

Łukasz Kuźmiński

WYBRANE ZAGADNIENIA KOSZTOWO-OPTYMALNEGO PLANOWANIA EKSPERYMENTÓW W ZAGADNIENIU REGRESJI

1. Wstęp

Planowanie eksperymentów jest działem statystyki rozwijanym od lat 30. wieku XX. Początkowo eksperymenty konstruowano z myślą o opracowaniu ich wyników z użyciem analizy wariancji, jednakże w latach 50. ubiegłego wieku rozwinęto planowanie eksperymentów w odniesieniu do zagadnień regresji w celu poprawienia własności estymatorów modelu regresji.

Szybki rozwój teorii planowania eksperymentów w wieku XX związany jest z dostrzeganiem coraz to szerszych możliwości zastosowania tej teorii do różnych zagadnień praktycznych, a nie tylko, jak to miało miejsce na początku, w rolnictwie. Zastosowania metod planowania eksperymentów to np.: ekonomika produkcji (maksymalizacja wydajności procesu), zarządzanie jakością (osiągnięcie założonego poziomu jakości produktu), medycyna (zmniejszenie kosztów badania leków), geografia (np. wybór położenia obserwatoriów meteorologicznych), badania marketingowe i wiele innych.

Równolegle do wzrostu liczby obszarów zastosowania metod planowania eksperymentów rozwijała się teoria planów optymalnych w odniesieniu do zagadnień regresji i planów kosztowo-optymalnych. Istotą tych teorii jest uzyskiwanie planów o możliwie najlepszych własnościach statystycznych przy limitowanej liczbie doświadczeń wyznaczonej przez ograniczone nakłady kosztowe.

W zakresie metod optymalizacji statystycznej kosztowo-optymalne planowanie eksperymentów pozwala:

- uzyskać optymalny lub quasi-optymalny plan eksperymentu przy zadanych nakładach kosztowych,
- uzyskać plan eksperymentu zapewniający żądane własności statystyczne, który minimalizuje koszty eksperymentu.

Teorie te zostaną zarysowane szerzej i będą zilustrowane przykładami empirycznymi w dalszej części pracy.

2. Eksperyment w zagadnieniu regresji

Eksperyment w zagadnieniu regresji polega na zbadaniu określonej liczby cech, które mogą być również określone mianem cech (zmiennych) objaśnianych w stosunku do innego zbioru cech (zmiennych), zwanych objaśniającymi. W praktyce często realizuje się eksperymenty z jedną zmienną objaśnianą w stosunku do zbioru wielu zmiennych objaśniających.

Tego rodzaju eksperyment może być przedstawiony jako ciąg doświadczeń, przy czym doświadczenie polega na uzyskaniu jednej obserwacji zmiennej objaśnianej dla jednocześnie zaobserwowanych (planowanie pasywne) albo ustalonych przez warunki tego doświadczenia wartości (lub stanów) poszczególnych cech objaśniających. W tej pracy rozpatrywane będzie jedynie planowanie aktywne, co oznacza, że wszystkie zmienne objaśniające są zmiennymi kontrolowanymi.

Istotną kwestią dotyczącą planowania aktywnego w zagadnieniu regresji jest ustalenie wartości wszystkich zmiennych objaśniających w poszczególnych doświadczeniach wchodzących w skład eksperymentu. Jeśli eksperyment składa się z n doświadczeń, a liczba zmiennych objaśniających wynosi m , to zbiór danych do eksperymentu można przedstawić w formie macierzy

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{21} & \dots & x_{m1} \\ x_{12} & x_{22} & \dots & x_{m2} \\ \vdots & \vdots & & \vdots \\ x_{1n} & x_{2n} & \dots & x_{mn} \end{bmatrix}. \quad (1)$$

Elementy ciągu $\mathbf{x}^{(i)} = [x_{1i} \ x_{2i} \ \dots \ x_{mi}]$ są odpowiednio równe kolejnym elementom i -tego ($i = 1, 2, \dots, n$) wiersza macierzy $\mathbf{X}$, zwanej macierzą planu eksperymentu lub macierzą planowania. Macierz (1) wyznacza jednocześnie plan eksperymentu P_n obejmujący n doświadczeń zdefiniowany jako zbiór

$$P_n = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(n)}\}, \quad (2)$$

gdzie: $\mathbf{x}^{(k)} = [x_{1k} \ x_{2k} \ \dots \ x_{mk}] \in D_1$,
 $k = 1, 2, \dots, n$.

Elementami zbioru P_n są te punkty obszaru eksperymentowania D_1 , w których należy wykonać poszczególne doświadczenia eksperymentu obejmującego n doświadczeń. Każdy z elementów planu P_n określa jednoznacznie wartości kontrolo-

wanych zmiennych objaśniających $X_1, X_2, \dots, X_m$ w jednym z n doświadczeń eksperymentu.

W przypadku zagadnienia regresji z jedną zmienną objaśnianą $Y(\mathbf{x})$ jej obserwacje uzyskane w wyniku realizacji planu P_n przedstawia się w postaci wektora

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} Y(\mathbf{x}^{(1)}) \\ Y(\mathbf{x}^{(2)}) \\ \vdots \\ Y(\mathbf{x}^{(n)}) \end{bmatrix}. \quad (3)$$

Omawiane związki pomiędzy zmienną objaśnianą a zbiorem zmiennych objaśniających można przedstawić w postaci równania regresji

$$Y(\mathbf{x}) = h(\mathbf{x}; \theta) + \varepsilon(\mathbf{x}). \quad (4)$$

Funkcja $h(\mathbf{x}; \theta)$ zwana jest powierzchnią odpowiedzi i wyraża się wzorem

$$h(\mathbf{x}, \theta) = \theta_1 f_1(\mathbf{x}) + \dots + \theta_r f_r(\mathbf{x}), \quad (5)$$

natomiast

$$\theta = [\theta_1 \ \theta_2 \ \dots \ \theta_r]^T \quad (6)$$

jest wektorem nieznanych parametrów rzeczywistych przedstawionego równania. Z kolei wektor

$$\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}) \ f_2(\mathbf{x}) \ \dots \ f_r(\mathbf{x})]^T \quad (7)$$

składa się z r znanych funkcji rzeczywistych $f_1(\mathbf{x}), f_2(\mathbf{x}), \dots, f_r(\mathbf{x})$ zmiennych $x_1, x_2, \dots, x_m$, które są ciągłe i liniowo niezależne w dziedzinie $D \subset R^m$ równania (4).

Przez $\varepsilon(\mathbf{X})$ oznaczony został wektor błędów losowych dla każdego z n doświadczeń

$$\varepsilon(\mathbf{X}) = \begin{bmatrix} \varepsilon(\mathbf{x}^{(1)}) & \varepsilon(\mathbf{x}^{(2)}) & \dots & \varepsilon(\mathbf{x}^{(n)}) \end{bmatrix}. \quad (8)$$

Niech

$$\mathbf{F}_\mathbf{x} = \begin{bmatrix} \mathbf{f}(\mathbf{x}^{(1)}) & \mathbf{f}(\mathbf{x}^{(2)}) & \dots & \mathbf{f}(\mathbf{x}^{(n)}) \end{bmatrix}^T. \quad (9)$$

Jeśli się przyjmie powyższe oznaczenia, to można zdefiniować macierz $\mathbf{F}_\mathbf{x}$, której elementy, gdy funkcje $f_1(\mathbf{x}), f_2(\mathbf{x}), \dots, f_r(\mathbf{x})$ są zadane, zależą wyłącznie od elementów macierzy $\mathbf{X}$

$$\mathbf{F}_x = \begin{bmatrix} f_1(\mathbf{x}^{(1)}) & f_2(\mathbf{x}^{(1)}) & \dots & f_r(\mathbf{x}^{(1)}) \\ f_1(\mathbf{x}^{(2)}) & f_2(\mathbf{x}^{(2)}) & \dots & f_r(\mathbf{x}^{(2)}) \\ \vdots & \vdots & & \vdots \\ f_1(\mathbf{x}^{(n)}) & f_2(\mathbf{x}^{(n)}) & \dots & f_r(\mathbf{x}^{(n)}) \end{bmatrix}. \quad (10)$$

3. Estymacja parametrów ogólnego równania liniowego regresji metodą najmniejszych kwadratów

Zakładamy, że celem planowania eksperymentów jest estymacja MNK wektora parametrów $\theta = [\theta_1 \ \theta_2 \ \dots \ \theta_r]^T$ lub powierzchni odpowiedzi (por. (5))

$$h(\mathbf{x}, \theta) = \theta_1 f_1(\mathbf{x}) + \dots + \theta_r f_r(\mathbf{x})$$

równania regresji (3).

W teorii optymalnych planów eksperymentów dla estymacji wektora (6) lub funkcji (5) zakłada się m.in., że:

1. Dziedzina D równania (4) jest zbiorem ograniczonym i domkniętym przestrzeni R^m .

2. Zbiór D_1 nazywany obszarem eksperymentowania $D_1 \subset D$ jest zbiorem ograniczonym i domkniętym.

3. Podzbiór D_2 zbioru D , na którym należy oszacować powierzchnię odpowiedzi (5), będzie nazywany obszarem prognozowania. Zbiór D_2 jest zbiorem ograniczonym i domkniętym.

4. Eksperyment mający na celu estymację wektora $\theta = [\theta_1 \ \theta_2 \ \dots \ \theta_r]^T$ realizowany jest według planu eksperymentu P_n o macierzy planowania (1) spełniającej warunek $\det(\mathbf{F}_x^T \mathbf{F}_x) > 0$.

5. $E\varepsilon(\mathbf{x}^{(i)}) = 0$, $i = 1, 2, \dots, n$, co oznacza, że wartości oczekiwane błędów losowych w każdym doświadczeniu są równe 0.

6. $V\varepsilon(\mathbf{x}^{(i)}) = \sigma^2$, $i = 1, 2, \dots, n$, co oznacza, że wariancje błędów losowych są w każdym z n doświadczeń jednakowe i niezależne od wartości zmiennych objaśniających w poszczególnych doświadczeniach.

7. $C[\varepsilon(\mathbf{x}^{(i)}); \varepsilon(\mathbf{x}^{(j)})] = 0$, $1 \leq i < j \leq n$, co oznacza, że błędy losowe poszczególnych doświadczeń nie są ze sobą skorelowane.

Z teorii estymacji wektora (6) MNK nieznanych parametrów θ wynika, że rozwiązanie problemów oszacowania równania regresji (4) wyraża się za pomocą wzoru

$$\hat{\theta} = (\mathbf{F}_x^T \mathbf{F}_x)^{-1} \mathbf{F}_x^T \mathbf{y}, \quad (11)$$

skąd estymator MNK powierzchni odpowiedzi $h(\mathbf{x}; \theta)$ wyraża się odpowiednio wzorem

$$\hat{h}(\mathbf{x}; \theta) = h(\mathbf{x}; \hat{\theta}). \quad (12)$$

4. Zarys teorii planowania eksperymentów dla estymacji MNK parametrów ogólnego modelu regresji liniowej

4.1. Pojęcia podstawowe

Zbiory możliwych wartości $x_1, x_2, \dots, x_m$ zmiennych objaśniających $X_1, X_2, \dots, X_m$ oznaczmy jako

$$\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_m. \quad (13)$$

Niech D_1 oznacza zbiór utworzony z punktów $\mathbf{x} = [x_1 \ x_2 \ \dots \ x_m]$, $x_j \in \mathbf{X}_j$, $j = 1, 2, \dots, m$, takich że w obrębie planowanego eksperymentu istnieje możliwość wykonania doświadczenia przy wartościach zmiennych kontrolowanych ustalonych tak, aby jednocześnie $X_1 = x_1, X_2 = x_2, \dots, X_m = x_m$. Zbiór D_1 nazwiemy obszarem eksperymentowania. Na ogół D_1 jest podzbiorem właściwym iloczynu kartezjańskiego $\mathbf{X}_1 \times \mathbf{X}_2 \times \dots \times \mathbf{X}_m$ (por. [4]).

Zbiór P_n , czyli plan eksperymentu określony przez wzór (2), zawiera te elementy (punkty) obszaru eksperymentowania D_1 , w których powinno się wykonać poszczególne doświadczenia eksperymentu obejmującego n doświadczeń zgodnie z planem P_n .

Plan eksperymentu zdefiniowany zgodnie ze wzorem (2) nie wyklucza sytuacji, w której

$$\mathbf{x}^{(i)} = \mathbf{x}^{(j)}, i \neq j, i, j = 1, 2, \dots, n, \quad (14)$$

co oznacza, że możliwe jest wykonanie większej liczby doświadczeń w jednym punkcie, czyli dla tego samego zbioru wartości zmiennych $X_1, X_2, \dots, X_m$.

Jak już wspomniano, plan eksperymentu P_n definiuje z dokładnością do kolejności wierszy pewną macierz planowania $\mathbf{X}$ o wymiarach $n \times m$ i odwrotnie: dowolna macierz planowania eksperymentu określa odpowiadający jej plan. Często w literaturze zamiast terminu „plan eksperymentu” stosuje się terminy „plan dokładny”, „plan konkretny” lub też „plan realny”.

Oznaczmy przez $\mathcal{P}_n$ rodzinę wszystkich planów eksperymentu o n doświadczeniach w obszarze eksperymentowania $D_1 \subset R^m$. Z kolei niech

$$\mathcal{P} = \bigcup_{n=1}^{\infty} \mathcal{P}_n \quad (15)$$

oznacza rodzinę wszystkich planów eksperymentu, które można zdefiniować na zbiorze D_1 . W praktyce planowania eksperymentów wystarcza zwykle ograniczenie do określonej podrodziny $\mathcal{P}_0 \subset \mathcal{P}$. Ograniczenie to jest spowodowane czynnikami natury zarówno ekonomicznej, jak i matematyczno-statystycznej. Ekonomia nakładów czasu, materiałów i kosztów przeznaczonych na eksperyment zwykle eliminuje plany z nadmierną liczbą doświadczeń. Zaakceptowana metoda statystyczna rozwiązania zagadnienia regresji może dopuszczać tylko takie plany należące do rodziny $\mathcal{P}$, które spełniają dodatkowe warunki wynikające złożeń przyjętej metody. W wyniku tego użyteczna w praktyce rodzina $\mathcal{P}_0$ składa się na ogół z takich planów $\mathcal{P}_n$, w których liczba doświadczeń spełnia nierówność $n_{\min} < n < n_{\max}$ i która ponadto spełnia inne założenia zastosowanej metody statystycznej. Wobec tego

$$\mathcal{P}_0 = \bigcup_{n=n_{\min}}^{n_{\max}} \mathcal{P}_n. \quad (16)$$

4.2. Macierz informacyjna planu

Macierzą informacyjną planu P_n nazywa się macierz

$$\mathbf{M}(P_n) = \frac{1}{n} \mathbf{F}_x^T \mathbf{F}_x. \quad (17)$$

Po uwzględnieniu wzoru (9) można macierz informacyjną planu P_n wyrazić w postaci

$$\mathbf{M}(P_n) = \frac{1}{n} \sum_{k=1}^n \mathbf{f}(\mathbf{x}^{(k)}) \mathbf{f}^T(\mathbf{x}^{(k)}). \quad (18)$$

Na uwagę zasługuje to, iż zarówno macierz kowariancji estymatora $\hat{\theta}$ dana wzorem:

$$\mathbf{C}(\hat{\theta}) = \sigma^2 (\mathbf{F}_x^T \mathbf{F}_x)^{-1}, \quad (19)$$

jak i wariancja estymatora $\hat{h}(\mathbf{x})$ opisana przez wzór

$$V\hat{h}(\mathbf{x}) = \sigma^2 \mathbf{f}^T(\mathbf{x}) (\mathbf{F}_x^T \mathbf{F}_x)^{-1} \mathbf{f}(\mathbf{x}) \quad (20)$$

są funkcjami planu P_n lub jego macierzy planowania. Zauważyć należy również, że prawe strony równań (19) i (20) nie zależą od θ .

Po uwzględnieniu przekształcenia $\mathbf{F}_x^T \mathbf{F}_x = n\mathbf{M}(P_n)$ otrzymujemy wzory na macierz kowariancji estymatora $\hat{\theta}$ oraz na wariancję estymatora $\hat{h}(x)$ postaci:

$$\mathbf{C}(\hat{\theta}) = \frac{\sigma^2}{n} \mathbf{M}^{-1}(P_n), \quad (21)$$

$$V\hat{h}(x) = \frac{\sigma^2}{n} \mathbf{f}^T(x) \mathbf{M}^{-1}(P_n) \mathbf{f}(x). \quad (22)$$

Własności statystyczne estymatora MNK wektora θ oraz powierzchni odpowiedzi $h(x; \theta)$ zależą jedynie od P_n . Ze wzorów (21) i (22) wynika, że stanowi to uzasadniony powód, aby do wyznaczania planów optymalnych posłużyć się kryterium, które będzie pewnym funkcjonałem prawej strony równania (21) lub (22). Funkcjonał taki powinien być pewną miarą dokładności wyznaczanych estymatorów, czyli miarą ich rozrzutu (zmienności). Miary rozrzutu wielowymiarowych estymatorów nieobciążonych są zwykle określane jako pewne funkcje rzeczywiste macierzy kowariancji, dla estymatorów jednowymiarowych zaś miary te są po prostu wariancjami lub funkcjami ich wariancji.

Wzory (21) i (22) opisujące $\mathbf{C}(\hat{\theta})$ i $V\hat{h}(x)$ pozwalają traktować takie miary rozrzutu jako funkcjonały zdefiniowane na danej rodzinie $\mathcal{P}_0 \subseteq \{\mathcal{P}_n \in \mathcal{P} : \det(\mathbf{F}_x^T \mathbf{F}_x) > 0\}$, wartości tych funkcjonałów są zależne od poszczególnych planów $P_n \in \mathcal{P}_0$ poprzez macierz $\mathbf{F}_x^T \mathbf{F}_x$ oraz liczbę doświadczeń n (ponieważ $\mathbf{F}_x$ jest macierzą o n wierszach).

Wynika stąd, iż dla różnych planów $P_n \in \mathcal{P}_0$ dana miara rozrzutu estymatorów $\theta_1, \theta_2, \dots, \theta_r$ osiąga różne wartości, a co za tym idzie – estymatory oszacowane na podstawie różnych planów $P_n \in \mathcal{P}_0$ mają różną dokładność.

Optymalne planowanie eksperymentów w celu estymacji MNK wektora (6) lub funkcji (5) będzie się wiązało z rozwiązaniem zadania optymalizacyjnego polegającego na wyznaczeniu planu $P_{n_0, \psi} \in \mathcal{P}_0$ spełniającego równanie $\psi(P_{n_0, \psi}) = \inf_{P_n \in \mathcal{P}_0} \psi(P_n)$, w którym funkcjonał ψ zależny jest od pewnej miary rozrzutu estymatora MNK, gdzie miara ta zależy od planu P_n z danej rodziny $\mathcal{P}_0$ poprzez liczbę doświadczeń n oraz macierz $\mathbf{F}_x^T \mathbf{F}_x$. Dokładność estymatora $\hat{\theta}$ lub $h(x; \hat{\theta})$ zwiększa się wraz ze zmniejszaniem się wartości funkcjonału ψ .

5. Kryteria optymalności planów

Definicja 1

Plan $P_{n_0, \psi} \in \mathcal{P}_0$ nazywa się ψ -optymalnym planem eksperymentu w klasie planów $\mathcal{P}_0$, jeśli

$$\psi(P_{n_0, \psi}) = \inf_{P_n \in \mathcal{P}_0} \psi(P_n). \quad (23)$$

Funkcjonał ψ , jak już wcześniej wspomniano, pełni funkcję kryterium optymalności planów eksperymentów $P_n \in \mathcal{P}_0$.

Precyzując to kryterium, należy podać *explicite* postać funkcjonału ψ . Szczegółowe kryteria optymalności planów eksperymentów zostaną przedstawione w dalszej części pracy.

Dokładność estymatora można rozumieć na wiele sposobów. Każdemu ze sposobów rozumienia dokładności można przyporządkować pewien specyficzny funkcjonal ψ , który określa się przez podanie *explicite* postaci tegoż funkcjonału. Uzyskane kryterium będzie określać szczegółowo statystyczny sens określonego rozumienia pojęcia dokładności estymatora MNK.

Poszczególne kryteria optymalności oznaczane są zwykle dużymi literami alfabetu łacińskiego w celu odróżnienia ich od ogólnego kryterium ψ . W tej pracy zdefiniowane będą tylko kryteria A - oraz D -optymalności.

Definicja 2

Plan $P_{n_0, A} \in \mathcal{P}_0$ nazywa się A -optymalnym planem eksperymentu w klasie $\mathcal{P}_0$, jeśli

$$\frac{1}{n} \text{tr} \mathbf{M}^{-1}(P_{n_0, A}) = \inf_{P_n \in \mathcal{P}_0} \frac{1}{n} \text{tr} \mathbf{M}^{-1}(P_n). \quad (24)$$

Z definicji 2 wynika, że przedstawiony w niej plan minimalizuje sumę $V\hat{\theta}_1 + V\hat{\theta}_2 + \dots + V\hat{\theta}_r$ (lub średnią arytmetyczną $\frac{1}{r} \sum_{i=1}^r V\hat{\theta}_i$) wariancji estymatorów

MNK parametrów $\theta_1, \theta_2, \dots, \theta_r$ w ogólnym modelu liniowym regresji. Ze wzoru (21) widać wyraźnie, że

$$V\hat{\theta}_1 + V\hat{\theta}_2 + \dots + V\hat{\theta}_r = \frac{\sigma^2}{n} \text{tr} \mathbf{M}^{-1}(P_n), \quad (25)$$

gdzie σ^2 jest stałą niezależną od P_n .

Definicja 3

Plan $P_{n_0, D} \in \mathcal{P}_0$ nazywa się D -optymalnym planem eksperymentu w klasie $\mathcal{P}_0$, jeśli

$$\frac{1}{n^r} \det \mathbf{M}^{-1}(P_{n_0}, D) = \inf_{P_n \in \mathcal{P}_0} \frac{1}{n^r} \det \mathbf{M}^{-1}(P_n). \quad (26)$$

Z definicji 3 wynika, że plan D -optymalny maksymalizuje dokładność estymacji wektora parametrów $\theta_1, \theta_2, \dots, \theta_r$ w takim sensie, że plan taki minimalizuje objętość r -wymiarowej elipsoidy koncentracji

$$(\hat{\theta} - \theta)^T n \mathbf{M}(P_n) (\hat{\theta} - \theta) \leq \sigma^2 (r + 2) \quad (27)$$

łącnego rozkładu estymatorów $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_r$.

Objętość opisanej elipsoidy jest proporcjonalna do pierwiastka kwadratowego z wyrażenia

$$\left\{ \det \left[\frac{n}{\sigma^2} \mathbf{M}(P_n) \right] \right\}^{-1} = \frac{\sigma^2}{n^r} \det \mathbf{M}^{-1}(P_n), \quad (28)$$

czyli z wyrażenia $\det \mathbf{C}(\hat{\theta})$. Wyrażenie (29) zostało nazwane przez Wilksa (1932 r.) wariancją uogólnianą rozkładu r -wymiarowego, a wielkość

$$\left[\sigma^{-2r} \det \mathbf{C}(\hat{\theta}) \right]^{0,5} = \left[\frac{1}{n^r} \det \mathbf{M}^{-1}(P_n) \right]^{0,5} \quad (29)$$

Frish (1929 r.) określił jako współczynnik rozrzutu rozkładu wielowymiarowego. Mówiąc inaczej, plan D -optymalny minimalizuje wariancję uogólnioną lub współczynnik rozrzutu r -wymiarowego rozkładu estymatorów $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_r$.

W niektórych przypadkach zamiast funkcjonału (26) wygodniejsze jest stosowanie funkcjonału

$$D'(\mathbf{M}(P_n), n) = \ln [D(\mathbf{M}(P_n), n)] = -r \ln(n) - \ln [\det \mathbf{M}(P_n)]. \quad (30)$$

Lokalizacja ekstremów obu funkcjonałów jest jednakowa, ponieważ funkcjonał (30) jest monotonicznym przekształceniem funkcji $D(\mathbf{M}(P_n), n)$. Zgodnie z definicją 3 plan D -optymalny jest planem spełniającym warunek:

$$-r \ln(n) - \ln [\det \mathbf{M}(P_{n_0}, D)] = \inf_{P_n \in \mathcal{P}_0} \left\{ -r \ln(n) - \ln [\det \mathbf{M}(P_n)] \right\}. \quad (31)$$

6. Kosztowo-optymalne plany eksperymentów

W planowaniu optymalnym eksperymentów, kierując się określonymi kryteriami, dąży się do wyznaczenia takich planów, które cechują się możliwie największą statystyczną dokładnością szacowanych parametrów równania regresji opisanego wzorem

$$Y(\mathbf{x}) = h(\mathbf{x}, \theta) + \varepsilon(\mathbf{x}). \quad (32)$$

W klasycznym podejściu do optymalnego planowania eksperymentów nie jest uwzględniony aspekt kosztowy. Wyznacza się plany spełniające ustalone kryteria optymalności, nie biorąc pod uwagę kosztu realizacji tych planów.

Na całkowity koszt eksperymentu wpływają takie same czynniki jak na dokładność planów eksperymentów, tzn. liczba doświadczeń oraz lokalizacja poszczególnych doświadczeń w obszarze eksperymentowania.

W teorii planów kosztowo-optymalnych można wyróżnić dwa przypadki. Pierwszy z nich, łatwiejszy, zachodzi, kiedy to całkowity koszt eksperymentu wzrasta proporcjonalnie do liczby doświadczeń. Oznacza to, że koszt przeprowadzenia obserwacji w każdym punkcie obszaru eksperymentowania D_1 jest taki sam. W takiej sytuacji proces kosztowej optymalizacji planu eksperymentu polega na warunkowej minimalizacji odpowiedniego kryterium optymalizacyjnego planu przy warunku ubocznym

$$nK \leq k_0, \quad (33)$$

gdzie: n – liczba doświadczeń,

K – koszt przeprowadzenia pojedynczej obserwacji,

k_0 – całkowity nakład kosztowy przeznaczony na przeprowadzenie eksperymentu.

W przypadku drugim problem staje się nieco bardziej skomplikowany, ponieważ całkowity koszt eksperymentu zależy dodatkowo od położenia punktów, w których wykonywane są obserwacje, w obszarze eksperymentowania. Inaczej mówiąc, koszt wykonania pojedynczej obserwacji nie jest taki sam w każdym punkcie obszaru D_1 . Doświadczenia wykonywane w różnych punktach powodują różne wielkości ponoszonych kosztów jednostkowych. Wtedy całkowity koszt eksperymentu przedstawia się jako sumę kosztów $k(x_i)$, $i = 1, 2, \dots, n$, poszczególnych doświadczeń:

$$K(P_n) = \sum_{i=1}^n k(x_i). \quad (34)$$

W poszukiwaniu planów kosztowo-optymalnych można wyróżnić dwa podejścia do rozwiązania tego problemu:

1. Poszukiwanie planu $P_{n,\psi}^* \in \mathcal{P}_0$, gdzie $\mathcal{P}_0 = \bigcup_{n=n_{\min}}^{n_{\max}} P_n$, dla którego $\psi(P_n) \leq \psi_0$

i całkowity koszt $K(P_{n,\psi}^*) = \inf_{P_{n,\psi} \in \mathcal{P}_0}$. Przy tym ψ_0 jest założoną z góry wartością dopuszczalną funkcjonału ψ .

2. Poszukiwanie planu $P_{n,\psi}^* \in \mathcal{P}_0$, gdzie $\mathcal{P}_0 = \bigcup_{n=n_{\min}}^{n_{\max}} P_n$, którego $K(P_{n,\psi}^*) \leq k_0$ i

$\psi(P_{n_0,\psi}) = \inf_{P_{n,\psi} \in \mathcal{P}_0} \psi(P_n)$. Przy tym $\psi(P_n)$ jest wartością funkcjonału ψ dla planu

P_n stanowiącego kryterium optymalności planów $P_n \in \mathcal{P}_0$.

Przykład 1

Przykład dotyczy sytuacji, w której poszukiwany będzie kosztowo-optymalny plan w przypadku, gdy całkowity koszt eksperymentu wzrasta proporcjonalnie do liczebności próby. Wykonanie obserwacji w każdym miejscu obszaru eksperymentowania jest tak samo kosztowne.

W tym przykładzie równanie regresji jest równaniem z jedną zmienną objaśniającą

$$h(x) = \theta_0 + \theta_1 x. \quad (35)$$

Zakłada się, że obszarem eksperymentowania jest zbiór $D_1 = \{-1; 1\}$.

Przez $\mathcal{P}_0$ w tym przykładzie oznaczony będzie zbiór wszystkich planów o spektrum ze zbioru $D_1 = \{-1; 1\}$, tzn. elementami rodziny $\mathcal{P}_0$ są plany postaci

$$P_n = \begin{bmatrix} -1 & 1 \\ n_1 & n_2 \\ n_1 + n_2 & n_1 + n_2 \end{bmatrix}, \quad (36)$$

gdzie $n_1 + n_2 = n$.

Poszukiwany jest plan $P_n^* \in \mathcal{P}_0$, dla którego suma wariancji estymatorów parametrów równania (35) będzie mniejsza niż $0,2\sigma^2$. Innymi słowy,

$$V\hat{\theta}_0 + V\hat{\theta}_1 \leq 0,2\sigma^2. \quad (37)$$

Plany, których dokładność wyrażona jest przez kryterium dotyczące sumy wariancji estymatorów rozpatrywanego równania regresji, związane są z kryterium A -optymalności.

Wyrażając warunek (37) w formie macierzy informacyjnej planu P_n , można ten warunek przedstawić w postaci

$$\frac{1}{n} \text{tr} \mathbf{M}(P_n)^{-1} \leq 0,2. \quad (38)$$

Zadanie w tym przypadku polega na wyznaczeniu planu $P_n^* \in \mathcal{P}_0$ o najmniejszej możliwej liczbie doświadczeń n , który spełniać będzie kryterium dokładności (38). Zakładamy dodatkowo, że $\mathcal{P}_0$ zawiera tylko takie plany, dla których istnieje $\mathbf{M}(P_n)^{-1}$ oraz zawiera co najmniej jeden plan spełniający kryterium (38).

Procedura poszukiwania planu spełniającego powyższe warunki polega na poszukiwaniu planu o założonej dokładności, poczynając od planu o możliwie najmniejszej liczbie doświadczeń n_1 . W przypadku, gdy $\psi_A(\mathbf{M}(P_{n_1})) \leq 0,2$ przerywamy procedurę i stwierdzamy, że plan $P_n^* = P_{n_1}^*$ jest rozwiązaniem sformułowanego zadania, czyli najmniej kosztownym planem spełniającym kryterium dokładności (38). W przypadku przeciwnym, gdy $\psi_A(\mathbf{M}(P_{n_1})) > 0,2$, zwiększamy i o 1, szukając dalej rozwiązania. Procedurę powtarzamy do momentu uzyskania planu o zadanej dokładności zdefiniowanej przez kryterium (38).

Przy zdefiniowanym wcześniej spektrum oraz ze względu na kryterium A -optymalności wystarczy rozważyć plany postaci

$$P_{n_i} = \begin{cases} \underbrace{(-1, \dots, -1)}_{n_i/2}, \underbrace{(1, \dots, 1)}_{n_i/2}, & \text{gdy } n_i \text{ jest liczbą parzystą,} \\ \underbrace{(-1, \dots, -1)}_{(n_i \pm 1)/2}, \underbrace{(1, \dots, 1)}_{(n_i \pm 1)/2}, & \text{gdy } n_i \text{ jest liczbą nieparzystą,} \end{cases} \quad (39)$$

dla ustalonej liczby doświadczeń n_i .

Z wyników teorii klasycznych planów optymalnych wiadomo bowiem, że dla spektrum o dwóch elementach A -optymalne plany dla liniowej funkcji regresji mają jednakową liczbę obserwacji w każdym punkcie spektrum.

Wartość funkcjonau ψ_A stosowanego jako kryterium A -optymalności, dla planów postaci (40) jest następująca

$$\psi_A(\mathbf{M}(P_{n_i})) = \begin{cases} \frac{2}{n_i}, & \text{gdy } n_i \text{ jest liczbą parzystą,} \\ \frac{2}{(n_i - 1)}, & \text{gdy } n_i \text{ jest liczbą nieparzystą.} \end{cases} \quad (40)$$

Zgodnie z omówioną procedurą rozpoczynamy poszukiwanie planu spełniającego przedstawione warunki, od liczebności $n_1 = 2$, dla której $\psi_A(\mathbf{M}(P_2^*)) = 1$. Oznacza to, że plan obejmujący $n_1 = 2$ doświadczenia nie spełnia warunku (38). Gdy postępuje się dalej zgodnie z procedurą, okazuje się, że po raz pierwszy warunek dokładności spełniony jest dla liczebności $n_9 = 10$, przy której $\psi_A(\mathbf{M}(P_{n_i}^*)) = 0,2$.

Poszukiwany plan ma więc postać

$$n^* = 10, P_n^* = (-1, -1, -1, -1, -1, 1, 1, 1, 1, 1). \quad (41)$$

Przykład 2

W tym przykładzie zaprezentowana zostanie procedura poszukiwania D -kosztowo-optymalnego planu w przypadku, gdy całkowity koszt eksperymentu mierzony jest funkcją kosztów postaci

$$K(P_n) = \sum_{i=1}^n x_i^2, \quad (42)$$

natomiast równanie regresji ma postać

$$h(x) = \theta_0 + \theta_1 x. \quad (43)$$

W tym przypadku macierz informacyjna dowolnego planu P_n ma postać

$$\mathbf{M}(P_n) = \begin{bmatrix} 1 & \bar{x} \\ \bar{x} & \frac{1}{n} \sum_{i=1}^n x_i^2 \end{bmatrix}. \quad (44)$$

Zakłada się tutaj, że obszar eksperymentowania $D_1 = [-1, 1]$. Przez $\mathcal{P}_0$ oznaczony zostanie zbiór wszystkich planów P_n , gdzie $n \geq 2$ określonych na odcinku $[-1; 1]$ o dokładnie określonych macierzach informacyjnych.

Zadanie polega na wyznaczeniu planu $P_n^* \in \mathcal{P}_0$, który będzie spełniał nierówność (45) odnoszącą się do definicji D -optymalności planu (por. def. (3))

$$\det(\mathbf{M}(P_n)^{-1}) \leq 2, \text{ tzn. } \det(\mathbf{M}(P_n)) \geq 0,5 \quad (45)$$

i jednocześnie będzie minimalizował funkcję kosztów (42)

Ze wzoru (44) wynika wówczas, że

$$\det(\mathbf{M}(P_n)) = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2 \geq 0,5. \quad (46)$$

Z tak sformułowanych ograniczeń wynika, że koszty eksperymentu muszą spełniać warunek

$$K(P_n) = \sum_{i=1}^n x_i^2 \geq n(0,5 + \bar{x}^2) \geq 1 \text{ dla } P_n \in \mathcal{P}_0. \quad (47)$$

Zatem minimalny koszt eksperymentu, który spełnia założone warunki dokładności, wynosi $K(P_n) = 1$. Minimum to uzyskane jest dla $n = 2$ doświadczeń. Wówczas z (42) oraz (47) otrzymujemy

$$\begin{cases} x_1^2 + x_2^2 = 1, \\ 2(0,5 + \bar{x}^2) = 1, \end{cases} \quad (48)$$

skąd $\bar{x} = 0$, czyli $x_1 = -x_2$. Wobec tego

$$2x_1^2 = 2x_2^2 = 1, \quad (49)$$

a więc

$$x_1^2 = x_2^2 = \frac{1}{2}, \quad (50)$$

czyli

$$x_1 = -\frac{\sqrt{2}}{2}, \text{ zaś } x_2 = \frac{\sqrt{2}}{2}.$$

Ostatecznie plan kosztowo-optimalny, spełniający założenia warunku (45), ma postać

$$P_2^* = \left\{ -\frac{1}{2}\sqrt{2}, \frac{1}{2}\sqrt{2} \right\}. \quad (51)$$

Dla porównania wyznaczony zostanie również plan D -optimalny o liczbie doświadczeń $n = 2$. Plan ten, zgodnie z definicją 3 i wzorem (45), maksymalizuje wyrażenie

$$\det(\mathbf{M}(P_n)) = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2. \quad (52)$$

Dla $n = 2$ otrzymujemy

$$\det(\mathbf{M}(P_n)) = \frac{x_1^2 + x_2^2}{2} - \left(\frac{x_1 + x_2}{2} \right)^2, \quad (53)$$

skąd po przekształceniach uzyskujemy

$$\det(\mathbf{M}(P_n)) = (x_1 - x_2)^2. \quad (54)$$

Uwzględniając fakt, iż plany eksperymentów realizowane są na odcinku $[-1, 1]$, wyrażenie (54) osiąga wartość maksymalną dla $x_1 = -1$ oraz $x_2 = 1$, z czego wynika, że szukany plan D -optimalny ma postać

$$P_{2,D} = \{-1, 1\}. \quad (55)$$

Warto zauważyć, że jeśli zaostrzymy warunki dotyczące dokładności planu, zastępując nierówność (46) przez warunek

$$\det(\mathbf{M}(P_n)^{-1}) \leq 1, \text{ tzn. } \det(\mathbf{M}(P_n)) \geq 1, \quad (56)$$

to minimalny koszt eksperymentu wynosi 2 i eksperyment taki obejmuje również $n = 2$ doświadczeń. Ze wzorów (46), (52) oraz (56) mamy bowiem

$$K(P_n) = \sum_{i=1}^n x_i^2 \geq n(1 + \bar{x}^2) \geq 2. \quad (57)$$

Po zastąpieniu nierówności w wyrażeniu (57) przez znak równości otrzymujemy równanie

$$x_1^2 + x_2^2 = 2, \quad (58)$$

skąd $x_1 = -1$, $x_2 = 1$ i szukany plan w ostateczności ma postać

$$P_2^* = \{-1, 1\} \quad (59)$$

i jest on identycznym planem jak wcześniej uzyskany D -optymalny plan $P_{2,D}$.

Na uwagę zasługuje to, iż uwzględnienie czynnika kosztowego w procedurze wyznaczania planów optymalnych w sposób znaczący wpływa na ich strukturę. Powoduje to, że plany optymalne różnią się na ogół od planów kosztowo-optymalnych. Niekiedy jednak, jak to pokazano w ostatniej części przykładu 2, plan optymalny może być równocześnie planem kosztowo-optymalnym.

Literatura

- [1] Bandemer H. i in., *Theorie und Anwendung der optimalen Versuchsplanung I*, Akademie Verlag, Berlin 1977.
- [2] Bandemer H., Nather W., *Theorie und Anwendung der optimalen Versuchsplanung II*, Akademie Verlag, Berlin 1980.
- [3] Jung W., *Kostenoptimale Versuchsplanung im Regressionsmodell*, „Biometrical Journal” 1974, nr. 16, s. 31-47.
- [4] Wawrzynek J., *Statystyczne planowanie eksperymentów w zagadnieniach regresji w warunkach małej próby*, AE, Wrocław 1994.

THE SELECTED ISSUES OF COST-OPTIMAL DESIGN OF EXPERIMENTS IN REGRESSION

Summary

This paper presents the bases of the theory of cost-optimal designs of experiment for the estimation of parameters of regression model. In this theory we can differentiate two cases. In the first case the total cost of experiment depends only on the number of observations. In the second case the total cost of experiment is additionally dependent on the allocation of points in the experimental area, in which the observations are done. In the article it is shown that the cost-optimal designs can be obtained according to two approaches. Both approaches are illustrated with examples.

Mgr inż. Łukasz Kuźmiński jest doktorantem w Katedrze Statystyki w Akademii Ekonomicznej we Wrocławiu.