УДК 519.222

ВІННИК О.О., ВОЛКОВ С.В. (КІІ ДонНТУ)

УЗАГАЛЬНЕННЯ МЕТОДІВ ПОБУДОВИ КОРЕЛЯЦІЙНО-РЕГРЕСІЙНИХ МОДЕЛЕЙ

У межах єдиного підходу до кореляційно-регресійного моделювання побудована узагальнена функція регресії, а також наведена методика та розроблена програма (за допомогою інтегрованого середовища програмування DELPHI) для пофакторного встановлення виду регресії та знаходження її параметрів, що дозволяє значно узагальнити відповідне моделювання.

У багатьох задачах потрібно установити та оцінити залежність деякого показника від одного чи, як правило, кількох інших показників. Очевидно, будь-які показники, зазвичай, перебувають під впливом випадкових факторів, а тому, з математичної точки зору, інтерпретуються як випадкові (стохастичні) величини. Необхідним етапом у моделюванні процесів є якісно певна постановка питання і вибір відповідного математичного апарату, що дозволяє дати досліджуваному процесу кількісну характеристику.

З теорії ймовірностей відомо, що випадкові величини можуть бути пов’язані функціональною чи статистичною залежністю або ж узагалі бути незалежними. Для забезпечення правильного вибору математичного апарату моделювання необхідно влучно установити специфічні особливості поставленого завдання, які полягають у тому, що кількісна оцінка впливу аргументів на досліджувану величину не може бути визначена за допомогою функціонального аналізу, оскільки дія відрізняється стохастичним імовірнісним характером.

Стохастичний характер зв'язку між змінними величинами проявляється у тому, що кожному значенню однієї або декількох змінних, прийнятих за незалежні величини, відповідає не одне, строго певне значення, а декілька значень з певною вірогідністю прояву.

Нагадаємо, що статистичною називають залежність, коли зі змінюванням однієї випадкової величини змінюється закон розподілу ймовірностей іншої. Зокрема, статистична залежність виявляється в тому, що зі змінюванням однієї величини змінюється середнє значення іншої. Така залежність називається кореляційною.

Знаходження математичних параметрів і форм, що відбивають наявність і конкретне вираження стохастичного зв'язку одного показника з іншим або з групою інших досягається за допомогою кореляційного і регресійного аналізу.

Використання методів кореляції і регресії, наприклад, у планово – економічних розрахунках дозволяє на основі узагальнення даних про роботу підприємств отримати виражені у математичній формі залежності і потім розрахувати за допомогою останніх найбільш вірогідні значення, які може прийняти досліджувана величина при певних значеннях інших техніко – економічних показників, прийнятих за аргументи моделі.

У ряді випадків розробляються і використовуються моделі, що виражають стохастичну залежність однієї досліджуваної величини Y від рівня іншої X, прийнятою за незалежну змінну (парна кореляція). У загальному вигляді ця залежність має вигляд:

$$Y = M\left(Y/X\right) = f\left(X\right),$$

де $M\left(Y/X\right)-$ умовне математичне сподівання.

Не зменшуючи значення парної кореляції для виявлення наявності і тісноти зв'язку між двома змінними, необхідно ж звернути увагу на ту обставину, що через складність сучасного виробництва визначення рівня якого-небудь показника тільки на основі даних про величину одного виробничого параметра недостатньо надійно.

Так, наприклад, на рівень трудомісткості робіт і собівартості продукції на вугільних шахтах може чинити вплив значна кількість ознак, чинників матеріально-технічних умов виробництва, починаючи від потужності пласта і закінчуючи рівнем продуктивної роботи. Віддати перевагу одному із них, відкинувши усі інші, неможна оскільки на величину трудових витрат і собівартості вугілля впливає в один і той же час увесь комплекс умов виробництва. Більше того, якщо навіть припустити, що вплив одного із аргументів на величину досліджуваної функції виявиться таким, що набагато перевищує дію усіх інших аргументів так, що останні можуть бути опущені без скільки-небудь помітного зниження точності моделі то і у цьому випадку передусім необхідно виявити кількісні взаємозв'язки між показником, що вивчається, і кожним із прийнятих аргументів, а також величину сукупного впливу усіх аргументів на функцію. Інакше відкидання яких-небудь ознак чинників матеріально-технічних умов виробництва, що трохи впливають на залежну змінну буде позбавлено достатніх підстав, а отримана модель навряд чи виявиться надійною.

Багатофакторний вплив на величину залежної змінної можна записати так:

$$Y = M\left(Y/X_1, X_2, ..., X_m\right) = f\left(X_1, X_2, ..., X_m\right)$$

де $Y-$ залежна змінна (результат);

$M\left(Y/X_1, X_2, ..., X_m\right)-$ умовне математичне сподівання, результату;

$f-$ стратегія розвитку результату;

$X_k \ \left(k=\overline{1,m}\right)-$ техніко-економічні показники (інформація), фактори, які обумовлюють результат.

Зв’язок між залежною та незалежними змінними, що описуються співвідношенням:

$$Y = f\left(X_1, X_2, ..., X_m, U\right) \tag{1}$$

називається регресійним рівнянням (моделлю).

Для побудови рівняння регресії (1) необхідно пройти сім кроків, які умовно можна розділити на чотири етапи див. рисунок 1.

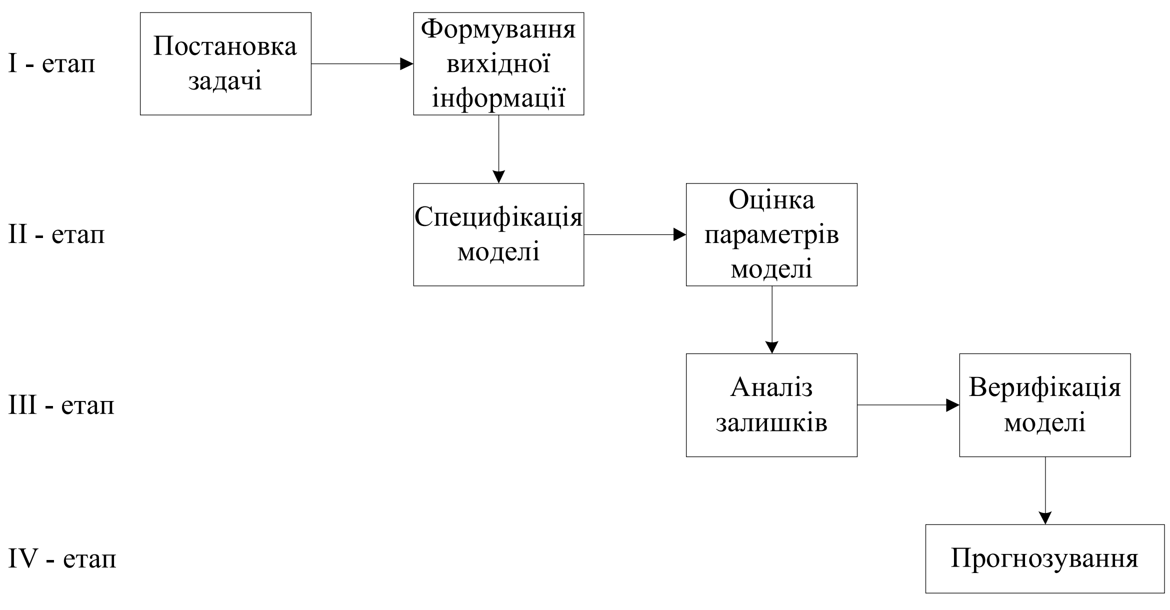

Рисунок 1. Схема побудови лінії регресії

Зазначимо, що дана робота присвячена саме вдосконаленню третього кроку побудови рівняння регресії (1).

Присутність у регресійних моделях (1) випадкового фактора U (відхилення) є обов'язковою. Найістотніші причини цього є: уведення у модель не всіх змінних; неправильний вибір функціональної форми моделі; агрегування змінних (у багатьох моделях розглядаються залежності між факторами, що самі є складною комбінацією інших, простіших змінних); помилки вимірювань; обмеженість статистичних даних; непередбачуваність людського фактора.

За звичай, для аналізу взаємозв'язків між показниками наводять клас функцій, які можуть описувати ці взаємозв'язки:

Функція	Вид	Функція	Вид
лінійна	$Y = a_0 + a_1 X_1 + \ldots + a_m X_m$	гіперболічна	$Y = a_0 + \frac{a_1}{X_1} + \frac{a_2}{X_2} + \ldots + \frac{a_m}{X_m}$
степенева	$Y = a_0 \cdot X_1^{a_1} \cdot X_2^{a_2} \ldots X_m^{a_m}$	квадратична	$Y = a_0 + a_1 X_1^2 + a_2 X_2^2 + \ldots + a_m X_m^2$
показникова	$Y = a_0 \cdot a_1^{X_1} \cdot a_2^{X_2} \ldots a_m^{X_m}$	логарифмічна	$Y = a_0 + a_1 \ln\left(X_1\right) + \ldots + a_m \ln\left(X_m\right)$

та ін..

Очевидно, наведені функції – стратегії, є однорідними відносно факторів, що в певній мірі обмежує глибокий аналіз результатів.

Автори пропонують:

По-перше: змінити класичний вигляд стратегій, функцій f. А саме розглядати суперпозицію класичних функцій. Тобто багатофакторний вплив на величину залежної змінної шукати у вигляді:

$$Y = \frac{1}{R} \sum_{k=1}^{m} r_k f_k\left(X_k\right) + U . \qquad (2)$$

де $r_k -$ коефіцієнт кореляції відповідної моделі f_k, а $R = \sum_{k=1}^{m} r_k$.

Схематично (2) можна зобразити, так:

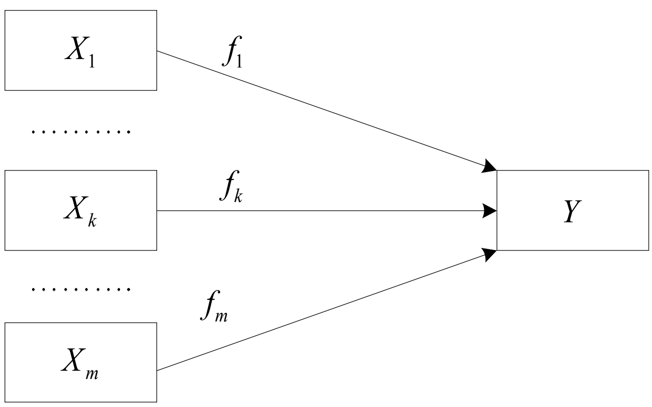

Рисунок 2. Схема моделі з пофакторним впливом

де, як правило:

$$f_k = a_0 + a_1 X_k ; f_k = a_0 X_k^{a_1} \left(a_0 X_k^2\right); f_k = a_0 a_1^{X_k} ; f_k = a_0 + \frac{a_1}{X_k};$$
$$f_k = \frac{1}{a_0 + a_1 X_k}; f_k = \frac{X_k}{a_0 + a_1 X_k}; f_k = a_0 + a_1 \ln X_k . \qquad (3)$$

По-друге: застосувати методику пофакторного визначення виду впливів інформацій X_k $\left(k=\overline{1,m}\right)$ на Y.

Встановлено, що необхідне відношення між середніми визначає вид f_k в 2. В основу цієї методики покладено відповідність середніх значень залежного та незалежного факторів, що обумовлюють наявність того чи іншого виду залежності, яка вказана у таблиці 1.

Нехай задані строго монотонні, додатні статистичні дані $Y\left(y_i\right), X_1\left(x_{i1}\right), X_2\left(x_{i2}\right),..., X_m\left(x_{im}\right)$ $\left(i=\overline{1,n}\right)$ для яких обчислені середні:

– арифметичні

$$S_a\left(Y\right)=\frac{1}{n}\sum_{i=1}^{n}Y\left(y_i\right),\quad S_a\left(X_k\right)=\frac{1}{n}\sum_{i=1}^{n}X_k\left(x_{ik}\right);$$

– геометричні

$$S_{ge}\left(Y\right)=\left(\prod_{i=1}^{n}Y\left(y_i\right)\right)^{\frac{1}{n}},\quad S_{ge}\left(X_k\right)=\left(\prod_{i=1}^{n}X_k\left(x_{ik}\right)\right)^{\frac{1}{n}};$$

– гармонічні

$$S_{gr}\left(Y\right)=\frac{n}{\sum\limits_{i=1}^{n}\frac{1}{Y\left(y_i\right)}},\quad S_{gr}\left(X_k\right)=\frac{n}{\sum\limits_{i=1}^{n}\frac{1}{X_k\left(x_{ik}\right)}}.$$

Таблиця 1. –
Необхідні умови наявності залежностей встановленого виду

№ з.п.	$S\left(X_k\right)\quad k=\overline{1,m}$	$S\left(Y\right)$	y_{sk}	$\left\lvert y_{sk}-S\left(Y\right)\right\rvert$	Вид залежності
1.	$S_a\left(X_k\right)$	$S_a\left(Y\right)$			$a_0+a_1X_k$
2.	$S_{ge}\left(X_k\right)$	$S_{ge}\left(Y\right)$			$a_0X_k^{a_1}\ \left(a_0X_k^2\right)$
3.	$S_a\left(X_k\right)$	$S_{ge}\left(Y\right)$			$a_0a_1^{X_k}$
4.	$S_{gr}\left(X_k\right)$	$S_a\left(Y\right)$			$a_0+\frac{a_1}{X_k}$
5.	$S_a\left(X_k\right)$	$S_{gr}\left(Y\right)$			$\frac{1}{a_0+a_1X_k}$
6.	$S_{gr}\left(X_k\right)$	$S_{gr}\left(Y\right)$			$\frac{X_k}{a_0+a_1X_k}$
7.	$S_{ge}\left(X_k\right)$	$S_a\left(Y\right)$			$a_0+a_1\ln X_k$

Заповнивши, враховуючи вхідні дані, перші два стовпчики необхідно $S\left(Y\right)$ порівняти з y_{sk}, що знаходиться як значення лінійної інтерполяції, яке відповідає $S\left(X_k\right)$, тобто:

$$y_{sk}=y_i+\frac{y_{i+1}-y_i}{x_{(i+1)k}-x_{ik}}\left(S\left(X_k\right)-x_{ik}\right),$$

де $x_{ik}-i-$та інформація $k-$го фактору, при чому $x_{ik}<S\left(X_k\right)<x_{(i+1)k}$;

$y_i -$ залежна змінна (результат), що відповідає $i-$тій інформації.

В результаті всіх обчислень кращою, більш підходящою, є залежність для якої $\left|y_{sk}-S(Y)\right|$ буде найменший.

Отже, для складання (2) необхідно, з урахуванням таблиця 1., за конкретною вибіркою отримати m пофакторних відповідностей

$$Y = f_k\left(X_k\right)+U_k,$$

тобто отримати такі оцінки m пар параметрів – коефіцієнтів, залежності f_k, щоб побудована регресія була найкращою у певному розумінні серед усіх інших. Іншими словами, побудована f_k має бути «найближчою» до точок спостережень за їх сукупністю.

Мірою якості знайдених оцінок можуть бути визначені композиції відхилень $U_k\left(u_{ik}\right)=Y-f_k\left(X_k\right)$. Наприклад, коефіцієнти регресії можуть бути оцінені за умови мінімізації однієї з таких сум:

$$\sum_{i=1}^{n}u_{ik}=\sum_{i=1}^{n}\left(Y-f_k\left(X_k\left(x_{ik}\right)\right)\right);\ \sum_{i=1}^{n}\left|u_{ik}\right|=\sum_{i=1}^{n}\left|Y-f_k\left(X_k\left(x_{ik}\right)\right)\right|;\ \sum_{i=1}^{n}u_{ik}^2=\sum_{i=1}^{n}\left(Y-f_k\left(X_k\left(x_{ik}\right)\right)\right)^2.$$

Метод визначення оцінок коефіцієнтів за умови мінімізації першої суми називається методом середніх (МС). Цей метод неоднозначно визначає рівняння регресії. Метод визначення оцінок коефіцієнтів за умови мінімізації другої суми називається методом найменших модулів (МНМ). Найпоширенішим і теоретично обґрунтованим є метод визначення коефіцієнтів, при якому мінімізується третя сума. Він дістав назву методу найменших квадратів (МНК). Останній метод оцінювання параметрів найпростіший з точки зору реалізації обчислень. Крім того, оцінки коефіцієнтів регресії, знайдені за МНК, при визначених передумовах, мають ряд оптимальних властивостей (незміщеність, ефективність, обґрунтованість). Серед інших методів визначення оцінок коефіцієнтів регресії можна виділити метод моментів (ММ) і метод максимальної правдоподібності (ММП).

Створена авторами програма «$freg$» реалізує побудову регресійних моделей виду (2), враховуючи аналіз пофакторного впливу незалежних змінних на результат, з оцінкою їх параметрів МНК.

Програма «$freg$» не потребує інсталяції, являється виконуючим файлом. На початку роботи необхідно запустити $freg.exe$, в результаті чого буде активоване вікно програми.

Інтерфейс програми достатньо простий у використанні. Вікно програми складається з робочої зони та меню. В верхній частині робочої зони передбачена можливість визначення кількості незалежних факторів статистики та кількості елементів кожного з них.

В меню програми «Файл» передбачено команди:

- «Відкрити статистику»: завантажує файл ***.txt чи створює таблицю для заповнення, за потребою (в програмі дозволяється аналізувати статистичні дані введені, як безпосередньо так і завантажені із файлу ***.txt.);
- «Зберегти»: зберігає статистику у файлі stat1.txt на робочому столі;
- «Вихід»: завершує роботу і закриває вікно програми (виконується команда «Зберегти» автоматично).

В меню «Друк» передбачено виведення результатів на друк. В меню «Довідка» передбачені коментарі по роботі з програмою.

В результаті відкриття, створення статистичних даних в робочій зоні з'явиться таблиця, що відображає відповідні дані і кнопка «Встановити вид залежності», після натискання якої з'явиться повідомлення про вид пофакторного зв'язку і кнопка «Розрахувати параметри». При натисканню останньої будуть обчислені параметри моделей МНК і відображено загальний вид цієї моделі (рис. 3).

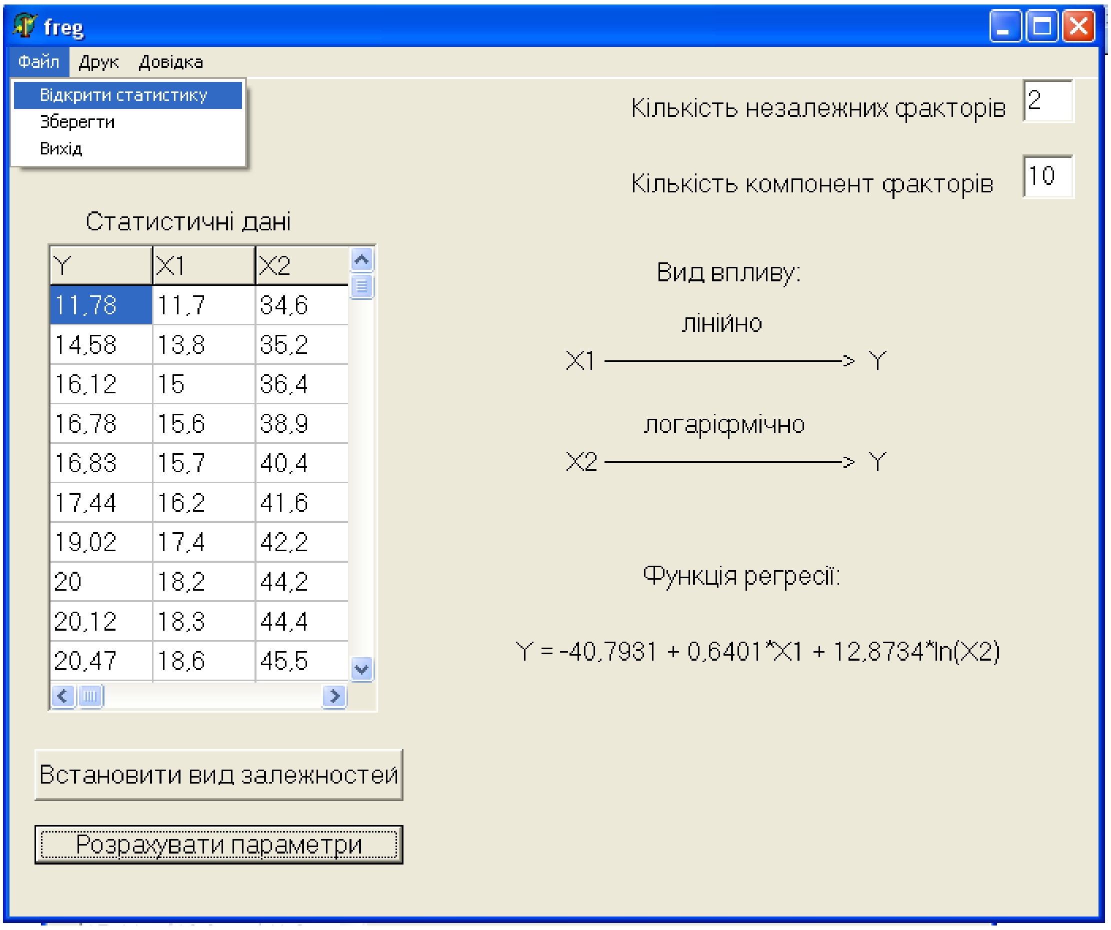

Рисунок 3. Інтерфейс програми

Автори розробленої програми «*freg*» рекомендують використовувати її на лабораторних та практичних заняттях з економічних та гірничих дисциплін таких, як «Основи геолого-математичного моделювання», «Дослідження операцій», «Економетрія», «Економіко-математичне моделювання» та ін., для отримання моделі багатофакторної регресії і використання цієї моделі для аналізу і прогнозування.

Список використаних джерел

1. Бережная Е.В., Бережной В.И. Математические методы моделирования экономических систем: Учеб. пособие. – М.: Финансы и статистика, 2001. – 368 с.
2. Демидович Б.П., Марон И.А., Шувалова Э.З. Численные методы анализа. Приближение функций, дифференциальные и интегральные уравнения. – М.: Наука, 1967. – 368 с.
3. Лещинський О. Л., Рязанцева В. В., Юнькова О.О. Економетрія: Навч. посіб. для студентів ВНЗ. – К.: МАУП, 2003. – 208 с.