КІБЕРНЕТИКА та КОМП'ЮТЕРНІ ТЕХНОЛОГІЇ

УДК 519.85, 51-76 DOI:10.34229/2707-451X.23.3.3

П.І. СТЕЦЮК, М.М. БУДНИК, І.О. СЕНЬКО, В.О. СТОВБА, І.А. ЧАЙКОВСЬКИЙ

ВИКОРИСТАННЯ МЕТОДУ ЕЛІПСОЇДІВ ДЛЯ ВИВЧЕННЯ ЗВ'ЯЗКІВ У МЕДИЧНИХ ДАНИХ

Робота присвячена використанню методу еліпсоїдів для вивчення зв'язків у медичних даних. На його основі реалізовано програму emlmp, метою якої є знаходження параметрів лінійної регресійної моделі для прогнозування психологічних індикаторів пацієнтів на основі кардіологічних даних. Продемонстровано переваги критерію на основі методу найменших модулів (МНМ) як порівняти з критерієм на основі методу найменших квадратів (МНК) за результатами обчислювальних експериментів. Запропоновано метрику для оцінки узгодженості масиву даних, яка дозволяє оцінити узгодженість кожного параметра окремо. Знайдено лінійний зв'язок між 4-ма психологічними параметрами та максимальною точністю регресійних моделей при оптимальній кількості параметрів у зазначених моделях.

Ключові слова: *лінійна регресія, опукла функція, метод еліпсоїдів, метод найменших модулів, прогнозування даних, GNU Octave.*

Вступ. Перебування України у стані широкомасштабної війни з усіма негативними соціальними наслідками призвело до погіршення психоемоційного стану різних верств населення та виникненню психогенних розладів. Суттєво зростає інтенсивність низки станів, які можна трактувати як переддепресійні. На фоні цих подій вкрай важливим завданням є професійна психологічна діагностика з використанням наявних та вимірних даних, які можна отримати. Це дозволить не лише вчасно поставити коректний діагноз, а й провести відповідне лікування і терапію.

За допомогою інноваційного програмно-апаратного комплексу, розробленого в ІК НАНУ [1], обстежено 90 військовослужбовців, які перебували на санаторному лікуванні у ЦВКС «Хмільник». Об'єктивна оцінка стану здійснювалася на основі поглибленого аналізу малих змін електрокардіограми та варіабельності ритму серця за методом, описаним у статті [2]. Результати аналізу ЕКГ порівнювалися з формалізованим висновком психолога та загальновідомої шкали Бека для оцінки тривожності. Мета цієї роботи – розробити математичний апарат для прогнозування психологічних висновків на основі кардіологічних даних.

У розділі 1 наведено опис методу еліпсоїдів для знаходження параметрів лінійної регресії з критерієм МНМ у степені p. Розділ 2 присвячено Octave-програмі **emlmp**, що реалізує цей метод, та результатам двох експериментів з її використанням. В розділі 3 описано механізм відбору змінних для найкращого прогнозування психологічного стану на основі кардіологічних даних. Розділ 4 містить результати обчислювального експерименту з використанням програми **emlmp** для критеріїв МНМ та МНК. У п'ятому розділі запропоновано метрику для оцінки узгодженості масиву даних, що дозволяє оцінити узгодженість для кожного параметра окремо. Знайдено лінійний зв'язок між 4-ма психологічними параметрами та максимальною точністю регресійних моделей при оптимальній кількості параметрів у зазначених моделях.

1. Метод еліпсоїдів для знаходження параметрів лінійної регресії за критерієм найменших модулів у степені p $(1 \le p \le 2)$. Нехай для оцінки n невідомих параметрів $x_1,..,x_n$ використовується m спостережень $y_1,..,y_m$ причому ці величини пов'язані співвідношенням:

$$y_i = \sum_{j=1}^{n} a_{ij} x_j + u_i, \quad i = 1,...,m, \tag{1}$$

де a_{ij} – відомі коефіцієнти, u_i – невідомі випадкові величини, що мають (приблизно) однакові функції розподілу, $m > n$. Рівняння (1) можна записати в матричній формі:

$$y = Ax + u, \tag{2}$$

де $y = (y_1,..,y_m)^T \in \mathbf{R}^m$ і $u = (u_1,..,u_m)^T \in \mathbf{R}^m$ – m-вимірні вектори, A – матриця розміру $m \times n$, $x = (x_1,..,x_n)^T \in \mathbf{R}^n$ – m-вимірний вектор параметрів, які потрібно оцінити.

Метод найменших модулів у степені p (відповідає знаходженню невідомого вектора x_p^* згідно з критерієм найменших модулів у степені p, де $1 \le p \le 2$) – задача математичного програмування:

$$f_{LMP}^* = f_{LMP}(x_p^*) = \min_{x \in \mathbf{R}^n} \left\{ f_{LMP}(x) = \sum_{i=1}^{m} \left| y_i - \sum_{j=1}^{n} a_{ij} x_j \right|^p \right\}, \tag{3}$$

де $|\cdot|$ – модуль (абсолютна величина) числа. Функція $f_{LMP}(x)$ є негладкою, якщо $p = 1$, та гладкою якщо $p > 1$.

Задача (3) – безумовна задача мінімізації опуклої функції $f_{LMP}(x)$, субградієнт якої у точці $\bar{x}$ обчислюється за такою формулою:

$$g_{f_{LMP}}(\bar{x}) = \begin{pmatrix} p \sum_{i=1}^{m} sign\left(\sum_{j=1}^{n} a_{ij} \bar{x}_j - y_i \right) \left| \sum_{j=1}^{n} a_{ij} \bar{x}_j - y_i \right|^{p-1} a_{i1}, \\ p \sum_{i=1}^{m} sign\left(\sum_{j=1}^{n} a_{ij} \bar{x}_j - y_i \right) \left| \sum_{j=1}^{n} a_{ij} \bar{x}_j - y_i \right|^{p-1} a_{i2}, \\ \dots \\ p \sum_{i=1}^{m} sign\left(\sum_{j=1}^{n} a_{ij} \bar{x}_j - y_i \right) \left| \sum_{j=1}^{n} a_{ij} \bar{x}_j - y_i \right|^{p-1} a_{in} \end{pmatrix}. \tag{4}$$

Якщо $p = 1$, то задача (3) переходить в задачу математичного програмування:

$$f_{LM}^* = \min_{x \in \mathbf{R}^n} \left\{ f_{LM}(x) = \sum_{i=1}^{m} \left| y_i - \sum_{j=1}^{n} a_{ij} x_j \right| \right\}. \tag{5}$$

Задача (5) – безумовна задача мінімізації опуклої кусково-лінійної функції $f_{LM}(x)$. Вона відповідає методу найменших модулів, який є робастним до аномальних спостережень або «викидів» [3–6].

Знаходження найкращого за критерієм найменших модулів вектора x^*, де x^* – розв'язок задачі (5), можна звести до розв'язання наступної ЛП-задачі: знайти

$$f_{LM}^* = \min_{z\in\mathbf{R}^n, z\geq 0} \sum_{i=1}^{m} z_i \text{ за обмежень } y_i - \sum_{j=1}^{n} a_{ij}x_j \leq z_i, -y_i + \sum_{j=1}^{n} a_{ij}x_j \leq z_i, i = 1,\ldots,m. \quad (6)$$

Для розв'язання ЛП-задачі (6) можна використовувати відповідні стандартні програми лінійного програмування. При цьому паралельно зі знаходженням самого вектора x^* знаходяться і оптимальні значення вектора $z^* = \left(z_1^*,..,z_m^*\right)^T$, компоненти якого задають оцінки для незалежної випадкової величини $u_i, i = 1,..,m.$

Якщо $p = 2$, то задача (3) переходить у задачу математичного програмування:

$$f_{LS}^* = \min_{x\in\mathbf{R}^n} \left\{ f_{LS}(x) = \sum_{i=1}^{m} \left(y_i - \sum_{j=1}^{n} a_{ij}x_j \right)^2 \right\}. \quad (7)$$

Задача (7) – безумовна задача мінімізації опуклої квадратичної функції $f_{LS}(x)$. Вона відповідає методу найменших квадратів. Якщо рядки матриці A лінійно незалежні, то задача (7) має аналітичний розв'язок $x^* = (A^T A)^{-1} Ay$.

Для розв'язання задачі (3) може бути застосований класичний метод еліпсоїдів [7, 8]. Його реалізація алгоритмом **emshor** для задачі мінімізації опуклої функції, мінімум якої знаходиться всередині кулі заданого радіуса, описана в [9]. Далі наведемо застосування алгоритму **emshor** для задачі мінімізації функції $f_{LMP}(x)$ при умові, що її точка мінімуму x_p^* локалізована у n-вимірній кулі радіуса r_0 з центром у точці $x_0 \in R^n$, тобто $\left\|x_0 - x_p^*\right\| \leq r_0$.

Алгоритм знаходження x_p^*. Вхідним параметром є величина $\varepsilon_f > 0$ – точність, з якою необхідно знайти $f_{LMP}^* = f(x_p^*)$.

Ініціалізація. Розглянемо $n\times n$-матрицю B і покладемо $B_0 := I_n$, де I_n – одинична $n\times n$-матриця. Перейдемо до першої ітерації зі значеннями x_0, r_0 і B_0.

Нехай на k-й ітерації знайдені значення $x_k \in R^n$, r_k, B_k. Перехід до $(k+1)$-ї ітерації полягає у такій послідовності дій.

Крок 1. Обчислимо $f_{LMP}(x_k)$ та субградієнт $g_{f_{LMP}}(x_k)$ у точці x_k за формулою (4). Якщо $r_k \left\|B_k^T g_{f_{LMP}}(x_k)\right\| \leq \varepsilon_f$, то "Зупинка: $k^* = k$ і $x_p^* = x_k$ ". Інакше переходимо до кроку 2.

Крок 2. Покладемо $\xi_k := \dfrac{B_k^T g_{f_{LMP}}(x_k)}{\left\|B_k^T g_{f_{LMP}}(x_k)\right\|}$.

Крок 3. Обчислимо чергову точку

$$x_{k+1} := x_k - h_k B_k \xi_k, \quad \text{де} \quad h_k = \frac{1}{n+1} r_k.$$

Крок 4. Обчислимо

$$B_{k+1} := B_k + \left(\sqrt{\frac{n-1}{n+1}} - 1 \right) (B_k \xi_k) \xi_k^T \quad \text{і} \quad r_{k+1} := r_k \frac{n}{\sqrt{n^2-1}}.$$

Крок 5. Переходимо до $(k+1)$-ї ітерації зі значеннями x_{k+1}, r_{k+1}, B_{k+1}.

Збіжність алгоритму знаходження x_p^* забезпечує теорема.

Теорема. Послідовність точок $\{x_k\}_{k=0}^{k^*}$ задовольняє нерівностям

$$\left\|B_k^{-1}(x_k - x_p^*)\right\| \le r_k, \quad k = 0,1,2,\ldots,k^*.$$

На кожній ітерації $k>0$ величина зменшення об'єму еліпсоїда $E_k = \left\{x \in R^n : \left\|B_k^{-1}(x_k - x)\right\| \le r_k\right\}$, в якому локалізована точка x_p^*, є величиною сталою і рівною

$$q = \frac{vol(E_k)}{vol(E_{k-1})} = \sqrt{\frac{n-1}{n+1}}\left(\frac{n}{\sqrt{n^2-1}}\right)^n < \exp\left\{-\frac{1}{2(n+1)}\right\} < 1.$$

З теореми випливає, що алгоритм знаходження x_p^* можна успішно застосовувати на сучасних комп'ютерах, якщо $n = 10 \div 30$ та $m = 100 \div 1000$. Дійсно, для зменшення в 10 разів об'єму еліпсоїда, в якому локалізовано точку x_p^*, потрібно зробити K ітерацій, де $K = -\frac{\ln 10}{\ln q} \approx (2\ln 10)(n+1) \approx 4.6(n+1)$, тобто, щоб на порядок покращити відхилення знайденого рекордного значення функції $f_{LMP}(x)$ від її оптимального значення f_{LMP}^* потрібно зробити $4.6(n+1)^2$ ітерацій алгоритму знаходження x_p^*.

Якщо $n = 30$ та $\varepsilon_f = 10^{-10} \times f(x_0)$, то максимальна кількість ітерацій алгоритму буде рівною $46(n+1)^2 = 46 \times 961 = 44206$. Тому, навіть прямолінійна матрично-векторна реалізація обчислення значення функції $f_{LMP}(x)$ та її субградієнта за формулою (4) дозволяє забезпечити швидку роботу алгоритму на сучасних комп'ютерах. Далі це підтвердимо результатами обчислювальних експериментів з використанням процесора Intel Core i5-9400f, 2.9 GHz, 16GB RAM та мови GNU Octave 5.1.0.

2. Octave програма emlmp та обчислювальні експерименти. Алгоритм знаходження x_p^* реалізовано octave програмою **emlmp** (**e**llipsoid **m**ethod **l**east **m**oduli to the power of **p**), код якої наведено далі.

```
function [xp,fp,itn,ist] = emlmp(A,y,p,x0,r0,epsf,maxitn,intp);        #row01
   n=columns(A); xp=x0; B=eye(n); r=r0;                                 #row02
   dn=double(n); beta=sqrt((dn-1.d0)/(dn+1.d0));                        #row03
   for (itn = 0:maxitn)                                                 #row04
      temp = A*xp - y; fp = sum(abs(temp).^p);                          #row05
      if((mod(itn,intp)==0)&&(intp<=maxitn))                            #row06
        printf(" itn %4d  fp %14.6e\n",itn,fp);                         #row07
      endif                                                             #row08
      g1 = p*A'*(sign(temp).*(abs(temp)).^(p-1));                       #row09
      g = B'*g1; dg = norm(g);                                          #row10
      if(r*dg < epsf) ist = 1;  return; endif                           #row11
      xi = (1.d0/dg)*g; dx = B * xi;                                    #row12
      hs = r/(dn+1.d0); xp -= hs * dx;                                  #row13
      B += (beta - 1) * B * xi * xi';                                   #row14
      r = r/sqrt(1.d0-1.d0/dn)/sqrt(1.d0+1.d0/dn);                      #row15
   endfor                                                               #row16
   ist = 4;                                                             #row17
endfunction                                                             #row18
```

Ядро програми **emlmp** зосереджене в циклі for (рядки 4–16). Спочатку обчислюється значення функції f (рядок 5) та її нормованого субградієнта у точці x_p (рядок 10). Якщо виконується умова зупинки (рядок 11) – алгоритм припиняє свою роботу. Інакше обчислюється наступна точка x_{k+1} (рядок 13), перераховуються матриця перетворення простору B_{k+1} (рядок 14) та радіус r_{k+1} (рядок 15).

Програма має такі вхідні параметри: A, y, p – стартові дані для МНМ степеня p $(1 \le p \le 2)$; x_0 – стартова точка; r_0 – радіус кулі з центром у точці x_0, яка локалізує точку мінімуму; ε_f, $maxitn$ – параметри зупинки (точність за значенням функції, що мінімізується, максимальна кількість ітерацій); $itnp$ – інтервал виводу (через кожні $itnp$ ітерацій).

Вихідні параметри: x_p – наближення до точки мінімуму; f_p – значення функції f у точці x_p; itn – кількість виконаних ітерацій; ist – код виходу $\left(1 = \varepsilon_f, 4 = maxitn\right)$.

Далі наведемо результати двох обчислювальних експериментів для розв'язання тестових задач (3) з кількістю невідомих параметрів $n \in \{10, 20, 30\}$, де кількість спостережень в десять раз більше ніж кількість параметрів, тобто $m = 10 \times n$. Мета першого експерименту – оцінити час розв'язання задач (3) для вказаних параметрів на персональному комп`ютері з процесором Intel Core i5-9400f, 2.9 GHz, 16GB RAM. Мета другого експерименту – продемонструвати робастність методу найменших модулів, а значить і розв'язків задачі (3), якщо p є близьким до одиниці.

Для першого експерименту матриця A та вектор y – вхідні дані для задачі (3) генерувалися випадковим чином з стандартним рівномірним розподілом $U(0,1)$ за такими формулами: **A = = 10*rand(m,n), y = A*xstar(n,1), xstar(n,1) = round(10*rand(n,1))**. Стартова точка x_0 вибиралась за правилом **x0(n,1) = round(5*rand(n,1))**, а радіус кулі, в якій локалізована точка $x_p^* = x_{star}$, вибирався за правилом **r0 = 5*norm(x0–xstar)**, тобто $r_0 = \|x_0 - x_{star}\|$. Перший експеримент реалізує наведений далі Octave-код.

```
# Test 1: emlmp runnning time for n = 30 and m = 10*n

n = 30, m = 10*n,
rand("seed", 2023);
A = 10*rand(m,n);

xstar=round(10.0*rand(n,1)+0.5); y = A*xstar;
x0 = round(5.0*rand(n,1)); r0 = 5*norm(x0 - xstar),
maxitn = 100000, intp = 100001,

printf("\n Test 1: emlmp runnning time for n = 30 and m = 10*n \n");
epsf0 = 1.e-6; ntest = 5; table = [];
for (i = 1:ntest)
   p = 1.d0 + (i - 1.d0)/(ntest - 1.d0),
   epsf = epsf0**(p); time0 = time();
   [xp,fp,itn,ist] = emlmp(A,y,p,x0,r0,epsf,maxitn,intp);
   time1 = time() - time0,
   dx = norm(xp-xstar);
   table = [table; p epsf time1 itn  ist fp dx];
endfor

n,m,
printf("   p     epsf   time   itn  ist      fp           dx  \n");
for (i = 1:ntest)
   printf(" %4.1f  %6.1e  %4.2f %6d %2d   %10.5e %10.1e\n",table(i,1:7))
endfor
```

Результати роботи програми **emlmp** для першого експерименту – затрати часу (time) на розв'язання задач (3) з точністю ε_f, кількість ітерацій методу itn, знайдене мінімальне значення функції f_p, dx – норма відхилення знайденого наближення до точки мінімуму від відомої точки мінімуму **xstar** наведено в табл. 1. Тут ε_f вибирається таким чином: при $p=1$ значення $\varepsilon_f = 10^{-6}$, при $p>1$ – вибираємо $\varepsilon_f = \left(10^{-6}\right)^p$.

ТАБЛИЦЯ 1. Перший експеримент: результати розв'язання задач (3) для різних n та m

n, m	p	ε_f	$time\,(sec)$	itn	f_p	dx
$n=10,\ m=100$	1.0	1.0e–06	0.34	4740	5.25552e–08	2.2e–10
	1.2	3.2e–08	0.37	4474	6.72815e–10	4.2e–10
	1.5	1.0e–09	0.33	4298	2.72608e–11	1.4e–09
	1.8	3.2e–11	0.34	4125	2.96872e–13	1.9e–09
	2.0	1.0e–12	0.27	3897	5.13470e–15	2.2e–09
$n=20,\ m=200$	1.0	1.0e–06	1.68	19668	2.31998e–08	5.3e–11
	1.2	3.2e–08	2.00	18005	3.76301e–10	4.4e–11
	1.5	1.0e–09	1.86	17577	1.40347e–11	6.8e–10
	1.8	3.2e–11	1.87	16833	2.63134e–13	9.8e–09
	2.0	1.0e–12	1.39	16384	1.00133e–14	2.5e–09
$n=30,\ m=300$	1.0	1.0e–06	4.59	45453	1.58523e–08	2.2e–11
	1.2	3.2e–08	5.91	42354	3.66854e–10	1.2e–10
	1.5	1.0e–09	4.87	40247	1.06924e–11	4.2e–10
	1.8	3.2e–11	5.31	38369	1.69296e–13	6.7e–10
	2.0	1.0e–12	3.76	37478	7.64327e–15	1.8e–09

З даної таблиці видно, що для розв'язання відповідних задач (3) за допомогою програми **emlmp** вимагається незначна кількість часу – максимальний час розв'язання задачі становить 5.91 секунди при $n=30$, $m=300$ та $p=1.2$. Результати колонки «dx» показують, що відхилення між знайденим розв'язком задачі (3) та її точкою мінімуму є дуже малим та відповідає параметру ε_f для зупинки алгоритму, вважаючи, що розвязок єдиний. Ця таблиця демонструє, що значення f_p, як правило, на декілька порядків менше, ніж ε_f. Це пов'язано з тим, що $A\cdot x_p^* = y$, $f_p^* = 0$.

Мета другого обчислювального експерименту – продемонструвати робастність методу найменших модулів, а значить така ж робастність буде характеризувати і розв'язки задачі (3), якщо p є близьким до одиниці. Тут матриця A, стартова точка x_0, радіус кулі r_0 вибиралися такими ж як і для першого тесту, вектор y коригувався так, щоб його непарні компоненти залишалися такими ж як у першому тесті, а парні компоненти домножувалися на величину **q = (1.0 + 1.0*sign(0.5 – rand)**. Таким чином, значення парних компонент вектора y можна вважати аномальними (неправильними) результатами спостережень.

Другий експеримент реалізує наведений далі Octave –код.

```
 # Test 2: Robustness of the Least Moduli Method

n = 20, m = 10*n,
rand("seed", 2023);
A = 10*rand(m,n);

xstar=round(10.0*rand(n,1)+0.5);
y = A*xstar;
x0 = round(5.0*rand(n,1)); r0 = 5*norm(x0 - xstar),
maxitn = 100000, intp = 100001,

m1 = m/2,
for i = 1:m1
  ind = (i-1)*2 + 1;
  y(ind) = y(ind)*(1.0 + 1.0*sign(0.5 - rand));
endfor

printf("\nTest 2: Robustness of the Least Moduli Method \n");
epsf0 = 1.e-6; ntest = 5; table = [];
for (in = 1:ntest)
  p = 1.d0 + (in - 1.d0)/(ntest - 1.d0),
  epsf = epsf0**(p);
  time0 = time();
  [xr,fr,itn,ist] = emlmp(A,y,p,x0,r0,epsf,maxitn,intp);
  time1 = time() - time0,
  dx = norm(xr-xstar);
  table = [table; p epsf time1 itn  ist fr dx];
endfor

n,m,
printf("   p     epsf   time   itn  ist      fr           dx  \n");
for (i = 1:ntest)
  printf(" %4.1f  %6.1e  %4.2f %6d %2d   %10.5e %10.1e\n",table(i,1:7))
endfor
```

Результати розрахунків для $n=20$ та $m=200$ наведено в табл. 2, де *dx* – норма відхилення знайденого наближення до точки мінімуму від точки **xstar**. Для всіх значень параметра p код $ist=1$, що свідчить про успішне завершення роботи програми.

ТАБЛИЦЯ 2. Другий експеримент: результати розв'язання задачі (3) при $n=20$, $m=200$ та різних p

p	ε_f	$time\,(sec)$	itn	f_p	dx	$\sqrt[p]{f_p}$
1.0	1.0e–06	1.55	18890	5.76784e+04	2.2e–10	5.76784e+04
1.2	3.2e–08	1.18	10978	2.81838e+05	1.2e+01	9.27915e+03
1.5	1.0e–09	1.19	12441	1.35286e+06	2.8e+01	1.22321e+04
1.8	3.2e–11	1.52	14321	6.50449e+06	3.4e+01	6.09703e+03
2.0	1.0e–12	1.33	16383	3.14988e+07	3.6e+01	5.61237e+03

З колонки «*dx*» даної таблиці видно, що при $p=1.0$ метод найменших модулів є робастним, адже відхилення від оптимального розв'язку становить $2.2\cdot10^{-10}$. Водночас при інших значеннях

$p>1$ відхилення є досить значним. Час розв'язання задач становить не більше 2 секунд, максимальна кількість ітерацій рівна 18890 і була виконана під час розв'язання задачі при $p=1.0$.

3. Зв'язок кардіологічних показників із психологічними індикаторами: відбір змінних. Для перевірки методу еліпсоїдів у реальних умовах та демонстрації його можливостей, використаємо задачу прогнозування психологічних висновків на основі кардіологічних даних, отриманих з використанням комплексу [1]. Для задач із медичними даними характерна особливість – високий ступінь невизначеності, тому наша мета це не отримання максимально точного прогнозу, а радше спроба визначення того, які з параметрів, що містяться в кардіологічних даних, можуть найкраще описувати психологічний стан пацієнтів. З іншої сторони можливості використання різних методів для знаходження коефіцієнтів моделей прогнозування таких як лінійна регресія, відкриває додаткові можливості для застосування подібних методів у зв'язку із різноманітністю умов щодо вхідних даних, які зустрічаються у прикладних задачах. Таким чином цікавим є застосування методу еліпсоїдів для знаходження коефіцієнтів регресії поряд із звичайним способом знаходження цих коефіцієнтів, який ґрунтується на методі найменших квадратів [10].

Задача знаходження зв'язку між медичними показниками може бути розв'язана багатьма методами, не лише побудовою прогнозуючої моделі з оцінкою відсотка поясненої невизначеності. Зокрема у [11] застосовувався непараметричний дискримінантний аналіз – тест Крускала – Уолліса – для встановлення діагностичної здатності досліджуваних у роботі діагностичних параметрів стосовно ішемічної хвороби серця у пацієнтів похилого віку.

За допомогою комплексу [1] було досліджено 90 пацієнтів, причому отримані вимірювання у нашому випадку становили 209 змінних за кардіологічними показниками до яких ми також додали вік пацієнтів: всього 210 змінних. Серед них було 175 числових змінних, включаючи вік, та 35 порядкових. Для прогнозування за допомогою регресії ми використали лише числові змінні, оскільки кількість спостережень у вибірці і так майже у 2 рази менша за кількість змінних. У подальшому також будуть відкинуті змінні із пропущеними значеннями у зв'язку з особливостями імплементації використаного методу відбору змінних. Таким чином числових змінних без пропущених значень у вибірці залишається 131 змінна.

Ми будемо розв'язувати задачу прогнозування оцінок психологічних шкал та формалізованого висновку психолога (висновок психолога групований, впорядкований та закодований числами від 1 до 4, з урахуванням впорядкованості за зміною стану) з використанням показників кардіограми та показників варіабельності серцевого ритму, отриманих за методикою описаною у [2]. Однією з проблем, яку доводиться вирішувати при застосуванні прогнозуючих методів до реальних даних є проблема відбору регресорів, або прогнозуючих ознак моделі. Далеко не всі методи спроєктовані так, щоб мати можливість відібрати найбільш підходящі ознаки у самому методі [12]. Оскільки для наших даних кількість ознак перевищує кількість спостережень, то застосовуючи прогнозування за допомогою регресії із використанням всіх наявних змінних у такому випадку, ми би зіткнулися з ефектом перенавчання, який для згаданого випадку – простий наслідок лінійної залежності системи векторів, де кількість векторів перевищує розмірність простору.

Тому для застосування знаходження коефіцієнтів регресії за допомогою методу еліпсоїдів ми застосуємо додаткову процедуру відбору ознак – Sequential Feature Selector [13], яка зокрема дозволяє використати додаткові, краще вивчені, алгоритми прогнозування для з'ясування, який набір ознак був би оптимальним для даної практичної задачі. Для обчислень використовується мова Python v3.10 [14] з бібліотеками роботи з даними NumPy v1.23 [15] та Pandas v1.5 [16], а алгоритм Sequential Feature Selector та звичайна лінійна регресія для відбору ознак береться з бібліотеки Scikit-Learn v1.2 [17]. Розрахунки проводилися у середовищі Google Colab [18].

З огляду на реалізації такого відбору ознак у даному класі в бібліотеці Scikit-Learn [19] ми будемо використовувати лише змінні без пропущених значень, оскільки алгоритм відбору не підтримує заповнення пропущених значень разом із крос-валідацією. Додатково при використанні лінійної регресії ми будемо використовувати стандартне масштабування ознак для зведення до нульового середнього та одиничної дисперсії. Для визначення оптимального набору змінних, будемо використовувати метрику R^2.

Використання для відбору ознак звичайної лінійної регресії дозволяє нам обмежити використання методу еліпсоїдів лише для знаходження коефіцієнтів для наперед визначених змінних, а отже дослідити лише цей аспект ефективності даного методу. Тим паче, що сам процес відбору змінних дозволяє використовувати різні підходи на різних етапах у процесі створення алгоритму відбору для конкретної задачі і визначеного набору даних [12].

Також, враховуючи використання на даному етапі регресії за методом найменших квадратів та метрики R^2, яка використовує вибіркові дисперсії, ми застосовуємо процедуру очистки даних від викидів із використанням методу Isolation Forest [20, 21] та його реалізації на базі Scikit-Learn [22]. Для набору із 131 числової змінної без пропусків та цільової змінної «шкала тривоги Бека» було відкинуто 8 спостережень, а отже залишилися дані по 82 пацієнтам. Для цільової змінної «формалізований висновок психолога» було відкинуто 6 спостережень і залишилися дані по 84 пацієнтам.

Застосовуючи Sequential Feature Selector та лінійну регресію по методу найменших квадратів, ми отримали додатні усереднені значення метрики R^2 на тестовій вибірці за допомогою крос–валідації по 5 вкладках для цільових змінних таких, як «шкала тривоги Бека» та «формалізований висновок психолога». Результати для отриманих середніх значень метрик на тренувальних та тестових вибірках з використанням крос-валідації на 5 вкладках наведено в табл. 3.

ТАБЛИЦЯ 3. Результати крос-валідації для вибору кількості змінних для прогнозу шкали тривоги Бека та формалізованого висновку психолога

Шкала тривоги Бека			**Формалізований висновок психолога**		
Кількість змінних	R^2 –метрика на тренувальній вибірці	R^2 –метрика на тестовій вибірці	Кількість змінних	R^2 –метрика на тренувальній вибірці	R^2 –метрика на тестовій вибірці
11	0.404470145	0.29707928	16	0.404763	0.184483144
12	0.418964541	0.303548047	17	0.431731	0.203951067
13	0.425250253	0.306413828	18	0.432256	0.207074532
14	0.425250253	0.306413828	19	0.443684	0.207097677
15	**0.425340358**	**0.306661804**	20	0.447223	0.210212644
16	0.425281984	0.306501178	**21**	**0.448175**	**0.213963134**
17	0.425272835	0.306455843	**22**	**0.448175**	**0.213963134**
18	0.42717088	0.293854387	**23**	**0.448175**	**0.213963134**
19	0.429288984	0.286335606	**24**	**0.448175**	**0.213963134**
20	0.445415707	0.261191929	**25**	**0.448175**	**0.213963134**

Отже, в результаті застосування Sequential Feature Selector отримані усереднені значення метрики R^2 для цільових змінних входять у визначення «Fair» за [23], з урахуванням того, що R^2

може розглядатися як квадрат множинного коефіцієнта кореляції між цільовою змінною та вибраним набором предикторів: $0.3 \le \sqrt{0.306}$, $\sqrt{0.213} < 0.6$.

Як найкращу кількість змінних ми візьмемо ту кількість яка максимізує метрику якості на тестовій вибірці, причому найменшу можливу таку кількість змінних. З метою запобігання перенавчання можна ще скорочувати оптимальну кількість змінних, використовуючи іншу оптимальну точку, аніж просто точку максимуму, проте ми обмежимося таким критерієм для перевірки застосування методу еліпсоїдів для обраних змінних.

Таким чином ми можемо для подальшого чисельного експерименту взяти 15 ознак для цільової змінної «шкала тривоги Бека» та 21 ознаку для цільової змінної «формалізований висновок психолога». А саме, «шкалу тривоги» Бека найкраще прогнозують кардіологічні спостереження: амплітуда Q (мкВ) (відв. II), амплітуда S (мкВ) (відв. III), амплітуда P (мкВ) (відв. III), амплітуда Q (мкВ) (відв. AvL), відношення амплітуд R/P (відв. II), амплітуда Q (мкВ) (відв. AvF), LF_n, амплітуда S (мкВ) (відв. AvF), індекс співвідношення фаз ЕКГ, стан резервів регуляції, код відведення AvR_init, комплексна оцінка виникнення суттєвих серцево-судинних подій_init, функціональний стан за Баєвським, код відведення I_univ, HF_n.

Цільову змінну «формалізований висновок психолога» найкраще прогнозують змінні: відношення амплітуд Q/R (відв. II), стан резервів регуляції, код відведення AvR_init, тривалість P (сек), поглиблений аналіз ЕКГ (6 відведень), амплітуда S (мкВ) (відв. II), активність підкоркових рівнів регуляції, тривалість Q (сек), амплітуда R (мкВ) (відв. AvF), $SDSD$, мс, $PNN50$, %, амплітуда Q (мкВ) (відв. AvF), амплітуда T (мкВ) (відв. I), підняття точки J над ізолінією (мкВ) (відв. I), відношення амплітуд R/T (відв. II), комплексна оцінка ступеню ураження міокарда_init, комплексна оцінка ступеню ураження міокарда_univ, $K_1 = (PQ + QTc)/RR$, інтегральний показник форми $ST - T$ (відв. I), комплексний показник стану міокарда_univ, кут альфа T у фронтальній площині.

Також, потрібно зауважити, що застосований алгоритм відбору ознак є жадібним [19], проте він вибирає набір ознак, що максимізують метрику якості (R^2 у нашому випадку), а тому дозволяє вибирати ознаки без додаткового попереднього відсіювання ознак на основі ступеню колінеарності, або інших критеріїв, за умови, що використаний метод прогнозування може впоратися із прогнозуванням за обраними ознаками. Таким чином ми скорочуємо обчислювальні витрати на перебір параметрів додаткових критеріїв, а також відмовляємося від ручного встановлення цих параметрів.

Тепер можемо застосувати метод еліпсоїдів для знаходження коефіцієнтів регресії для вибраних ознак відповідно до цільових змінних, щоби порівняти його результати у наближеному розв'язанні системи лінійних рівнянь, з якої шукаються коефіцієнти лінійної регресії із результатами звичайної регресії.

Додатково зазначимо, що відшукання коефіцієнтів є наближеним, оскільки системи рівнянь, які описують регресійну задачу зазвичай є перевизначеними, і додатково спотвореними:

а) за рахунок неточності у вимірюваннях. Такі задачі можна розв'язувати за допомогою моделей з похибками у змінних, які розглядаються у [24] для множинної одновимірної лінійної регресії, та, наприклад, у [25, 26] для векторної лінійної регресії;

б) за рахунок того, що регресійна модель (не обов'язково лінійна) є лише наближенням істинної невідомої залежності між факторами моделі [10];

в) та за рахунок того, що важливі для прогнозування та опису досліджуваного явища змінні можуть бути недоступні або внаслідок обмежень на дизайн експерименту, або особливостей тех-

нологій вимірювання та спостереження, або внаслідок ролі самого експерименту у пошуку оптимальних прогнозуючих змінних.

Тому, враховуючи наближеність коефіцієнтів регресійної залежності, нашою метою буде перевірка можливості застосування методу еліпсоїдів до знаходження коефіцієнтів регресії на тренувальній вибірці, із відкладенням побудови більш комплексних систем прогнозування для подальших досліджень.

4. Прогнозування психологічних індикаторів стану пацієнта на основі кардіологічних даних. Розглянемо результати обчислювального експерименту з використанням відібраних змінних з кардіологічних даних, наведених у попередньому розділі, для прогнозування цільової змінної «формалізований висновок психолога». Вхідні дані включають 84 пацієнтів з 90 (номери 7, 39, 53, 58, 59 та 87 були вилучені на етапі відбору змінних) та 22 показники. Для визначення параметрів лінійної регресійної моделі використовувалась програма **emlmp** зі значеннями параметра $p=1$ та $p=2$, де перший випадок відповідає МНМ, а другий – МНК. Матриця A_0, що включає у себе 84 спостереження для 22 показників з нумерацією у першому стовпчику, – така:

```
A0 = [
1 0.0 62 1 0.072 55 −148 4 0.026 97 9 0 0 223 −28 0.95 0.0 100 0.687 100 79 41 1;
2 0.0 41 2 0.098 80 0 5 0.0 384 22 11 0 173 −21 2.68 0.5 84 0.814 70 60 26 3;
3 0.0 48 2 0.106 86 −138 3 0.018 185 7 0 0 190 −17 4.22 0.5 84 0.891 100 64 5 3;
4 0.0 34 2 0.11 60 −104 4 0.0 0 5 0 −67 244 −22 1.53 2.2 26 0.753 45 44 9 1;
5 0.07 67 1 0.098 73 0 3 0.0 560 15 4 −55 185 −68 2.5 0.0 100 0.908 75 55 46 1;
6 0.0 67 2 0.108 65 −70 3 0.0 49 11 1 0 178 −33 1.23 1.2 60 0.748 62 54 37 2;
8 0.0 48 1 0.106 86 0 3 0.028 86 66 40 0 324 127 1.02 0.8 74 0.645 74 63 17 1;
9 0.0 81 3 0.094 84 0 3 0.0 105 22 14 0 53 10 1.07 1.2 60 0.764 38 40 67 1;
10 0.0 59 1 0.12 61 0 3 0.0 0 13 2 −65 251 118 0.9 2.2 26 0.914 60 48 5 1;
11 0.0 67 2 0.172 73 0 3 0.0 0 10 1 0 190 166 2.93 1.2 60 0.849 58 50 0 2;
12 0.0 66 1 0.098 86 0 3 0.0 509 28 24 0 279 −32 1.34 1.3 56 0.608 90 69 52 2;
13 0.0 47 2 0.076 67 0 3 0.0 611 17 1 0 61 −21 2.46 2.2 25 0.648 42 41 78 1;
14 0.0 74 1 0.118 84 0 3 0.0 0 12 2 0 143 44 3.02 1.7 43 0.752 80 52 61 1;
15 0.0 40 2 0.102 77 0 3 0.024 222 5 0 0 182 58 2.43 0.5 84 0.8 62 54 22 1;
16 0.0 77 1 0.112 70 −228 4 0.026 132 14 2 0 196 −22 1.28 0.0 100 0.826 76 51 45 3;
17 0.08 59 3 0.084 65 0 3 0.0 735 10 0 −74 97 −13 7.05 1.2 60 0.804 48 48 30 4;
18 0.0 59 1 0.062 100 0 3 0.0 522 23 1 −42 241 −36 2.95 0.0 100 0.839 90 66 19 2;
19 0.0 57 2 0.104 89 −174 3 0.0 555 10 0 0 113 −45 6.83 1.7 43 0.586 83 55 7 2;
20 0.12 66 1 0.066 81 0 3 0.0 381 17 3 −54 248 3 1.85 0.0 100 0.857 77 62 24 1;
21 0.03 63 1 0.078 89 0 3 0.0 834 20 7 −36 162 4 3.26 0.5 84 0.755 94 71 58 1;
22 0.04 71 1 0.07 100 0 3 0.026 527 20 13 0 321 43 2.43 0.8 74 0.509 80 61 22 3;
23 0.05 89 1 0.092 81 0 3 0.0 889 25 11 −70 322 62 7.11 1.2 60 0.658 91 57 −6 2;
24 0.12 45 2 0.086 50 0 3 0.022 39 4 0 0 149 5 1.46 1.2 60 0.98 88 62 41 1;
25 0.0 74 2 0.116 89 −79 3 0.0 402 13 3 0 93 −27 3.61 1.2 60 0.883 45 46 53 1;
26 0.0 69 1 0.16 78 0 3 0.0 172 15 3 0 254 66 1.72 0.5 84 0.726 85 73 39 1;
27 0.09 30 1 0.114 100 0 4 0.0 355 3 0 −34 291 30 1.78 0.0 100 1.075 60 58 21 1;
28 0.0 79 1 0.094 35 −102 3 0.0 320 15 3 −42 198 −19 1.47 0.0 100 0.717 90 63 39 1;
29 0.0 61 2 0.068 90 0 3 0.0 214 10 0 0 88 7 2.92 1.2 60 0.624 73 50 50 1;
30 0.0 29 3 0.084 89 0 3 0.0 216 9 0 0 79 17 2.84 1.2 60 0.968 57 46 48 1;
31 0.0 61 1 0.16 87 0 3 0.0 339 11 1 0 366 55 3.22 1.2 60 0.781 82 75 3 2;
32 0.0 56 2 0.1 41 −146 3 0.0 82 20 10 0 144 −28 0.64 2.0 34 0.737 70 51 52 2;
33 0.05 66 2 0.09 83 0 3 0.0 396 13 1 −25 105 −35 2.21 1.2 60 0.74 80 47 61 1;
34 0.04 65 1 0.102 80 0 3 0.0 479 12 0 0 347 19 3.07 0.0 100 0.759 71 64 16 4;
35 0.0 60 1 0.13 100 0 3 0.0 448 9 0 0 206 34 1.95 0.0 100 0.716 90 67 47 1;
36 0.0 71 3 0.106 82 0 3 0.0 94 38 7 0 93 38 1.98 1.2 60 0.737 64 50 30 2;
37 0.0 50 1 0.112 61 −169 3 0.026 98 9 1 0 258 −16 0.66 0.5 84 0.709 63 59 2 2;
38 0.0 54 2 0.094 73 −136 3 0.026 187 7 0 0 165 −1 2.75 0.5 84 1.03 73 56 40 4;
40 0.12 78 1 0.094 86 0 3 0.03 109 29 14 0 231 59 1.72 1.4 53 0.645 81 70 39 2;
```

```
41 0.0 73 2 0.086 100 0 2 0.026 325 22 10 0 101 36 4.12 0.5 84 0.707 87 59 28 1;
42 0.0 70 1 0.12 87 0 3 0.0 398 12 2 0 500 129 3.22 1.2 60 0.841 96 56 3 2;
43 0.0 71 2 0.12 79 0 3 0.0 305 18 3 0 178 6 3.59 0.5 84 0.694 81 57 19 2;
44 0.0 76 2 0.074 76 0 3 0.0 155 25 19 0 176 -24 2.25 0.5 84 0.723 70 55 30 1;
45 0.0 71 2 0.116 51 0 2 0.028 44 87 17 0 135 74 0.69 1.3 56 0.502 11 48 41 3;
46 0.0 40 2 0.124 100 0 3 0.0 518 3 0 0 122 -15 2.38 1.2 60 0.976 84 58 62 3;
47 0.0 70 1 0.12 80 0 3 0.0 27 12 1 0 255 57 0.97 0.5 84 0.612 56 51 5 3;
48 0.0 61 2 0.096 67 0 2 0.024 447 10 1 0 143 -8 7.57 0.5 84 0.978 62 50 -7 1;
49 0.0 33 2 0.108 64 -332 3 0.02 236 5 0 0 128 -11 1.9 0.5 84 0.96 49 47 53 2;
50 0.06 67 3 0.096 80 0 3 0.0 247 19 8 0 128 12 5.86 1.2 60 0.674 49 49 -1 3;
51 0.0 47 2 0.1 100 0 4 0.0 502 20 13 0 121 72 3.26 1.2 60 0.72 39 54 54 2;
52 0.09 72 2 0.098 100 0 3 0.0 581 23 2 -71 112 -20 4.64 1.2 60 0.832 95 64 45 4;
54 0.07 60 1 0.098 71 -67 3 0.0 483 9 0 -49 245 -17 2.15 0.0 100 0.874 96 72 48 1;
55 0.14 52 2 0.074 86 -50 3 0.0 708 8 1 -104 137 -6 3.53 1.2 60 1.11 55 59 57 3;
56 0.0 76 1 0.082 81 0 3 0.016 38 17 6 0 255 -8 1.48 0.0 100 0.736 46 49 11 1;
57 0.1 60 1 0.112 71 0 4 0.026 310 25 15 -34 240 46 2.08 0.0 100 0.646 67 69 36 2;
60 0.0 42 2 0.078 28 -86 5 0.0 197 29 12 0 142 -19 1.14 1.2 60 0.845 77 59 30 2;
61 0.0 54 2 0.098 72 0 4 0.0 171 8 0 0 160 65 2.1 0.5 84 0.78 53 51 30 1;
62 0.0 76 2 0.114 75 0 3 0.0 71 16 3 0 199 34 3.61 1.2 60 0.691 100 72 20 2;
63 0.0 71 2 0.086 100 0 3 0.0 565 15 3 0 151 -6 3.06 1.2 60 0.959 94 81 50 2;
64 0.0 66 2 0.066 90 0 3 0.0 274 9 0 0 163 -39 1.78 0.5 84 0.935 94 59 44 1;
65 0.0 73 2 0.096 87 0 3 0.02 128 18 8 0 185 67 1.68 0.5 84 0.76 60 55 30 2;
66 0.04 50 2 0.058 72 0 3 0.0 483 17 0 -38 201 9 15.65 1.2 60 0.697 54 48 -22 1;
67 0.0 72 2 0.118 81 0 4 0.026 0 14 2 -94 213 104 2.54 2.2 25 0.707 51 48 -10 2;
68 0.64 40 1 0.102 38 0 3 0.02 0 6 0 -75 197 -11 0.75 2.2 26 0.912 64 55 37 3;
69 0.0 51 3 0.11 100 0 4 0.0 163 8 0 0 82 24 2.05 1.2 60 0.752 76 51 49 3;
70 0.04 81 1 0.104 100 0 3 0.024 557 16 4 0 210 -1 3.85 0.0 100 0.848 65 70 35 3;
71 0.0 77 1 0.084 89 0 4 0.0 355 22 3 0 277 -37 2.09 0.5 84 0.709 90 65 14 3;
72 0.0 50 1 0.102 88 0 3 0.0 405 6 0 0 263 2 4.13 1.2 60 0.964 100 68 7 2;
73 0.0 74 1 0.094 91 0 4 0.0 342 11 2 -27 226 -17 1.9 0.0 100 0.746 94 77 37 2;
74 0.0 67 1 0.098 80 -284 3 0.0 134 19 4 0 256 18 1.72 0.5 84 0.936 94 63 22 2;
75 0.1 32 1 0.072 100 0 5 0.0 503 5 0 -85 203 39 1.93 0.5 84 0.813 60 58 53 3;
76 0.0 43 1 0.12 86 0 3 0.0 205 4 0 0 318 13 2.19 0.5 84 1.129 80 59 16 1;
77 0.0 57 1 0.082 51 -65 4 0.022 53 11 0 0 437 95 1.1 0.0 100 0.72 93 61 22 2;
78 0.0 83 2 0.108 66 0 3 0.0 141 15 3 0 165 -29 2.0 1.2 60 0.899 53 48 14 4;
79 0.0 83 2 0.134 88 0 3 0.0 169 30 17 0 86 26 1.35 2.2 25 0.741 36 44 56 2;
80 0.03 62 1 0.094 88 0 3 0.022 396 9 0 0 226 -5 3.45 0.5 84 0.715 69 67 30 3;
81 0.0 76 2 0.116 100 0 3 0.0 525 19 6 0 89 36 3.16 1.2 60 0.913 100 66 65 3;
82 0.11 65 1 0.094 100 0 3 0.0 843 11 2 -129 236 0 6.65 0.0 100 0.896 96 68 11 1;
83 0.0 74 1 0.11 86 0 3 0.024 164 14 3 0 332 8 2.55 1.2 60 0.973 95 67 5 2;
84 0.0 54 1 0.084 74 -136 3 0.024 239 11 1 0 217 2 1.86 0.0 100 0.637 85 61 43 1;
85 0.0 49 2 0.162 84 0 5 0.0 106 9 1 0 156 160 7.04 1.2 60 0.698 100 71 -5 2;
86 0.0 46 3 0.0 68 0 5 0.0 262 20 9 0 53 11 2.16 2.2 25 0.672 75 51 64 1;
88 0.0 76 2 0.088 75 0 3 0.024 224 11 1 0 160 -37 1.89 0.5 84 0.856 64 59 30 1;
89 0.0 44 1 0.092 33 -312 5 0.0 40 11 2 0 236 -43 0.96 0.5 84 0.609 78 53 15 1;
90 0.0 76 2 0.094 100 0 3 0.0 97 60 8 0 126 11 0.9 1.2 60 0.616 38 56 43 3];
```

Результати роботи програми **emlmp** наведено в табл. 4, яка містить час розв'язання задачі (рядок 3), кількість виконаних ітерацій (рядок 4), значення функції в точці x_ε^* (рядок 5) та розв'язок задачі x_ε^* (рядок 6) для чотирьох значень точності та двох значень параметра p.

З табл. 4, видно, що для розв'язання задачі при $p=1$, $\varepsilon_f = 10^{-6}$ та $p=2$, $\varepsilon_f = 10^{-12}$ програмі **emlmp** необхідно виконати близько 20 тисяч ітерацій. При збільшенні точності до 10^{-20} (при $p=1$) та 10^{-40} (при $p=2$) кількість ітерацій зростає майже вдвічі: до 35517 та 41486 ітерацій відповідно. Значення f_p для фіксованого p залишається незмінним. В останньому рядку цієї таблиці

наведено значення x_p^*, причому жирним виділено ті значення, які відрізняються при порівнянні 2 і 3 стовпчика ($p=1$) та 4 і 5 стовпчиків ($p=2$). Добре видно, що при $p=1$ таких відмінностей всього 4, а при $p=2$ всього 7, причому відповідні елементи вектора x_p^* відрізняються лише в 7 знаку після коми.

ТАБЛИЦЯ 4. Результати роботи програми **emlmp** при $n=22$, $m=84$, $p=1.0;\ 2.0$ та різних точностей ε_f

p	1		2	
ε_f	1.0e–06	1.0e–20	1.0e–12	1.0e–40
$time\,(sec)$	1.65	2.52	1.50	2.94
itn	20980	35517	19304	41486
f_p	4.14520e+01	4.14520e+01	4.00051e+01	4.00051e+01
x_p^*	4.2793599	4.2793599	**4.7386891**	**4.7386890**
	0.0260423	0.0260423	0.0214248	0.0214248
	0.5373529	**0.5373528**	0.7679384	0.7679384
	9.3570427	9.3570427	**10.1775695**	**10.1775693**
	0.0102424	0.0102424	0.0118134	0.0118134
	–0.0028067	–0.0028067	–0.0023647	–0.0023647
	0.7964202	0.7964202	0.5846114	0.5846114
	22.1611534	**22.1611538**	**20.0497737**	**20.0497739**
	0.0011158	0.0011158	0.0020131	0.0020131
	0.0372800	0.0372800	0.0324861	0.0324861
	–0.0680685	–0.0680685	–0.0606403	–0.0606403
	0.0035819	0.0035819	0.0049247	0.0049247
	0.0041393	0.0041393	0.0045200	0.0045200
	–0.0040721	–0.0040721	–0.0033057	–0.0033057
	–0.0276459	–0.0276459	–0.0751805	–0.0751805
	23.0675760	**23.0675765**	**13.8245946**	**13.8245945**
	0.6751749	0.6751749	0.4014796	0.4014796
	–0.0736514	–0.0736514	**0.3411751**	**0.3411752**
	–0.0132269	–0.0132269	–0.0095270	–0.0095270
	0.0107583	0.0107583	**0.0114221**	**0.0114220**
	0.0049044	0.0049044	–0.0027800	–0.0027800
	–74.4511848	**–74.4511864**	**–46.9434310**	**–46.9434309**

Результати прогнозування за допомогою програми **emlmp** наведено в табл. 5. У 3 і 4 стовпчиках наведено величини $y_1 = A \cdot x_1^*$ ($p=1$) та $y_2 = A \cdot x_2^*$ ($p=2$) відповідно, у стовпчиках 5 і 6 – величини $(y-y_1)/y$ та $(y-y_2)/y$ відповідно.

ТАБЛИЦЯ 5. Результати прогнозування програмою **emlmp** при $p=1$ та $p=2$ ($n=22$, $m=84$)

1	1.00000	1.76485	1.47986	0.76000	0.48000
2	3.00000	2.67279	2.66744	−0.11000	−0.11000
3	3.00000	1.70136	2.05923	−0.43000	−0.31000
4	1.00000	1.80410	1.78499	0.80000	0.78000
5	1.00000	1.14873	1.42311	0.15000	0.42000
6	2.00000	1.94310	1.88760	−0.03000	−0.06000
8	1.00000	1.00000	1.07904	−0.00000	0.08000
9	1.00000	1.92455	1.84936	0.92000	0.85000
10	1.00000	0.39202	0.56921	−0.61000	−0.43000
11	2.00000	1.34877	1.73515	−0.33000	−0.13000
12	2.00000	0.77811	1.19919	−0.61000	−0.40000
13	1.00000	1.00000	1.32359	−0.00000	0.32000
14	1.00000	1.00000	0.58369	−0.00000	−0.42000
15	1.00000	1.27186	1.56426	0.27000	0.56000
16	3.00000	2.87753	2.46477	−0.04000	−0.18000
17	4.00000	2.22912	2.73179	−0.44000	−0.32000
18	2.00000	0.96880	1.44967	−0.52000	−0.28000
19	2.00000	2.05693	2.46037	0.03000	0.23000
20	1.00000	1.00000	1.38122	−0.00000	0.38000
21	1.00000	1.51863	1.76470	0.52000	0.76000
22	3.00000	2.37171	2.56878	−0.21000	−0.14000
23	2.00000	2.02919	2.62488	0.01000	0.31000
24	1.00000	1.07479	1.31018	0.07000	0.31000
25	1.00000	2.56887	2.56732	1.57000	1.57000
26	1.00000	1.89772	1.85861	0.90000	0.86000
27	1.00000	1.32020	1.84799	**0.32000**	**0.85000**
28	1.00000	0.61832	0.61858	−0.38000	−0.38000
29	1.00000	0.98927	0.94706	−0.01000	−0.05000
30	1.00000	0.85537	1.31391	−0.14000	0.31000
31	2.00000	2.24553	2.76288	0.12000	0.38000
32	2.00000	2.00000	1.61917	−0.00000	−0.19000
33	1.00000	1.80647	1.88951	0.81000	0.89000
34	4.00000	1.65856	2.18967	−0.59000	−0.45000
35	1.00000	1.00000	1.40407	**0.00000**	**0.40000**
36	2.00000	2.42722	2.63573	0.21000	0.32000
37	2.00000	2.00000	1.97611	−0.00000	−0.01000
38	4.00000	2.09056	2.13237	−0.48000	−0.47000
40	2.00000	2.26800	2.19695	0.13000	0.10000
41	1.00000	1.00000	1.33892	−0.00000	0.34000
42	2.00000	2.00000	2.67035	−0.00000	0.34000
43	2.00000	2.04644	2.25358	0.02000	0.13000
44	1.00000	1.04672	1.06443	0.05000	0.06000
45	3.00000	3.00000	2.78058	−0.00000	−0.07000
46	3.00000	1.36578	1.91012	−0.54000	−0.36000
47	3.00000	1.48254	1.39976	−0.51000	−0.53000
48	1.00000	1.00000	1.55624	−0.00000	0.56000
49	2.00000	2.18034	2.09103	0.09000	0.05000
50	3.00000	2.04381	2.48769	−0.32000	−0.17000
51	2.00000	2.00000	2.09182	−0.00000	0.05000
52	4.00000	2.52921	2.82648	−0.37000	−0.29000
54	1.00000	1.08596	1.43601	**0.09000**	**0.44000**
55	3.00000	2.11789	2.54563	−0.29000	−0.15000

Закінчення таблиці 5

56	1.00000	1.15702	1.14151	0.16000	0.14000
57	2.00000	2.00000	1.97589	−0.00000	−0.01000
60	2.00000	1.96790	1.79679	−0.02000	−0.10000
61	1.00000	1.88433	1.78728	0.88000	0.79000
62	2.00000	1.57454	1.74542	−0.21000	−0.13000
63	2.00000	2.23528	2.73527	0.12000	0.37000
64	1.00000	1.48333	1.68271	0.48000	0.68000
65	2.00000	2.00000	2.03090	−0.00000	0.02000
66	1.00000	1.00000	1.09865	−0.00000	0.10000
67	2.00000	2.00000	2.23373	−0.00000	0.12000
68	3.00000	3.00000	3.13146	−0.00000	0.04000
69	3.00000	2.31331	2.47469	−0.23000	−0.18000
70	3.00000	2.51788	2.80616	−0.16000	−0.06000
71	3.00000	3.00000	2.90820	−0.00000	−0.03000
72	2.00000	0.83897	1.42089	−0.58000	−0.29000
73	2.00000	1.75069	1.79015	−0.12000	−0.10000
74	2.00000	2.00000	1.95943	−0.00000	−0.02000
75	3.00000	2.37770	2.07861	−0.21000	−0.31000
76	1.00000	1.00000	1.39829	−0.00000	0.40000
77	2.00000	1.57523	1.60176	−0.21000	−0.20000
78	4.00000	2.12837	2.27335	−0.47000	−0.43000
79	2.00000	1.53737	1.72094	−0.23000	−0.14000
80	3.00000	2.27581	2.25950	−0.24000	−0.25000
81	3.00000	1.95228	2.30273	−0.35000	−0.23000
82	1.00000	1.05546	1.78005	**0.06000**	**0.78000**
83	2.00000	2.22513	2.52256	0.11000	0.26000
84	1.00000	1.00000	1.01213	−0.00000	0.01000
85	2.00000	2.00000	2.08707	0.00000	0.04000
86	1.00000	1.02668	1.11495	0.03000	0.11000
88	1.00000	2.58891	2.59787	1.59000	1.60000
89	1.00000	2.54835	1.87245	**1.55000**	**0.87000**
90	3.00000	3.60511	3.28394	0.20000	0.09000

Результати табл. 5 демонструють, що при використанні критерія МНМ (п'ятий стовпчик) відсоткові значення відхилення значно менші, ніж при використання МНК (шостий стовпчик). Зокрема, у стовпчику 5 налічується 22 нульових значення (що означає, що розв'язок знайдено з установленою точністю), тоді як у стовпчику 6 нульових значень немає взагалі. Окрім цього, в стовпчику 5 міститься більше значень, які за модулем досить близькі до нуля, ніж у шостому стовпчику. Такі факти означають, що програма **emlmp** з використанням критерія найменших модулів впоралась з прогнозуванням суттєво краще, ніж при використанні критерія найменших квадратів.

5. Аналіз узгодженості масиву даних. Отже, загальний масив даних включає групу $n=90$ пацієнтів, які описуються $m=191$ діагностичними параметрами, серед яких 186 кількісних та 5 психологічних. Попередньо у 3-х параметрів (загальний рівень біоенергетики (TP), VLF, LF) були виявлені аномальні дані, які були замінені на середні арифметичні. Для подальшого відбракування параметрів, які мають значний розкид та погано узгоджуються з іншими параметрами потрібно знайти метрику, яка описує ступінь узгодженості як всього масиву даних, так і кожного пацієнта чи параметра окремо. При цьому вибракувати можна як пацієнтів, так і параметри, рівень узгодженості яких нижче певного порогу.

Для всього масиву відома рангова метрика узгодженості, така як ранговий коефіцієнт конкордації (КК) Кендала W [27]. Це число від 0 до 1, при $W < 0.3$ масив вважається неузгодженим, при $0.3 < W < 0.7$ маємо середню узгодженість, а при $W > 0.7$ – високу. Але наданий масив замість рангів включає різнорідні параметри, які мають числові значення, тому рангові метрики не підходять [28]. Потрібно врахувати також, що вибракування пацієнтів небажане, бо дані кожного пацієнта є цінними, а в першу чергу потрібно вибраковувати окремі параметри, яких, як правило, значно більше, ніж пацієнтів. Отже, цей підхід дозволить провести відбір параметрів та зменшити їх кількість.

Для коефіцієнта узгодженості (КУ) j-го параметра K_j $\left(j = \overline{1,m}\right)$ пропонуємо вираз

$$K_j = \frac{S_{\max,\, j} - S_j}{S_{\max,\, j} + S_j}, \tag{8}$$

де S_j – дисперсія за вибіркою пацієнтів для j-го параметра, $S_{\max,\, j}$ – максимальна дисперсія за вибіркою пацієнтів для j-го параметра. При повній узгодженості, коли значення певного параметра для всіх пацієнтів співпадає, тобто $S_j = 0$, він рівний 1, а при повній неузгодженості, коли параметр має максимальний розкид, тобто для половини пацієнтів він має мінімальне значення, а для іншої половини – максимальне, тоді $S_j = S_{\max,\, j}$, а КУ мінімальний та рівний 0. Шкалу ступенів узгодженості трохи модифікуємо порівняно з КК Кендала. Будемо вважати, що при $K < 0.5$ маємо низьку ступінь узгодженості, при $0.5 < K < 0.7$ середню ступінь, а при $K > 0.7$ – високу.

Дисперсію S_j для нормованого j-го параметра R_{ij} обчислюємо згідно

$$S_j = \sum_{i=1}^{n}\left(R_{ij} - 1\right)^2, \quad R_{ij} = \frac{X_{ij}}{X_{ave,\, j}}, \quad X_{ave,\, j} = \frac{1}{n}\sum_{i=1}^{n} X_{ij}, \tag{9}$$

де X_{ij} – значення j-го параметра для i-го пацієнта $\left(i = \overline{1,n}\right)$, $X_{ave,\, j}$ – середнє значення j-го параметра за вибіркою пацієнтів, а також враховано, що середнє значення нормованого j-го параметра $R_{ave,\, j} = 1$.

Відповідно, вираз для максимальної дисперсії $S_{\max,\, j}$ має вигляд

$$S_{\max,\, j} = \sum_{i=1}^{n}\left(R_{\max,\, j} - R_{ave,\max,\, j}\right)^2 = \frac{n}{4}\left(R_{\max,\, j} - R_{\min,\, j}\right)^2, \quad R_{ave,\max,\, j} = \frac{R_{\max,\, j} + R_{\min,\, j}}{2}, \tag{10}$$

де $R_{\max,\, j}$ та $R_{\min,\, j}$ – максимальне та мінімальне значення нормованого j–го параметра за вибіркою пацієнтів, $R_{ave,\max,\, j}$ – його середнє значення при максимальному розкиді.

За результатами аналізу узгодженості масиву даних згідно виразів (8), (9), (10), середній КУ дорівнює 79 %. Це свідчить про досить високу ступінь узгодженості різнорідних параметрів для даної вибірки пацієнтів загалом, бо він більше 70 %. В табл. 6 зведено КУ погано узгоджених параметрів, з якої видно, що один найгірший параметр «Тривалість Q (сек)» має низьку (< 50 %) ступінь узгодженості біля 44 %, та ще 20 параметрів демонструють середню ступінь, меншу 70 %, отже ці 21 параметрів потрібно вилучити при подальшому проведенні аналізу іншими математичними методами.

ТАБЛИЦЯ 6. Параметри, які підлягають вибракуванню

№	Назва параметра	КУ, %	№	Назва параметра	КУ, %
1	Тривалість Q (сек)	44,2	12	Комплексна оцінка ступеня ураження міокарда_init	67,2
2	Код CIIS 6_univ	58,9	13	Відношення амплітуд Q/R (відведення AvF)	67,2
3	Первинний висновок психолога	59,8	14	Код відведення I_init	67,5
4	Тривалість S (сек)	61,0	15	Код відведення I_verb	67,6
5	Код CIIS 6_verb	62,4	16	Код відведення AvR_init	68,2
6	Ознаки серцевої недостатності за даними ЕКГ_univ	63,4	17	Код відведення AvR_verb	68,2
7	Амплітуда S (мкВ) (відведення AvR)	63,8	18	Амплітуда R (мкВ) (відведення III)	69,5
8	Амплітуда S (мкВ) (відведення AvL)	64,0	19	Відношення площ P/QRS (відведення I)	69,6
9	Шкала тривоги Бека	65,2	20	Індекс співвідношення фаз ЕКГ	69,8
10	Підсумкова оцінка_verb	65,9	21	Психоемоційний стан_verb	69,8
11	Амплітуда P (мкВ) (відведення AvL)	67,1			

В контексті даного дослідження психологічні параметри № 3 (первинний висновок психолога) та № 9 (шкала тривоги Бека) не можна викидати, бо для них ставиться задача обчислення регресій. Тоді доцільно виявити, чи корелює узгодженість з максимальною точністю регресійних моделей для 4-х психологічних параметрів, отриманих раніше в розділі 3, де в якості метрики взято множинний коефіцієнт кореляції R.

Ці дані наведено в табл. 7, звідки видно, що КУ з великою точністю лінійно зв'язаний з квадратом коефіцієнта кореляції (метрикою R^2) на етапі навчання згідно виразу, де К – це КУ згідно (8). Це означає, що чим менший КУ (більша дисперсія) психологічного параметра, тим точнішу регресійну модель для нього можна побудувати. Між метрикою R^2 на етапі тестування та КУ суттєвої кореляції не виявлено.

$$R^2\,(\%) = 107.6 - 1.07 * K\,. \tag{11}$$

ТАБЛИЦЯ 7. Узгодженість та точність регресійної моделі для 4-х психологічних параметрів

№	Назва параметра	Метрика R^2		КУ
		навчання	тестування	
1	Первинний висновок психолога (4)	44,8	21,4	59,8
2	Шкала тривоги Бека (2)	42,5	30,7	65,2
3	Шкала самооцінки PCL–M–в/в (0)	27,9	2,8	71,1
4	Анкета здоров'я пацієнта PHQ–9 (1)	27,5	4,5	72,9

Висновки. У статті описано математичний інструмент на основі методу еліпсоїдів та методу найменших модулів для визначення параметрів лінійної регресійної моделі. Ця модель використовується для виявлення зв'язків у медичних кардіологічних даних та прогнозування психологічних індикаторів стану пацієнта. Реалізовано програму **emlmp** мовою Octave, яка реалізує цей алгоритм, та проведено низку обчислювальних експериментів з нею. Результати першого експерименту де-

монструють, що для знаходження розв'язку з заданою точністю при декількох сотнях спостережень та декількох десятків параметрів програмі **emlmp** необхідно не більше 6 секунд. Результати другого експерименту доводять робастність розв'язків задачі при значеннях параметра p, близьких до одиниці, а отже і методу найменших модулів.

Розглянуто приклад прогнозування психологічних індикаторів («формалізований висновок психолога», «шкала тривоги Бека») на основі даних кардіологічних обстежень. Проведено відбір ознак за допомогою процедури Sequential Feature Selector з Python-бібліотеки Scikit-learn з використанням метрики R^2. Для відібраних даних, які налічують 84 спостережень та 22 показники, проведено обчислювальний експеримент з використанням програми **emlmp** та двох значень параметра p: 1 та 2. Показано, що використання критерія на основі методу найменших модулів дозволяє отримати стійкі розв'язки за меншу кількість затрачених ітерацій та часу роботи програми як порівняти з критерієм на основі методу найменших квадратів.

Запропоновано метрику (коефіцієнт узгодженості, КУ) для оцінки узгодженості масиву даних, яка дозволяє оцінити узгодженість для кожного параметра окремо. В результаті аналізу знайдено, що із 191 параметрів 21 (див. табл. 7), які мають низьку чи середню ступінь узгодженості, повинні бути вилучені для подальшого аналізу. Також знайдено лінійний зв'язок (11) між КУ 4-х психологічних параметрів та максимальною точністю регресійних моделей при оптимальній кількості параметрів у зазначених моделях.

Використання методу еліпсоїдів є досить ефективним інструментом для задач регресійного аналізу на основі методу найменших квадратів, які розглядаються в монографії [29]. На його основі можна розробити алгоритми для визначення параметрів умовної лінійної та нелінійної регресійних моделей з обмеженнями на параметри у вигляді рівностей та нерівностей не лише для методу найменших квадратів, а й для методу найменших модулів.

Подяка. Роботу підтримано грантом Національного фонду досліджень України № 2021.01/0136, грантом Volkswagen Foundation № 97775 та проєктом дослідних робіт молодих вчених №07–02/03–2023/ВМ120.34.

Список літератури

1. Чайковський І., Прімін М., Казмірчук А. Розроблення та впровадження в медичну практику нових інформаційних технологій і метрик для аналізу малих змін в електромагнітному полі серця людини. *Вісник НАН України*. 2021. Т. **2**. С. 33–43.
2. Chaikovsky I. Electrocardiogram scoring beyond the routine analysis: subtle changes matters. *Expert Review of Medical Devices*. 2020. Vol. **17**, N 5. P. 379–382.
3. Хьюбер Дж.П. Робастность в статистике. М.: Мир, 1984. 304 с.
4. Стецюк П.И., Колесник Ю.С., Лейбович М.М. О робастности метода наименьших модулей. *Компьютерная математика*. 2002. С. 114–123.
5. Мудров В.И., Кушко В.Л. Метод наименьших модулей. М.: Издательство «Знание». 1971. 64 с.
6. Стецюк П.І., Стецюк М.Г., Брагін Д.О., Молодик М.О. Використання r-алгоритму Шора в лінійних задачах робастної оптимізації. *Cybernetics and Computer Technologies*. 2021. Т. **1**. С. 29–42. https://doi.org/10.34229/2707–451X.21.1.3
7. Шор Н.З. Метод отсечения с растяжением пространства для решения задач выпуклого программирования. *Кибернетика*. 1977. Т. **1**. С. 94–95.
8. Стовба В.О. Метод еліпсоїдів для знаходження параметрів лінійної регресії. *Cybernetics and Computer Technologies*. 2020. Т. **3**. С. 14–24. https://doi.org/10.34229/2707–451X.20.3.2
9. Fischer A., Khomyak O., Stetsyuk P. The ellipsoid method and computational aspects. *Commun. Optim. Theory*. Vol. **21**. 2023. P. 1–14. http://cot.mathres.org/archives/1548
10. James G., Witten D., Hastie T., Tibshirani R., Taylor J. An Introduction to Statistical Learning with Applications in Python. New York: Springer. 2023. 613 p.

11. Жарінова В.Ю., Табакович-Вацеба В.О., Сенько І.О. Діагностичні та прогностичні можливості кардіотропних аутоантитіл у пацієнтів похилого віку з ішемічною хворобою серця з різною скоротливою здатністю міокарда. *Український кардіологічний журнал.* 2015. Т. **4**. С. 81–86. https://ucardioj.com.ua/index.php/UJC/issue/view/29
12. Kumar V., Minz S. Feature selection: a literature review. *SmartCR*. 2014. **4** (3). P. 211–229. https://faculty.cc.gatech.edu/~hic/CS7616/Papers/Kumar–Minz–2014.pdf
13. Ferria F. J., Pudilb P., Hatefc M., Kittlerca J. Comparative Study of Techniques for Large–Scale Feature Selection. *Machine Intelligence and Pattern Recognition.* 1994. Vol. **16**. P. 403–413. https://doi.org/10.1016/B978–0–444–81892–8.50040–7
14. Rossum G.V., Drake F.L. Python 3 Reference Manual. Scotts Valley, CA: CreateSpace. 2009.
15. Harris C.R., Millman K.J., van der Walt S.J. et al. Array programming with NumPy. *Nature*. 2020. Vol. **585**. P. 357–362. https://doi.org/10.1038/s41586–020–2649–2
16. McKinney W. Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference*. 2010. Vol. **445**. P. 56–61. https://doi.org/10.25080/Majora–92bf1922–00a
17. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay, E. Scikit–learn: Machine Learning in Python. *Journal of Machine Learning Research.* 2011. Vol. **12**. P. 2825–2830. https://doi.org/10.5555/1953048.2078195
18. Google Colaboratory. https://colab.research.google.com/ (звернення: 02.10.2023)
19. Scikit–Learn: Sequential Feature Selection. https://scikit–learn.org/1.2/modules/feature_selection.html#sequential–feature–selection (звернення: 02.10.2023)
20. Liu F.T., Ting K.M., Zhou Z.-H. Isolation Forest. Eighth IEEE International Conference on Data Mining, Pisa, Italy, 2008. P. 413–422. https://doi.org/10.1109/ICDM.2008.17
21. Liu F.T., Ting K.M., Zhou Z.–H. Isolation–Based Anomaly Detection. *ACM Transactions on Knowledge Discovery from Data*. 2012. Vol. **6**, N 1. P. 1–39. https://doi.org/10.1145/2133360.2133363
22. Scikit-Learn: Isolation Forest. https://scikit–learn.org/1.2/modules/outlier_detection.html#isolation–forest (звернення: 02.10.2023)
23. Akoglu H. User's guide to correlation coefficients. *Turk J Emerg Med.* 2018. **18** (3). P. 91–93. https://doi.org/10.1016/j.tjem.2018.08.001
24. Fuller W. A. Measurement Error Models. New York: Wiley. 1987.
25. Sen'ko I.O. Consistency of an adjusted least–squares estimator in a vector linear model with measurement errors. *Ukrainian mathematical journal*. 2013. Vol. **64**, N 11. P. 1739–1751. https://doi.org/10.1007/s11253–013–0748–z
26. Sen'ko I. The asymptotic normality of an adjusted least squares estimator in a multivariate vector errors–in–variables regression model. *Theory of Probability and Mathematical Statistics.* 2014. Vol. **88**. P. 175–190. https://doi.org/10.1090/S0094–9000–2014–00929–1
27. Snedecor G.W., Cochran W.G. Statistical Methods. Eighth Edition. Iowa State University Press. 1989.
28. Грабовецький Б.Є. Економічне прогнозування і планування: навчальний посібник. Вінниця: ВДТУ. 2000. 163 с.
29. Knopov P.S., Korkhin A.S. Regression Analysis Under A Priori Parameter Restrictions. *Springer Optimization and Its Applications*. 2013. Vol. **54**. 234 p.

Одержано 12.09.2023

Стецюк Петро Іванович,
доктор фізико–математичних наук, завідувач відділу методів негладкої оптимізації
Інституту кібернетики імені В.М. Глушкова НАН України, Київ,
stetsyukp@gmail.com
https://orcid.org/0000–0003–4036–2543

Будник Микола Миколайович,
доктор технічних наук, головний науковий співробітник
відділу сенсорних пристроїв, систем та технологій безконтактної діагностики
Інституту кібернетики імені В.М. Глушкова НАН України, Київ,
budnyk@meta.ua
http://orcid.org/0000–0002–4020–0213

Сенько Іван Олександрович,
кандидат фізико–математичних наук, науковий співробітник
відділу методів негладкої оптимізації
Інституту кібернетики імені В.М. Глушкова НАН України, Київ,
statistic.roots.2013@gmail.com
https://orcid.org/0000–0002–2432–4582

Стовба Віктор Олександрович,
доктор філософії, науковий співробітник
відділу методів негладкої оптимізації
Інституту кібернетики імені В.М. Глушкова НАН України, Київ,
vik.stovba@gmail.com
https://orcid.org/0000–0003–3023–5815

Чайковський Ілля Анатолійович,
доктор медицини (Німеччина), провідний науковий співробітник
відділу сенсорних пристроїв, систем та технологій безконтактної діагностики
Інституту кібернетики імені В.М. Глушкова НАН України, Київ.
illya.chaikovsky@gmail.com

УДК 519.85, 51–76

П.І. Стецюк [*], М.М. Будник, І.О. Сенько, В.О. Стовба, І.А. Чайковський

Використання методу еліпсоїдів для вивчення зв'язків у медичних даних

Інститут кібернетики імені В.М. Глушкова НАН України, Київ
[*] *Листування: stetsyukp@gmail.com*

Погіршення морально-психологічного стану на фоні повномасштабної війни спостерігається у багатьох верств населення. Своєчасне виявлення різного роду переддепресійних станів і відповідна терапія є критично важливим завданням у теперішній час. Окрім цього, не менш важливою задачею є виявлення зв'язків між фізичними та психологічними показниками здоров'я. Встановлення таких закономірностей дозволить виявляти тривожні стани, уникаючи безпосереднього профільного тестування пацієнта.

Стаття присвячена побудові математичного апарата для прогнозування психологічних висновків на основі кардіологічних даних. Для цього використовується лінійна регресійна модель та метод еліпсоїдів для визначення її параметрів з критерієм на основі методу найменших модулів (МНМ), процедура відбору ознак та метрика для оцінки узгодженості масиву даних.

Матеріал статті викладено в 5 розділах. У розділі 1 наведено опис методу еліпсоїдів для знаходження параметрів лінійної регресії з критерієм найменших модулів у степені p. Вказано розмірності задач, які можна успішно розв'язувати за допомогою методу еліпсоїдів на сучасних комп'ютерах.

Другий розділ присвячено Octave-програмі emlmp, яка реалізує метод еліпсоїдів, та результатам двох обчислювальних експериментів з її використанням. Отримані результати демонструють робастність отриманих розв'язків при використанні значень параметра p, близьких до одиниці.

У третьому розділі описується механізм відбору змінних для найкращого прогнозування психологічного стану пацієнтів на основі кардіологічних даних. Проведено відбір змінних за допомогою Python-процедури Sequential Feature Selector для прогнозування двох психологічних індикаторів – шкали тривоги Бека та формалізованого висновку психолога.

Четвертий розділ містить результати обчислювального експерименту з використанням програми emlmp з критеріями МНМ та методу найменших квадратів (МНК) для прогнозування формалізованого висновку психолога на основі 84 відібраних пацієнтів та 22-х показників. Наведено отримані розв'язки та прогнози для порівняння критеріїв на основі МНМ та МНК.

У п'ятому розділі запропоновано метрику для оцінки узгодженості масиву даних, яка дозволяє оцінити узгодженість для кожного параметра окремо. Знайдено лінійний зв'язок між 4-ма психологічними

параметрами та максимальною точністю регресійних моделей при оптимальній кількості параметрів у зазначених моделях.

Ключові слова: лінійна регресія, опукла функція, метод еліпсоїдів, метод найменших модулів, прогнозування даних, GNU Octave.

MSC 62J05

Petro Stetsyuk [*], Mykola Budnyk, Ivan Sen'ko, Viktor Stovba, Illia Chaikovsky

Using the Ellipsoid Method to Study Relationships in Medical Data

V.M. Glushkov Institute of Cybernetics of the NAS of Ukraine, Kyiv
[*] Correspondence: stetsyukp@gmail.com

Deterioration of moral and psychological state on the background of a full–scale war is observed in many social groups. Timely detection of various types of pre–depressive states and appropriate therapy is a critically important task nowadays. In addition, an equally important task is to identify relationships between physical and psychological indicators of health. Establishing such regularities will allow detecting anxious states, avoiding direct profile testing of a patient.

The article is devoted to construction of a mathematical apparatus for predicting psychological conclusions based on cardiological data. For this, a linear regression model and the ellipsoid method are used to determine its parameters with a criterion based on least moduli method (LMM), a feature selection procedure and a metric for assessing consistency of a data set.

Material of the article is presented in 5 sections. Section 1 describes the ellipsoid method for finding parameters of linear regression with the least moduli method as a criterion in the power of p. Problem dimensions that can be successfully solved using the ellipsoid method on modern computers are indicated.

The 2nd Section is devoted to Octave program emlmp, which implements the ellipsoid method, and the results of two computational experiments with its use. The obtained results demonstrate robustness of the obtained solutions when parameter p values are close to one.

The 3rd Section describes mechanism of variable selection for the best prediction of psychological state of patients based on cardiological data. Variable selection was carried out using the Python Sequential Feature Selector procedure for predicting two psychological indicators – Beck's anxiety scale and psychologist's formalized conclusion.

The 4th Section contains the results of a computational experiment using the emlmp program with LMM and least square method (LSM) criteria for predicting a psychologist's formalized conclusion based on 84 selected patients and 22 parameters. Obtained solutions and forecasts for comparing criteria based on LMM and LSM are given.

In the 5th Section, a metric for evaluation consistency of a data set is proposed, which allows to evaluate consistency for each parameter separately. A linear connection was found between 4 psychological parameters and the maximum accuracy of regression models with optimal number of parameters in specified models.

Keywords: linear regression, convex function, ellipsoid method, least moduli method, data prediction, GNU Octave.