УДК 519.6
DOI: 10.15587/2312-8372.2019.175020

ЗНАХОДЖЕННЯ НАЙВПЛИВОВІШИХ ФАКТОРНИХ ОЗНАК ПРИ ПОБУДОВІ ЛІНІЙНИХ РЕГРЕСІЙНИХ МОДЕЛЕЙ

Гече Ф. Е., Мулеса О. Ю., Гриненко В .В., Смоланка В. Ю.

1. Вступ

Однією з найбільш актуальних проблем функціонування і розвитку підприємств є виявлення шляхів підвищення ефективності їх діяльності. Перспективними дослідженнями у цьому напрямі є визначення та аналіз чинників, які впливають на результуючі показники підприємств, через яких визначається ефективність функціонування підприємства.

Більшість результуючих економічних показників підприємства формується під впливом багатьох факторів. Виділення основних з цих факторів та встановлення зв'язків між ними дозволяє визначити основні зусилля, необхідні для їх результативного використання і, як наслідок, забезпечення ефективної діяльності підприємства.

Це актуально і для економічних задач, які можна описати за допомогою кореляційної моделі, що визначає кореляційну (регресійну) залежність між факторними ознаками і результуючими показниками.

Важливим етапом побудови кореляційної моделі є етап вибору впливових факторних ознак серед великої кількості вимірюваних показників. Від якості виконання такого вибору залежить адекватність побудованої регресійної моделі. В загальному випадку процес визначення ступеня впливу ознаки на результуючий показник є трудомістким і вимагає великих обчислювальних затрат. У зв'язку з цим, розробка нових ефективних алгоритмів знаходження найвпливовіших факторних ознак є актуальною і практично важливою задачею.

2. Об'єкт дослідження та його технологічний аудит

Об'єктом дослідження є задача побудови лінійної регресійної моделі, яка виникає в процесі вирішення проблеми прогнозування значень залежної змінної від сукупності незалежних факторних ознак.

Особливістю проблеми прогнозування економічних показників діяльності підприємств є те, що в процесі аналізу їх діяльності доводиться приймати рішення про доцільність включення в модель багатьох факторів. Розв'язання такої задачі спонукає до використання апарату кореляційного аналізу. Для застосування відомих методів та алгоритмів розв'язання задач визначення ступеня впливовості ознаки на досліджуваний показник необхідним є залучення математичного апарату визначення законів розподілу деяких параметрів моделі. Такий підхід веде до суттєвого збільшення обчислювальної складності методів та алгоритмів визначення впливовості ознаки при збільшенні кількості досліджуваних факторних ознак. Це вказує на можливість

виникнення потенційних проблем з тривалістю пошуку розв'язку задачі в умовах дослідження економічних показників діяльності реальних підприємств.

Звідси випливає потреба в розробці нових ефективних алгоритмів визначення найвпливовіших факторних ознак на результуючий показник, використання яких на практиці дозволить уникнути описаних вище труднощів. Важливим при цьому є досягнення адекватності побудованих моделей та передбачення можливості виявлення випадку, у якому жоден набір з розглядуваних ознак не може забезпечити необхідний рівень адекватності.

3. Мета та задачі дослідження

Мета роботи – дослідити задачу побудови рівняння множинної лінійної регресії та запропонувати нові підходи до знаходження найвпливовіших факторних ознак для побудови лінійних регресійних моделей. Для досягнення поставленої мети було поставлено такі завдання:

1. Дослідити залежності між результуючими і факторними ознаками в процесі побудови лінійних регресійних моделей.
2. Розробити ефективний алгоритм вибору факторних ознак, використання яких може забезпечити високий рівень адекватності побудованих регресійних моделей.
3. Виконати експериментальну верифікацію розробленого алгоритму.

4. Дослідження існуючих рішень проблеми

Задачі кількісної оцінки взаємозалежностей між економічними показниками досліджуються в багатьох наукових джерелах. Так, наприклад, [1, 2] присвячені дослідженню ефективності застосування кореляційно-регресійного аналізу в процесі вирішення проблем, пов'язаних із встановленням взаємозв'язків між факторними ознаками різних видів. Широкого застосування при цьому набуває побудова рівняння множинної лінійної регресії [3], тому саме цьому питанню присвячена велика кількість наукових робіт. Так, в [4, 5] наведено математичний апарат, який застосовується в процесі побудови лінійних регресійних моделей. Тут наведені основні поняття, означення та твердження, які є ключовими в кореляційно-регресійному аналізі. В [6] описано процедуру побудови моделі кореляційного аналізу для дослідження багатофакторних процесів і явищ. При побудові рівнянь множинної лінійної регресії, для оцінки ступеня зв'язку між факторами, в цьому дослідженні пропонується застосовувати коефіцієнти частинної кореляції вищих порядків, що робить запропонований підхід складним для застосування. В роботі [7] запропоновано генетичний алгоритм вибору факторних ознак для побудови регресійних моделей. Проте, як показано в дослідженні, такий підхід в загальному випадку не дозволяє досягнути або й покращити результати роботи класичних алгоритмів вибору факторних ознак. Розробці евристичних підходів до скорочення множини факторних ознак присвячені дослідження [8, 9]. Особливістю запропонованих підходів є те, що при їх застосуванні виникає необхідність роботи з матрицями великої розмірності, що у випадку їх поганої обумовленості дуже ускладнює процес обчислень. Дослідження [10] присвячене застосуванню апарату теорії вибору та теорії нечітких множин при виборі найвпливовіших ознак. Такий підхід

призводить до того, що основним при обчисленні рівня впливу ознаки на результуючий показник є думка експерта, на основі якої відбувається побудова нечітких множин. Таким чином, при залученні різних експертів для розв'язання однієї і тієї ж задачі, з великою імовірністю, буде отримано різні результати.

Проведене дослідження наукових джерел свідчить про те, що при аналізі результатів економічної діяльності підприємств широке застосування має кореляційно-регресійний аналіз в частині побудови рівнянь множинної лінійної регресії. Важливою проблемою при цьому є вибір найвпливовіших факторних ознак, які будуть включені в модель. Відомі алгоритми або мають велику обчислювальну складність, або залежать від умов їх застосування. Таким чином, перспективним залишається питання розробки нових підходів до вирішення зазначеної проблеми.

5. Методи дослідження

Побудова економетричних моделей з використанням математичного апарату кореляційного аналізу складається з наступних етапів:

1. Вибір незалежної змінної (фактор-аргументу).
2. Оброблення статистичної інформації.
3. Встановлення міри залежності між результуючими змінними (ознаками) і факторними змінними (показниками).
4. Визначення найпливовіших факторних ознак для результуючого показнику.
5. Побудова рівняння регресії відносно найпливовіших факторних ознак і перевірка його на адекватність.
6. Аналіз кореляційної моделі.

Вибір факторних ознак (впливових факторів) проводиться на основі логічного аналізу: із сукупності технічних, технологічних, організаційних і соціально-економічних умов функціонування підприємств.

При побудові лінійних регресійних моделей, як правило, до факторних ознак ставляться наступні вимоги:

– факторні змінні повинні мати кількісне вимірювання (валовий прибуток, чистий прибуток, матеріальні витрати, амортизація тощо);

– факторні ознаки мають бути лінійно незалежними;

– їх значення визначаються за даними поточної та оперативної звітності (квартальні, річні звіти, диспетчерські документи тощо).

Після визначення залежних змінних (результуючих показників) і факторних ознак (незалежних змінних) за їхніми статистичними даними будуються:

– відповідні варіаційні (інтервальні варіаційні) ряди;

– розраховуються числові характеристики варіаційних рядів;

– видалення певних значень результуючих та факторних ознак, які «суттєво» відрізняються від основної маси спостережень;

– перевіряється вибірка на репрезентативність.

6. Результати дослідження

6.1. Аналіз залежностей між результуючими і факторними ознаками

Математичні моделі багатьох економічних задач містять цілу групу факторних ознак. Якщо кількість факторних ознак дорівнює одиниці, то такі моделі відносяться до класу парних, а в протилежному випадку до багатовимірних кореляційних моделей. При досліджені непарних кореляційних моделей використовується математичний апарат багатовимірного кореляційного аналізу.

Припустимо, що маємо багатовимірну сукупність ознак, серед яких є одна результуюча *y* і *m* факторних ознак $x_1, x_2, ..., x_m$.

Нехай маємо вибірку об'єму *n*, тобто маємо *n* точок:

$$\left(y_i, x_{1i}, x_{2i}, ..., x_{mi}\right),\ i = 1,2,...,n,$$

в *m*+1-вимірному векторному просторі.

Між кожною парою ознак можна встановити вибірковий парний коефіцієнт кореляції. Наприклад, $\tilde{r}_{yx_j}$ – вибірковий парний коефіцієнт між результуючим показником *y* і факторною змінною $x_j\left(j \in \{1,2,...,m\}\right)$, $\tilde{r}_{x_i x_j}$ – вибірковий коефіцієнт між факторними ознаками x_i і x_j.

Зв'язок між ознаками можна задати за допомогою кореляційної матриці:

$$Q_{m+1} = Q\left(y, x_1, ..., x_m\right),$$

елементами якої є вибіркові парні коефіцієнти кореляції:

$$Q_{m+1} = \begin{pmatrix} 1 & \tilde{r}_{yx_1} & ... & \tilde{r}_{yx_m} \\ \tilde{r}_{x_1 y} & 1 & ... & \tilde{r}_{x_1 x_m} \\ ... & ... & ... & ... \\ \tilde{r}_{x_m y} & \tilde{r}_{x_m x_1} & ... & 1 \end{pmatrix}. \quad (1)$$

Для того, щоб знайти вплив тільки одної факторної ознаки x_j на результуючий показник *y*, необхідно зафіксувати значення ознак $x_1, ..., x_{j-1}, x_{j+1}, ..., x_m$ і отримати вибірку об'єму *n* при різних значеннях ознаки x_j. Зрозуміло, що в даному випадку варіація *y* пояснюється змінною x_j. Коефіцієнт кореляції, отриманий на основі цієї вибірки, називається частинним коефіцієнтом кореляції і позначається $\tilde{r}_{yx_j \cdot \{x_1,...,x_m\} \setminus \{x_j\}}$. Частинні коефіцієнти кореляції можуть бути знайдені за допомогою кореляційної матриці Q_{m+1} за наступною формулою [5]:

$$\tilde{r}_{yx_j\cdot\{x_1,\ldots,x_m\}\setminus\{x_j\}} = \frac{-A_{1\,j+1}}{\sqrt{A_{11}\cdot A_{j+1\,j+1}}}, \qquad (2)$$

де A_{sq} – алгебраїчне доповнення до елементу a_{sq} кореляційної матриці, що знаходиться на перетині s-го рядка і q-го стовпчика кореляційної матриці Q_{m+1}.

Для оцінки значущості частинних коефіцієнтів кореляції $\tilde{r}_{yx_j\cdot\{x_1,\ldots,x_m\}\setminus\{x_j\}}\ (j=1,2,\ldots,m)$ знаходяться відповідні значення параметрів t_j:

$$t_j = \frac{\tilde{r}_{yx_j\cdot\{x_1,\ldots,x_m\}\setminus\{x_j\}}\sqrt{k}}{\sqrt{1-\tilde{r}^2_{yx_j\cdot\{x_1,\ldots,x_m\}\setminus\{x_j\}}}},$$

де t_j – випадкові величини, які розподілені за законом Стьюдента зі ступенем свободи $k=n-m-1$, і порівнюються з критичним значенням $t_{k,\alpha}$ при рівні значущості α. Якщо $|t_j| > t_{k,\alpha}$, тоді частинний коефіцієнт $\tilde{r}_{yx_j\cdot\{x_1,\ldots,x_m\}\setminus\{x_j\}}$ є значущим.

Множинний кореляційний зв'язок між результуючим показником y і факторними ознаками $x_1, x_2, \ldots, x_m$ можна визначити через коефіцієнт множинної кореляції $R_{y.\,x_1,x_2,\ldots,x_m}$ [4, 5], який знаходиться за наступною формулою:

$$R_{y.\,x_1,x_2,\ldots,x_m} = \sqrt{1-\frac{|Q_{m+1}|}{A_{11}}}, \qquad (3)$$

де $|Q_{m+1}|$ – визначник (детермінант) кореляційної матриці Q_{m+1}.

Після виявлення впливових факторних ознак будується рівняння множинної лінійної регресії:

$$\begin{aligned}&\tilde{y}(x_1,x_2,\ldots,x_m) = m_y(x_1,x_2,\ldots,x_m) = \\ &= a_0 + a_1x_1 + \ldots + a_mx_m,\end{aligned} \qquad (4)$$

де $m_y(x_1,x_2,\ldots,x_m)$ – умовне математичне сподівання.

Нехай $y_i = \tilde{y}_i + \tilde{u}_i\ (i=1,2,\ldots,n)$, тоді коефіцієнт детермінації [5] визначається так:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \tilde{u}_i^{\,2}}{\sum_{i=1}^{n} (y_i - \tilde{m}_y)^2},$$

де $\tilde{m}_y$ – вибіркове математичне сподівання *y*.

Адекватність лінійної регресійної моделі (4) перевіряється за параметром:

$$F = \frac{R^2}{1 - R^2} \cdot \frac{n - m - 1}{m}.$$

Розраховане значення параметра *F* порівнюється з критичним значенням $F_{кр}(m, n-m-1)$ [5] і за критерієм Фішера досліджувана лінійна регресійна модель буде адекватною, якщо $F > F_{кр}(m, n-m-1)$.

6.2. Алгоритм визначення найвпливовіших факторних ознак при побудові множинної лінійної регресійної моделі

Для розв'язання задачі вибору найвпливовіших факторних ознак пропонується наступний алгоритм.

Крок 1. Відносно факторних ознак (незалежних змінних) $x_1, x_2, ..., x_q$ на основі емпіричної таблиці (вибірки) знаходимо вибіркові коефіцієнти кореляції:

$$\tilde{r}_{yx_i} = \frac{\tilde{\mathrm{cov}}(y, x_i)}{\tilde{\sigma}_y \cdot \tilde{\sigma}_{x_i}}, \ (i = 1, 2, ..., q),$$

де y – результуюча ознака (залежна змінна); $\tilde{\mathrm{cov}}(y, x_i)$ – вибіркова коваріація досліджуваних показників y та x_i; $\tilde{\sigma}_y, \tilde{\sigma}_{x_i}$ – вибіркові середньоквадратичні відхилення досліджуваних показників.

За допомогою t-статистики Стьюдента визначимо підмножину факторних ознак $\{x_{j_1}, x_{j_2}, \ldots, x_{j_k}\} \subset \{x_1, x_2, \ldots, x_n\}$, для яких:

$$t_{j_s} = \frac{\left|\tilde{r}_{yx_{j_s}}\right| \cdot \sqrt{n-2}}{\sqrt{1 - r^2_{yx_{j_s}}}} > t_{n-2,\alpha}, \qquad (5)$$

де *n* – кількість спостережень; $t_{n-2,\alpha}$ – критичне значення параметра t_{j_s} при рівні значущості α зі ступенем свободи $n-2$.

Виконання нерівності (5) свідчить про значущість вибіркового парного коефіцієнта кореляції $\tilde{r}_{yx_{j_s}}$.

Крок 2. На основі вибіркових даних для факторних ознак $x_{j_1}, x_{j_2},..., x_{j_k}$ і результуючого показнику y побудуємо кореляційну матрицю $Q_{k+1}\left(y, x_{j_1},..., x_{j_k}\right)$:

$$Q_{k+1}\left(y, x_{j_1},..., x_{j_k}\right)=\begin{pmatrix} 1 & \tilde{r}_{yx_{j_1}} & ... & \tilde{r}_{yx_{j_k}} \\ \tilde{r}_{x_{j_1}y} & 1 & ... & \tilde{r}_{x_{j_1}x_{j_k}} \\ ... & ... & ... & ... \\ \tilde{r}_{x_{j_k}y} & \tilde{r}_{x_{j_k}x_{j_1}} & ... & 1 \end{pmatrix}$$

і знаходимо частинні вибіркові коефіцієнти кореляції:

$$\tilde{r}_{yx_{j_s}\cdot\{x_{j1},...,x_{j_k}\}\setminus\{x_{j_s}\}}=\frac{-A_{1s+1}}{\sqrt{A_{11}\cdot A_{s+1s+1}}},\ (s=1,2,...,k).$$

За допомогою t-статистики Стьюдента визначимо підмножину $\{z_1, z_2,..., z_m\}\subset\{x_{j_1}, x_{j_2},..., x_{j_k}\}$, для яких:

$$t_{yz_i\cdot\{x_{j1},...,x_{j_k}\}\setminus\{z_i\}}=\frac{\left|\tilde{r}_{yz_i\cdot\{x_{j1},...,x_{j_k}\}\setminus\{z_i\}}\right|\sqrt{n-k-1}}{\sqrt{1-\tilde{r}^2_{yz_i\cdot\{x_{j1},...,x_{j_k}\}\setminus\{z_i\}}}}>t_{n-k-1,\alpha},\qquad(6)$$

де $t_{n-k-1,\alpha}$ – критичне значення параметра $t_{yz_i\cdot\{x_{j1},...,x_{j_k}\}\setminus\{z_i\}}$ при рівні значущості і ступенем свободи $n-k-1$.

Крок 3. Знаходимо множинний коефіцієнт детермінації:

$$R^2_{y.z_1,...,z_m}=1-\frac{\left|Q_{m+1}(y, z_1,..., z_m)\right|}{A_{11}},$$

де $\left|Q_{m+1}(y, z_1,..., z_m)\right|$ – детермінант матриці $Q_{m+1}(y, z_1,..., z_m)$.

Якщо значення коефіцієнта $R^2_{y.z_1,...,z_m}$ показує, що варіація y у достатній мірі пояснюється варіаціями факторних ознак $z_1,..., z_m$, то переходимо до кроку 4. А в протилежному випадку зробимо висновок, що існують факторні змінні, які суттєво впливають на результуючу ознаку y і вони не враховані. Тому при побудові множинної лінійної регресійної моделі треба провести додаткове дослідження, щоб виявити нові значущі факторні ознаки $x_{q+1}, x_{q+2},...$, і можуть бути включені у лінійну регресійну модель. Або збільшити кількість

спостережень для уточнення зв'язків між результуючим показником y і факторними ознаками $z_1,...,z_m$, якщо це можливо.

Крок 4. Відносно факторних ознак $z_1, z_2,..., z_m$ побудуємо рівняння множинної лінійної регресії:

$$\tilde{y} = a_0 + a_1 z_1 + ... + a_m z_m . \qquad (7)$$

Крок 5. На основі вибіркових значень факторних ознак $z_1, z_2,..., z_m$ побудуємо матрицю:

$$Z(y, z_1,..., z_m) = \begin{pmatrix} 1 & z_{11} & ... & z_{1m} \\ 1 & z_{21} & ... & z_{2m} \\ ... & ... & ... & ... \\ 1 & z_{n1} & ... & z_{nm} \end{pmatrix},$$

де $i+1$ – стовпчик матриці складається із відповідних значень факторної змінної z_i (z_{ji} – значення факторної змінної z_i при її j-ому спостереженні).

Оцінимо статистичну значущість коефіцієнтів a_i ($i = 0,1,...,m$) множинної лінійної регресії (7).

Для цього знаходимо оцінену коваріаційну матрицю [5]:

$$\tilde{\sigma}_u^2 (Z' \cdot Z)^{-1} = \begin{pmatrix} \tilde{\sigma}^2_{a_0} & & * & * \\ & \tilde{\sigma}^2_{a_1} & & * \\ * & & \ddots & \\ * & * & & \tilde{\sigma}^2_{a_m} \end{pmatrix}, \qquad (8)$$

на діагоналі якої знаходяться оцінки $\tilde{\sigma}^2_{a_0}, \tilde{\sigma}^2_{a_1},..., \tilde{\sigma}^2_{a_m}$ дисперсій параметрів $a_0, a_1,..., a_m$, $\sigma_u = \sqrt{\dfrac{\sum_{i=1}^{n} (y_i - \tilde{y}_i)^2}{n-m-1}}$; Z' – транспонована матриця матриці Z.

Обчислюючи квадратні корені цих оцінок $\tilde{\sigma}^2_{a_0}, \tilde{\sigma}^2_{a_1},..., \tilde{\sigma}^2_{a_m}$, знаходимо стандартні похибки $\tilde{\sigma}_{a_0}, \tilde{\sigma}_{a_1},..., \tilde{\sigma}_{a_m}$ коефіцієнтів $a_0, a_1, \ldots, a_m$ і через них відповідно визначимо параметри $t_{a_0}, t_{a_1},..., t_{a_m}$:

$$t_{a_0} = \frac{|a_0|}{\tilde{\sigma}_{a_0}}, t_{a_1} = \frac{|a_1|}{\tilde{\sigma}_{a_1}},..., t_{a_m} = \frac{|a_m|}{\tilde{\sigma}_{a_m}}.$$

До знайдених параметрів $t_{a_i}(i=1,2,...,m)$ застосовуємо t-статистику Стьюдента. Якщо $t_{a_i} > t_{n-m-1,\alpha}$, то коефіцієнт a_i є значущим і його залишаємо у рівнянні (7), а в протилежному випадку, вважаємо, що коефіцієнт a_i є незначущим і його видалимо з рівняння (7). Після таких перетворень рівняння (7) запишеться так:

– якщо $t_{a_0} > t_{n-m-1,\alpha}$, то:

$$\tilde{y} = a_0 + a_{j_1} z_{j_1} + ... + a_{j_h} z_{j_h}, \tag{9}$$

– або у протилежному випадку:

$$\tilde{y} = a_{j_1} z_{j_1} + ... + a_{j_h} z_{j_h}, \tag{10}$$

де $\{z_{j_1},...,z_{j_h}\} \subset \{z_1,...,z_m\}$ і переходимо до кроку 6.

Крок 6. Відносно факторних ознак $z_{j_1}, z_{j_2},..., z_{j_h}$ знаходимо множинний коефіцієнт детермінації $R^2 = R^2_{y.z_{j_1},z_{j_2},...,z_{j_h}}$:

$$R^2_{y.z_{j_1},z_{j_2},...,z_{j_h}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{\sum_{i=1}^{n}(y_i - \tilde{m}_y)^2}, \tag{11}$$

де $\tilde{m}_y$ – вибіркове математичне сподівання y.

Перевіримо рівняння (9) або (10) на значущість у цілому. Для цього знаходимо величину:

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-h-1}{h}, \tag{12}$$

і порівнюємо його з критичним значенням $F_{кр}(h, n-s-1, \alpha)$.

За F-статистикою Фішера рівняння (9) або (10) буде значущим з рівнем значущості α, якщо $F > F_{кр}(h, n-h-1, \alpha)$ і алгоритм закінчено.

Зауваження. Якщо жодна факторна ознака не задовольняє умови (5) і (6), то робимо висновок, що факторні ознаки не є значущими і відносно цих факторних ознак рівняння множинної лінійної регресії не будуємо.

6.3. Аналіз результатів роботи алгоритму вибору найвпливовіших факторних ознак

Для проведення експериментальної верифікації розробленого алгоритму вибору найвпливовіших факторних ознак було розглянуто задачу побудови множинної лінійної регресії на основі показників діяльності підприємств. Дані для задачі отримані з [11]. Було розглянуто діяльність двох підприємств на протязі 2003–2012 років. Система ознак включала в себе 17 елементів, серед яких були виробничі запаси, грошові кошти та їх еквіваленти, кредиторська заборгованість за товари, роботи та послуги та інші.

Метою дослідження було визначити вплив цих ознак на такі результуючі показники, як:

– *y1* – фондовіддача основних засобів;

– *y2* – рентабельність основних засобів;

– *y3* – рентабельність капіталу при фіксованих значеннях параметрів *ST*;

– *ST* – рівень значущості факторних ознак за критерієм Стьюдента;

– *F* – рівень адекватності рівняння множинної лінійної регресії за критерієм Фішера;

– *PM* – «поріг мультиколінеарності».

Результати роботи алгоритму при (*ST*=0,1; *F*=0,1; *PM*=0,8) наведені у табл. 1, 2.

Таблиця 1

Множинні лінійні регресійні моделі для результуючих показників *y1*, *y2*, *y3* при аналізі першого підприємства

Результуючі показники	Рівняння множинної лінійної регресії	R^2
y1	*y1*=0,807086+0,000191*x13*–0,000123*x6*	0,936022
y2	*y2*=–0,024015+0,000125*x11*+0,000656*x2*	0,934932
y3	*y3*=+4,7*e*–05*x11*+0,000253*x2*	0,90165

Таблиця 2

Множинні лінійні регресійні моделі для результуючих показників *y1*, *y2*, *y3* при аналізі другого підприємства

Результуючі показники	Рівняння множинної лінійної регресії	R^2
y1	*y1*=12,329052–0,005155*x6*	0,939498
y2	*y2*=–0,005758+0,000411*x11*	0,977696
y3	*y3*=–0,00275+0,000258*x17*	0,992574

Як видно з табл. 1, 2, всі побудовані рівняння множинної лінійної регресії результуючих показників *У1*, *У2*, *У3* є адекватними і мають високий коефіцієнт детермінації для досліджуваних підприємств.

Якщо дослідити залежність між *y2* і факторними ознаками *x1, x2,…, x17*, то бачимо, що для кожного досліджуваного підприємства достовірність побудованих рівнянь є високою і у відсотках вони відповідно становлять:

– 93,4932 % (перше підприємство);
– 97,7696 % (друге підприємство).

Для порівняння *Y2* (первинні дані) та *y2* (дані отримані за рівнянням регресії) на кожному підприємстві використовуємо їх графічне зображення (рис. 1, 2).

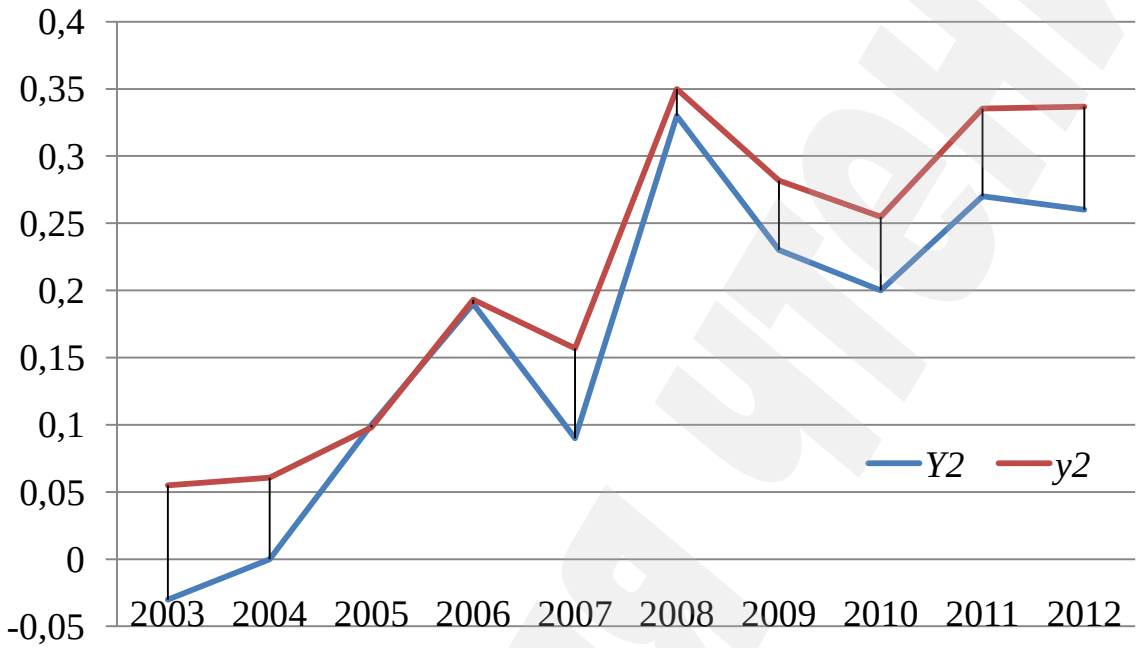

Рис. 1. Порівняння показників рентабельності основних засобів (*Y2*) та (*y2*) для першого підприємства

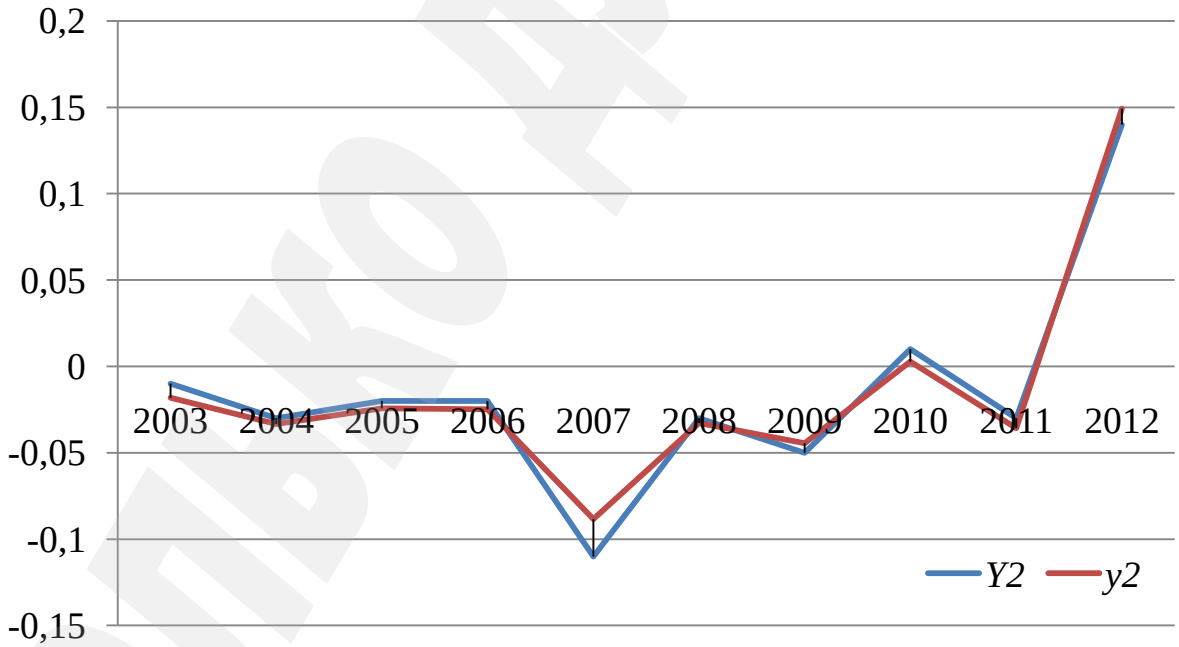

Рис. 2. Порівняння показників рентабельності основних засобів (*Y2*) та (*y2*) для другого підприємства

Отримані числові результати для модельних задач свідчать про ефективність застосування розробленого алгоритму на етапі визначення впливових факторних ознак.

7. SWOT-аналіз результатів дослідження

Strengths. Розроблений алгоритм вибору найвпливовіших факторних ознак в задачі побудови рівнянь множинної лінійної регресії базується на властивостях частинних коефіцієнтів кореляції і забезпечує якісний вибір факторних ознак. Такий підхід дозволяє зменшити обчислювальну складність процесу вибору факторних ознак в порівнянні з відомими алгоритмами.

Weaknesses. Запропонований підхід до вибору найвпливовіших факторних ознак допускає мультиколінеарність між факторними ознаками у побудованій моделі із заданим порогом.

Opportunities. Побудовані множинні лінійні регресійні моделі можуть успішно використовуватися при вирішенні проблем, пов'язаних з прогнозуванням значень економічних показників діяльності підприємств.

Threats. При розв'язуванні реальних задач вибору найвпливовіших факторних ознак можливі випадки, коли задана система ознак не може забезпечити побудову адекватної лінійної регресійної моделі. Розроблений алгоритм виявляє такі випадки, але не дає відповіді на питання щодо методики розширення системи факторних ознак з метою забезпечення можливості побудови адекватної лінійної регресійної моделі.

8. Висновки

1. Досліджено процес аналізу залежностей між результуючими і факторними ознаками, який застосовується при побудові лінійних регресійних моделей. Наведено робочі розрахункові формули для обчислення відповідних параметрів, через які визначається значущість факторних ознак, які включаються у лінійну регресійну модель.

2. Розроблено новий ефективний алгоритм вибору факторних ознак, який базується на властивостях частинних коефіцієнтів кореляції. Застосування розробленого алгоритму дозволяє зменшувати обчислювальну складність процесу вибору факторних ознак в порівнянні з відомими алгоритмами.

3. Результати роботи розробленого алгоритму продемонстровано на модельних прикладах. Показано, як з множини запропонованих факторних ознак було вибрано тільки найвпливовіші, що забезпечило досягнення достовірності побудованого рівняння множинної лінійної регресії на рівні, що перевищує 90 %.

Література

1. Smeekes, S., Wijler, E. (2018). Macroeconomic forecasting using penalized regression methods. *International Journal of Forecasting, 34 (3),* 408–430. doi: https://doi.org/10.1016/j.ijforecast.2018.01.001

2. Alvarez-Diaz, M., Alvarez, A. (2010). Forecasting exchange rates using local regression. *Applied Economics Letters, 17 (5),* 509–514. doi: https://doi.org/10.1080/13504850801987217

3. Cleland, A. C., Earle, M. D., Boag, I. F. (2007). Application of multiple linear regression to analysis of data from factory energy surveys. *International Journal of Food Science & Technology, 16 (5),* 481–492. doi: https://doi.org/10.1111/j.1365-2621.1981.tb01841.x

4. Гече, Ф. Е. (2019). *Теорія ймовірностей і математична статистика.* Ужгород: АУТДОР-ШАРК, 235.

5. Baltagi, B. (2008). *Econometric analysis of panel data.* John Wiley & Sons, 388.

6. Shojima, K., Usami, S., Hashimoto, T., Todo, N., Takano, K. (2018). Understanding Differences in Statistical Models. *The Annual Report of Educational Psychology in Japan, 57,* 302–308. doi: https://doi.org/10.5926/arepj.57.302

7. Depczynski, U., Frost, V. J., Molt, K. (2000). Genetic algorithms applied to the selection of factors in principal component regression. *Analytica Chimica Acta, 420 (2),* 217–227. doi: https://doi.org/10.1016/s0003-2670(00)00893-x

8. Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. Journal of the *Royal Statistical Society: Series B (Methodological), 58 (1),* 267–288. doi: https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

9. Mulesa, O. (2016). Development of evolutionary methods of the structural and parametric identification for tabular dependencies. *Technology audit and production reserves, 4 (2 (30)),* 13–19. doi: https://doi.org/10.15587/2312-8372.2016.74482

10. Azadeh, A., Ziaei, B., Moghaddam, M. (2012). A hybrid fuzzy regression-fuzzy cognitive map algorithm for forecasting and optimization of housing market fluctuations. *Expert Systems with Applications, 39 (1),* 298–315. doi: https://doi.org/10.1016/j.eswa.2011.07.020

11. *Агентство з розвитку інфраструктури фондового ринку України.* Available at: https://smida.gov.ua/