\mathbf{x})$ – zmienna losowa objaśniana,

$g_j\left(\mathbf{x}, \theta^{(j)}\right)$ – funkcja regresji dla modelu j $(j = 1,..,r)$,

$\mathbf{x}=[x_1 \ x_2 \ \cdots \ x_k]^T$ – wektor wartości zmiennych objaśniających $X_1, X_2, ..., X_k$,

$\theta^{(j)} = \left[\theta_1^{(j)} \ \theta_2^{(j)} \ \cdots \ \theta_{k_j}^{(j)}\right]^T$ – wektor k_j parametrów modelu j $(j = 1,..,r)$,

$\varepsilon_j(\mathbf{x})$ – błąd losowy dla modelu j $(j = 1,..,r)$.

Niech M oznacza dyskretną zmienną losową określającą, który spośród rozpatrywanych modeli regresji wyraża związek pomiędzy zmienną objaśnianą Y a zmiennymi objaśniającymi $X_1, X_2, ..., X_k$. Zmienna losowa M może przyjmować wartości od 1 do r, przy czym jeśli $M = j$, oznacza to prawdziwość modelu j $(j = 1,..,r)$.

Zmienna losowa M może zostać łatwo ujęta w równaniu regresji. Zamiast posługiwać się równaniem (1) właściwym dla pojedynczego modelu, można zdefiniować jedno rozszerzone równanie regresji, które będzie zawierało wszystkie modele szczegółowe, a wśród nich jeden model prawdziwy. Taki rozszerzony model regresji można wyrazić wzorem

$$Y(\mathbf{x}) = \sum_{j=1}^{r} \lambda_j \left(g_j\left(\mathbf{x}, \theta^{(j)}\right) + \varepsilon_j(\mathbf{x})\right), \quad \lambda_j \in \{0,1\}, \quad \sum_{j=1}^{r} \lambda_j = 1, \tag{2}$$

przy czym jedyny spośród r współczynników λ_j $(j = 1,..., r)$, który jest równy 1, identyfikuje numer prawdziwego modelu, a pozostałe zerowe współczynniki λ_j identyfikują $r-1$ modeli nieprawdziwych. Z podanej interpretacji wynika, że $p(M = j) = p\left(\lambda_j = 1, \lambda_{\neq j} = 0\right)$ dla $j = 1,..., r$.

Parametry λ_j, podobnie jak parametry $\theta^{(j)} = \left[\theta_1^{(j)} \ \theta_2^{(j)} \ \cdots \ \theta_{k_j}^{(j)}\right]^T$, traktowane są zatem jako zmienne losowe, co oznacza, że możemy za pomocą wzoru Bayesa wyznaczać prawdopodobieństwo prawdziwości modelu (wymaga to uprzedniego określenia rozkładu *a priori* $p(\lambda_1,..., \lambda_r)$). Wnioskowanie o prawdopodobieństwie prawdziwości modelu regresji można więc zastąpić wnioskowaniem o rozkładzie parametrów modelu regresji.

Prawdopodobieństwo prawdziwości modelu j można zgodnie z wzorem Bayesa wyznaczyć następująco (zakłada się, że zmienne objaśniające są nielosowe, a zatem $p(\mathbf{y}, \mathbf{X}) = p(\mathbf{y}|\mathbf{X})$):

$$p(M = j|\mathbf{X},\mathbf{y}) = \frac{p(M = j)\cdot p(\mathbf{y}|\mathbf{X}, M = j)}{p(\mathbf{y}|\mathbf{X})}, \tag{3}$$

gdzie: $\mathbf{y} = [y_1 \quad y_n]^T$ – wektor n realizacji zmiennej objaśnianej Y,

$\mathbf{X} = \begin{bmatrix} x_{11} & x_{1k} \\ \vdots & \vdots \\ x_{n1} & \cdots \quad x_{nk} \end{bmatrix}$ – macierz n wartości k zmiennych objaśniających $X_1, X_2, ..., X_k$,

$p(M = j|\mathbf{X},\mathbf{y})$ – prawdopodobieństwo, że prawdziwy jest model j, jeśli zaobserwowano dane $\mathbf{X}$, $\mathbf{y}$,

$p(M = j)$ – prawdopodobieństwo *a priori* prawdziwości modelu j,

$p(\mathbf{y}|\mathbf{X}, M = j)$ – prawdopodobieństwo zaobserwowania danych $\mathbf{X}$, $\mathbf{y}$, jeśli prawdziwy jest model j.

Wzór (3) można rozwinąć następująco:

$$p(M = j|\mathbf{X},\mathbf{y}) = \frac{p(M = j)\,p(\mathbf{y}|\mathbf{X}, M = j)}{p(\mathbf{y}|\mathbf{X})} =$$

$$\frac{p(M = j)\int\limits_{\Theta^{(j)}} p(\mathbf{y},\theta^{(j)}|\mathbf{X}, M = j)d\theta^{(j)}}{p(\mathbf{y}|\mathbf{X})} =$$

$$\frac{p(M = j)}{p(\mathbf{y}|\mathbf{X})}\int\limits_{\Theta^{(j)}} p(\mathbf{y}|\mathbf{X}, M = j,\theta^{(j)})p(\theta^{(j)}|M = j)d\theta^{(j)}.$$

Zauważmy, że jeżeli prawdopodobieństwa *a priori* $p(M = j)$ są identyczne dla każdego modelu, to wpływ na prawdopodobieństwo modelu ma wyłącznie składnik:

$$p(\mathbf{y}|\mathbf{X}, M = j) = \int\limits_{\Theta^{(j)}} p(\mathbf{y}|\mathbf{X}, M = j,\theta^{(j)})p(\theta^{(j)}|M = j)d\theta^{(j)}. \tag{4}$$

Obliczanie prawdopodobieństw modeli wymaga w praktyce zastosowania metod numerycznych (algorytmów Monte Carlo).

3. Wybór zmiennych do modelu

Szczególnym przypadkiem problemu wyboru modelu jest zagadnienie wyboru zmiennych do modelu regresji liniowej. Załóżmy, że zmienna zależna Y może być determinowana przez pewien podzbiór zbioru zmiennych niezależnych $X_1, X_2, ..., X_k$.

Równanie regresji związane z jednym z modeli będzie miało postać

$$Y(\mathbf{x}) = \sum_{j=1}^{r} \lambda_j \left(\left(\mathbf{x}^{(j)} \right)^T \theta^{(j)} + \varepsilon_j \left(\mathbf{x}_j \right) \right), \tag{5}$$

gdzie: $\theta^{(j)} = \left[\theta_1^{(j)} \quad \theta_2^{(j)} \quad \cdots \quad \theta_{k_j}^{(j)} \right]^T$,

k_j – liczba parametrów modelu j,

R – liczba modeli,

$\mathbf{x} = \left[x_1 \quad x_2 \quad \cdots \quad x_k \right]^T$ – wektor wartości zmiennych objaśniających,

$\mathbf{x}^{(j)}$ – podzbiór wartości wektora zmiennych objaśniających o k_j składowych,

$\varepsilon_j \left(\mathbf{x}_j \right)$ – błąd modelu j.

Poza parametrami modelu regresji liniowej $\boldsymbol{\theta}^{(j)}$ w modelu występują dodatkowo parametry $\lambda_1, \ldots, \lambda_r$. Prawdopodobieństwo modelu może zostać obliczone za pomocą wzoru (3).

4. Praktyczne sposoby obliczania prawdopodobieństw modeli

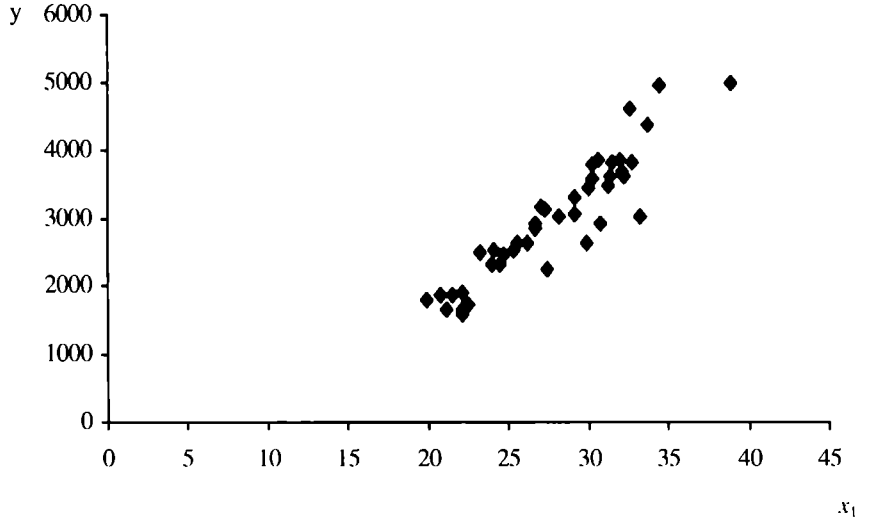
Wybór zmiennych do modelu. Załóżmy, że zmienna Y może być opisana przez dwie potencjalne zmienne objaśniające X_1 i X_2 oraz że wystarczające jest rozpatrzenie dwóch modeli: pierwszy z nich będzie uwzględniać zmienną X_1 (model 1), drugi zmienną X_2 (model 2). Wówczas równanie (5) można rozpisać następująco:

$$Y(\mathbf{x}) = \lambda_1 \left(\theta_1^{(1)} + \theta_2^{(1)} x_1 + \varepsilon(x_1) \right) + \lambda_2 \left(\theta_1^{(2)} + \theta_2^{(2)} x_2 + \varepsilon(x_2) \right).$$

Przyjmijmy, że zmienna losowa M będzie równa 1, jeśli prawdziwy jest model 1, lub 2, jeśli prawdziwy jest model 2. Założony zostanie normalny rozkład błędów: $\varepsilon(x_1) \sim N(0, \sigma)$ i $\varepsilon(x_2) \sim N(0, \tau)$.

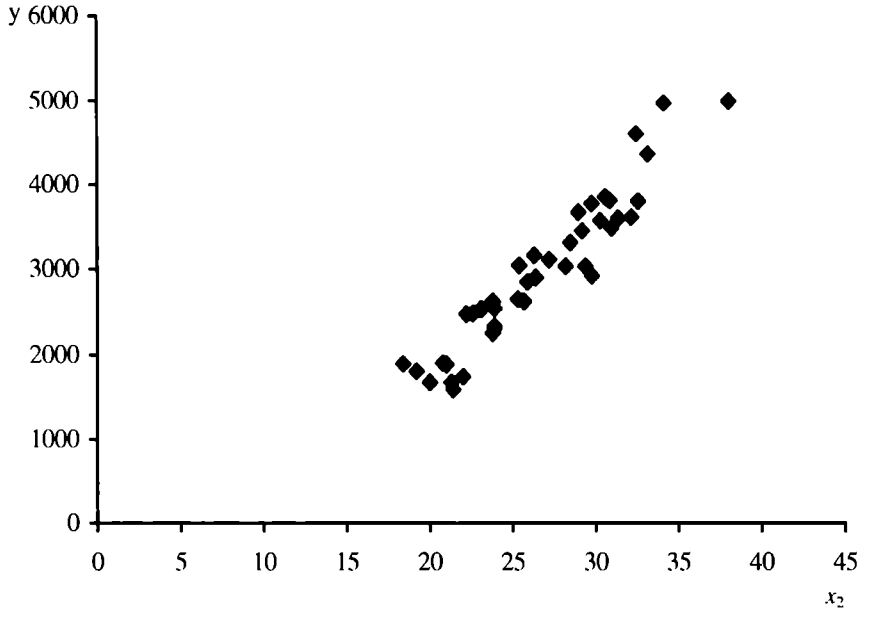
Na rysunkach 1 i 2 przedstawione są wykresy korelacyjne zmiennych x_1 i x_2 względem zmiennej y.

Aby obliczyć prawdopodobieństwa prawdziwości poszczególnych modeli, należy określić prawdopodobieństwa *a priori* modeli oraz ich parametrów (dane do przykładu oraz zastosowane rozkłady pochodzą z pracy [Carlin, Chib 2001]):



Rys. 1. Wykres korelacyjny zmiennej y i x_1

Źródło: opracowanie własne.



Rys. 2. Wykres korelacyjny zmiennej y i x_2

Źródlo: opracowanie własne.

$$p(M=1)=\frac{1}{2}, \qquad p(M=2)=\frac{1}{2},$$

$$\left(\theta_1^{(1)}\middle|M=1\right)\sim N(3000,1000), \qquad \left(\theta_1^{(2)}\middle|M=2\right)\sim N(3000,1000),$$

$$\left(\theta_2^{(1)}\middle|M=1\right)\sim N(185,100), \qquad \left(\theta_2^{(2)}\middle|M=2\right)\sim N(185,100),$$

$$\left(\sigma^2\middle|M=1\right)\sim IG(3,180000), \qquad \left(\tau^2\middle|M=2\right)\sim IG(3,180000).$$

Symbolem $IG(\cdot)$ oznaczono odwrócony rozkład gamma. Przyjęto założenie, że rozkłady parametrów są *a priori* niezależne.

Rozkłady *a priori* parametrów mają zatem postać:

$$p\left(\theta_1^{(1)}\middle|M=1\right)=\frac{1}{1000\sqrt{2\pi}}\exp\left(-\frac{\left(\theta_1^{(1)}-3000\right)^2}{2\cdot 1000^2}\right),$$

$$p\left(\theta_2^{(1)}\middle|M=1\right)=\frac{1}{100\sqrt{2\pi}}\exp\left(-\frac{\left(\theta_2^{(1)}-185\right)^2}{2\cdot 100^2}\right),$$

$$p\left(\theta_1^{(2)}\middle|M=1\right)=\frac{1}{1000\sqrt{2\pi}}\exp\left(-\frac{\left(\theta_1^{(2)}-3000\right)^2}{2\cdot 1000^2}\right),$$

$$p\left(\theta_2^{(2)}\middle|M=1\right)=\frac{1}{100\sqrt{2\pi}}\exp\left(-\frac{\left(\theta_2^{(2)}-185\right)^2}{2\cdot 100^2}\right),$$

$$p\left(\sigma^2\middle|M=1\right)=\frac{180\,000^3}{\Gamma(3)}\left(\sigma^2\right)^{-2}e^{-\frac{180\,000}{\sigma^2}},$$

$$p\left(\tau^2\middle|M=2\right)=\frac{180\,000^3}{\Gamma(3)}\left(\tau^2\right)^{-2}e^{-\frac{180\,000}{\tau^2}}.$$

Wyznaczmy następnie funkcje wiarygodności dla poszczególnych modeli:

$$p\left(\mathbf{y}\middle|M=1,\theta_1^{(1)},\theta_2^{(1)},\sigma^2\right)=\prod_{i=1}^{n}p\left(y_i\middle|M=1,\theta_1^{(1)},\theta_2^{(1)},\sigma^2\right)=$$

$$\prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{\left(\theta_1^{(1)}x_{i2}+\theta_2^{(1)}-y_i\right)^2}{2\sigma^2}\right)=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n\exp\left(-\frac{\sum_{i=1}^{n}\left(\theta_1^{(1)}x_{i2}+\theta_2^{(1)}-y_i\right)^2}{2\sigma^2}\right),$$

$$P\left(\mathbf{y}\middle|M=2,\theta_1^{(2)},\theta_2^{(2)},\tau^2\right)=\prod_{i=1}^{n}p\left(y_i\middle|M=2,\theta_1^{(2)},\theta_2^{(2)},\tau^2\right)=$$

$$\prod_{i=1}^{n}\frac{1}{\tau\sqrt{2\pi}}\exp\left(-\frac{\left(\theta_1^{(2)}x_{i2}+\theta_2^{(2)}-y_i\right)^2}{2\tau^2}\right)=\left(\frac{1}{\tau\sqrt{2\pi}}\right)^n\exp\left(-\frac{\sum_{i=1}^{n}\left(\theta_1^{(2)}x_{i2}+\theta_2^{(2)}-y_i\right)^2}{2\tau^2}\right).$$

Ostatecznie prawdopodobieństwa modeli można obliczyć następująco:

$$p\left(M=1\middle|\mathbf{y}\right)\propto$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{0}^{\infty}p\left(\mathbf{y}\middle|\theta_1^{(1)},\theta_2^{(1)},\sigma^2,M=1\right)p\left(\theta_1^{(1)}\middle|M=1\right)\cdot$$

$$P\left(\theta_2^{(1)}\middle|M=1\right)P\left(\sigma^2\middle|M=1\right)d\theta_1^{(1)}d\theta_2^{(1)}d\sigma^2=$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{0}^{\infty}\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n\exp\left(-\frac{\sum_{i=1}^{n}\left(\theta_1^{(1)}x_{i1}+\theta_2^{(1)}-y_i\right)^2}{2\sigma^2}\right)\cdot\frac{1}{1000\sqrt{2\pi}}\exp\left(-\frac{\left(\theta_1^{(1)}-3000\right)^2}{2\cdot1000^2}\right)\cdot$$

$$\frac{1}{100\sqrt{2\pi}}\exp\left(-\frac{\left(\theta_2^{(1)}-185\right)^2}{2\cdot100^2}\right)\cdot\frac{180.000^3}{\Gamma(3)}\left(\sigma^2\right)^{-2}e^{-\frac{180.000}{\sigma^2}}d\theta_1^{(1)}d\theta_2^{(1)}d\sigma^2,$$

$$p\left(M=2\middle|\mathbf{y}\right)\propto$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{0}^{\infty}p\left(\mathbf{y}\middle|\theta_1^{(2)},\theta_2^{(2)},\tau^2,M=2\right)p\left(\theta_1^{(2)}\middle|M=2\right)\cdot$$

$$P\left(\theta_2^{(2)}\middle|M=2\right)P\left(\tau^2\middle|M=2\right)d\theta_1^{(2)}d\theta_2^{(2)}d\tau^2=$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{0}^{\infty}\left(\frac{1}{\tau\sqrt{2\pi}}\right)^n\exp\left(-\frac{\sum_{i=1}^{n}\left(\theta_1^{(2)}x_{i2}+\theta_2^{(2)}-y_i\right)^2}{2\tau^2}\right)\cdot\frac{1}{1000\sqrt{2\pi}}\exp\left(-\frac{\left(\theta_1^{(2)}-3000\right)^2}{2\cdot1000^2}\right)\cdot$$

$$\frac{1}{100\sqrt{2\pi}}\exp\left(-\frac{\left(\theta_2^{(2)}-185\right)^2}{2\cdot100^2}\right)\cdot\frac{180\,000^3}{\Gamma(3)}\left(\tau^2\right)^{-2}e^{-\frac{180\,000}{\tau^2}}d\theta_1^{(2)}d\theta_2^{(2)}d\tau^2.$$

Obliczenia powyższych całek wykonano numerycznie przy użyciu programu WINBUGS. Uzyskane wyniki są następujące:

$p(M = 1|\mathbf{X}, \mathbf{y}) = 0,2688,$

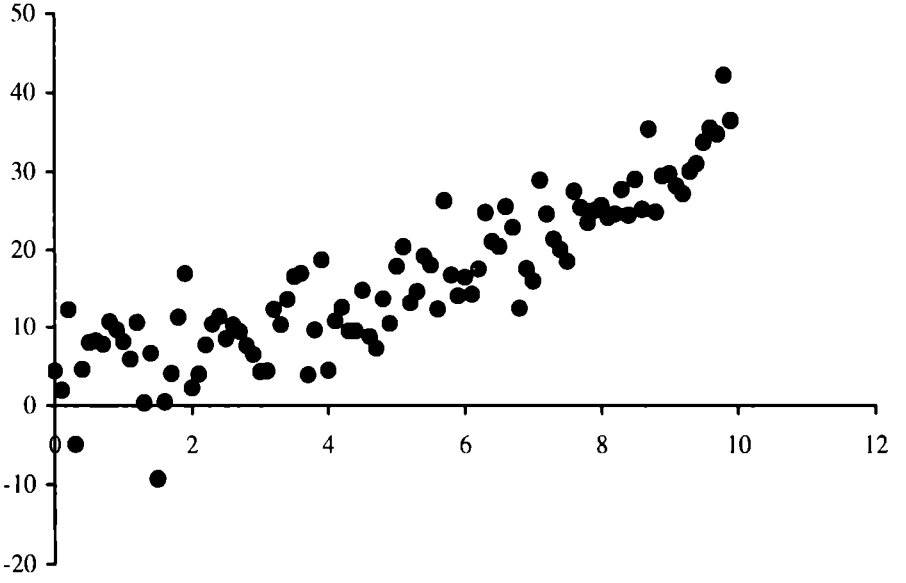
$p(M = 2|\mathbf{X}, \mathbf{y}) = 0,6312.$

Oznacza to, że za bardziej prawdopodobny został uznany model 2, który uwzględnia zmienną x_2 jako zmienną objaśniającą.

Wybór postaci analitycznej modelu. Załóżmy, że cecha Y może być opisana przez pojedynczą zmienną objaśniającą X. Zależność pomiędzy tymi zmiennymi może być liniowa lub kwadratowa. Zgodnie z wzorem (2) model przyjmuje postać

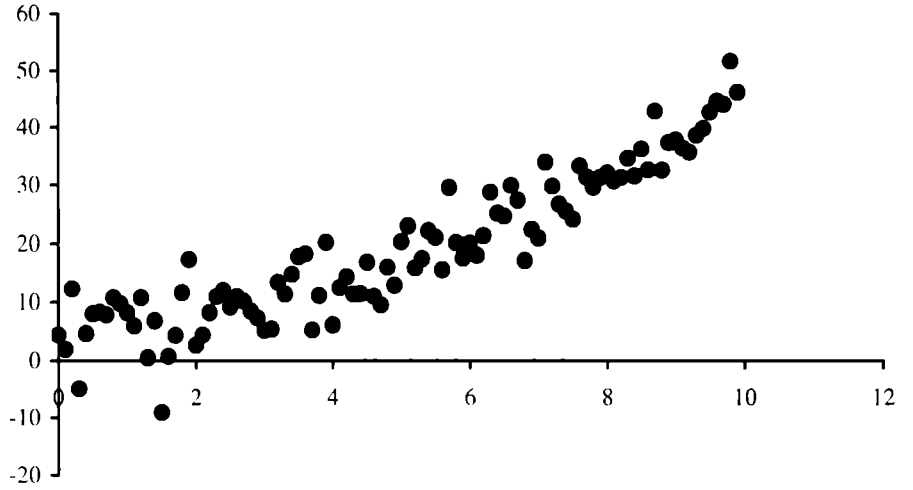
$$Y(x) = \lambda_1\left(\theta_1^{(1)} + \theta_2^{(1)}x + \theta_3^{(1)}x^2 + \varepsilon_1(x)\right) + \lambda_2\left(\theta_1^{(2)} + \theta_2^{(2)}x + \varepsilon_2(x)\right). \qquad (6)$$

Analizie zostały poddane cztery zestawy danych, wygenerowane zgodnie z zależnością $Y(x) = ax^2 + x + 4 + \varepsilon(x)$, dla zmiennej losowej $\varepsilon(x) \sim N(0,5)$ oraz dla parametru a równego 0,2; 0,3; 0,4 i 0,5. Wykresy korelacyjne otrzymanych danych przedstawiają rys. 3 do 6.



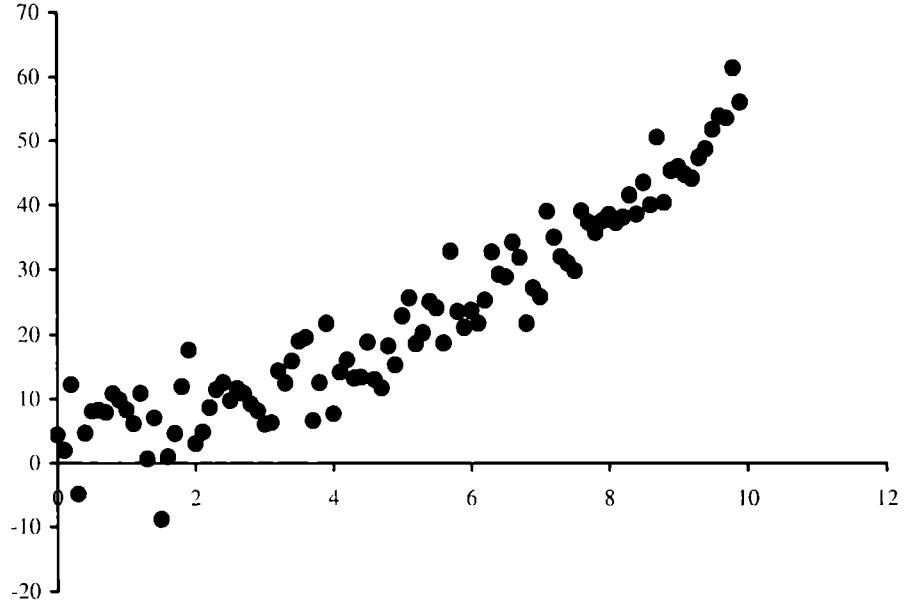
Rys. 3. Wykres korelacyjny dla $a = 0,2$

Źródło: opracowanie własne.



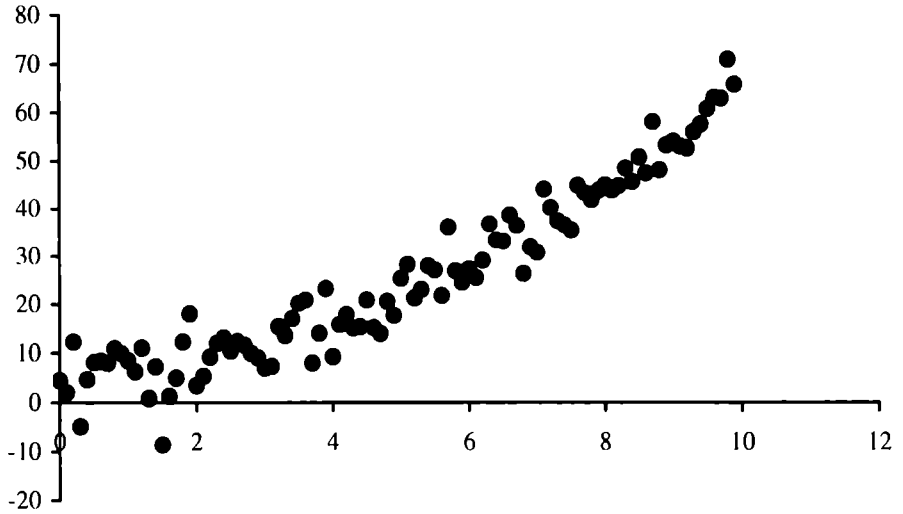
Rys. 4. Wykres korelacyjny dla $a = 0,3$

Źródło: opracowanie własne.



Rys. 5. Wykres korelacyjny dla $a = 0,4$

Źródło: opracowanie własne.



Rys. 6. Wykres korelacyjny dla $a = 0,5$

Źródło: opracowanie własne.

Do obliczenia prawdopodobieństw zastosowana została metoda zaproponowana w [George, McCulloch 1993]. Obliczenia wykonane zostały za pomocą programu WINBUGS.

Tabela 1. Prawdopodobieństwa modelu kwadratowego dla różnych wartości parametru a

Wartość parametru	Prawdopodobieństwo modelu kwadratowego
$a = 0,2$	0,076
$a = 0,3$	0,697
$a = 0,4$	0,985
$a = 0,5$	1

Źródło: obliczenia własne.

Literatura

Bernardo J. M., *Bayesian Theory*, John Wiley & Sons, New York 1995.

Carlin B., Chib S., *Bayesian Model Choice via Markov Chain Monte Carlo Methods*, „Journal Royal Statistical Society" Series B, 2001 vol. 75 nr 3, s. 473-484.

Congdon P., *Bayesian Statistical Modelling*, John Wiley & Sons, New York 2001.
George E., McCulloch R., *Variable Selection* via *Gibbs Sampling*, „Journal American Statistical Association" 1993 vol. 88 nr 423, s. 881-889.
Koop G., *Bayesian Econometrics*, John Wiley & Sons, New York 2003.

BAYESIAN APPROACH TO MODEL AND VARRIABLE SELECTION

Summary

The paper describes Bayesian approach to statistical model identification and to the problem of variable choice in linear regression model. The model choice bases on the criterion of posterior probability of model reliability which is understood as the accurateness of identification of data generation mechanism. Theoretical basis for this approach have been presented as well as several examples of its application.