УДК 631:331.4; 63:658.34; 63:331.34

В.М. Степанишин, Л.О. Тисовський
Національний університет "Львівська політехніка",
кафедра ОП
Національний лісотехнічний університет України

ПОБУДОВА МОДЕЛІ КОРЕЛЯЦІЙНОГО АНАЛІЗУ ДЛЯ ДОСЛІДЖЕННЯ БАГАТОФАКТОРНИХ ПРОЦЕСІВ І ЯВИЩ

Наведено основні показники і процедуру застосування методів множинного кореляційно-регресійного аналізу для дослідження багатофакторних процесів і явищ. Отримані результати, зокрема, будуть використані для побудови математичної моделі оптимального планування охорони праці на підприємствах лісової галузі України з метою зменшення рівня травматизму.

Ключові слова: кореляційно-регресійний аналіз, множинна регресія, коефіцієнт множинної кореляції, парні і часткові коефіцієнти кореляції, коефіцієнт еластичності, бета-коефіцієнт, дельта-коефіцієнт, довірчі інтервали для параметрів регресії.

Given the basic parameters and the procedure using the methods of multiple correlation-regression analysis to study multifactor processes and phenomena. The results, in particular, are used to construct mathematical models of optimal planning of work for the forestry sector of Ukraine in order to reduce injuries.

Key words: correlation and regression analysis, multiple regression, the coefficient of multiple correlation, paired and partial correlation coefficients, coefficient of elasticity, the beta coefficient, delta ratio, confidence intervals for regression parameters.

Постановка проблеми

На діяльність будь-якого підприємства впливають деякі фактори. Оцінити результати їх дії можливо методами статистики, основу яких становлять побудова і аналіз відповідної математичної моделі. Для багатофакторних моделей чи явищ доцільно використовувати методи множинного кореляційно-регресійного аналізу, які дають змогу вивчити та кількісно оцінити внутрішні і зовнішні наслідкові зв'язки між утворюючими модель факторами та встановити закономірності функціонування і тенденції розвитку досліджуваної результативної ознаки.

Аналіз останніх досліджень і публікацій

Методи статистичного аналізу вже давно знайшли своє застосування в різних сферах людської діяльності. І, хоча ґрунтуються вони на чистій математиці, проте в реальному житті використовуються в техніці, економіці, праві, медицині тощо [1–6]. Проте, слід зазначити, що незважаючи на велику кількість публікацій, для вирішення кожної конкретної проблеми слід будувати свою математичну модель процесу чи явища, яка би враховувала основні їх аспект і ґрунтувалася на певних засобах обчислювального апарата статистики.

Цілі досліджень

Ціллю цього дослідження є побудова розрахункової процедури багатофакторного кореля-ційно-регресійного аналізу.

Виклад основного матеріалу

У реальному житті трапляється так, що один визначальний фактор залежить від кількох різних чинників, між якими не можна встановити явного зв'язку. У цьому випадку доцільно на

основі методів багатофакторного кореляційно-регресійного аналізу створити математичну модель процесу чи явища, яка дала би змогу оцінити ступінь впливу на досліджуваний результативний показник кожного із введених у модель факторів при фіксованому положенні на середньому рівні інших факторів.

Під час побудови багатофакторних регресійних моделей можна виділити такі етапи.

Вибір всіх можливих факторів, які впливають на показник (або процес), що досліджується. Для кожного фактора потрібно визначити його числові характеристики. Якщо деякі фактори неможливо кількісно чи якісно визначити або для них недоступна статистика, то їх вилучають з подальшого розгляду.

Вибір вигляду регресійної чи багатофакторної моделі, тобто знаходження аналітичного виразу, котрий найкраще відображував би зв'язок факторних ознак з результативною, тобто вибір функції:

$$\hat{Y} = f(X_1, X_2, X_3, \ldots, X_n),$$

де $\hat{Y}$ – результативна ознака-функція; $X_1, X_2, X_3, \ldots, X_n$ – факторні ознаки.

Важливою проблемою в цьому випадку є вибір аналітичного вигляду для функції f, яка зв'язує наявні фактори з результативною ознакою-функцією. Ця функція має краще за інші відображати реальні зв'язки між досліджуваними показниками і факторами. Емпіричне обґрунтування типу функції за допомогою графічного аналізу зв'язків для багатофакторних моделей непридатне. Якщо враховувати, що будь-яку функцію багатьох змінних шляхом логарифмування або заміни змінних можна звести до лінійного вигляду, то на практиці рівняння множинної регресії подають у лінійному вигляді:

$$\hat{Y} = a_0 + a_1X_1 + a_2X_2 + \ldots + a_nX_n. , \qquad (1)$$

де $a_0, a_1, \ldots, a_n$ – параметри рівняння, які підлягають визначенню.

Якщо для кожного фактора, зокрема і результативну ознаку, відомі n значень $y_i, X_{1i}, X_{2i}, \ldots, X_{ni}$, $i=1,2,\ldots,m$, то використовуючи стандартну процедуру методу найменших квадратів, для оцінки параметрів регресійного рівняння отримаємо систему лінійних алгебраїчних рівнянь

$$\begin{cases} a_o m + a_1 \sum_{j=1}^{m} x_{1j} + a_2 \sum_{j=1}^{m} x_{2j} + \ldots + a_n \sum_{j=1}^{m} x_{nj} = \sum_{j=1}^{m} y_j; \\ a_o \sum_{j=1}^{m} x_{1j} + a_1 \sum_{j=1}^{m} x_{1j}^2 + a_2 \sum_{j=1}^{m} x_{1j}x_{2j} + \ldots + a_n \sum_{j=1}^{m} x_{1j}x_{nj} = \sum_{j=1}^{m} x_{1j}y_j; \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_o \sum_{j=1}^{m} x_{nj} + a_1 \sum_{j=1}^{m} x_{nj}x_{1j} + a_2 \sum_{j=1}^{m} x_{nj}x_{2j} + \ldots + a_n \sum_{j=1}^{m} x_{nj}^2 = \sum_{j=1}^{m} x_{nj}y_j \end{cases} \qquad (2)$$

Отриману систему n+1 рівнянь з n+1 невідомими $a_0, a_1, \ldots, a_n$ можна розв'язувати методами лінійної алгебри. Для великої кількості рівнянь найдоцільніше використовувати метод Гаусса із вибором головного елемента. Оскільки матриця цієї системи лінійних алгебраїчних рівнянь є симетричною, то її розв'язок завжди існує, причому єдиний. Якщо ж кількість рівнянь є невеликою, то з успіхом можна використовувати для розв'язання задачі метод оберненої матриці.

Перевірка адекватності отриманої моделі. Для цього необхідно обчислити:

– залишки моделі, тобто розбіжності між спостереженими та розрахунковими значеннями:

$$u_i = y_i - \hat{y}_i = y_i - (a_0 + a_1X_{1i} + a_2X_{2i} + \ldots + a_nX_{ni}), \quad i = 1,2,\ldots,m;$$

– відносну похибку залишків та її середнє значення:

$$\delta_i = \frac{u_i}{y_i} \cdot 100\%, \qquad \delta = \frac{\sum_{i=1}^{m} d_i}{m};$$

– середньоквадратичну помилку дисперсії збурень:

$$\sigma_u=\sqrt{\frac{\sum_{i=1}^{m} u_i^2}{m-n-1}};$$

– коефіцієнт детермінації:

$$R^2=1-\frac{\sum_{i=1}^{m} u_i^2}{\sum_{i=1}^{m}(y_i-\bar{y})^2} \quad \text{або} \quad R^2=1-\frac{\sum_{i=1}^{m}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{m}(y_i-\bar{y})^2};$$

– коефіцієнт множинної кореляції R , який є основним показником щільності кореляційного зв’язку узагальненого показника з факторами:

$$R=\sqrt{1-\frac{\sum_{i=1}^{m}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{m}(y_i-\bar{y})^2}}.$$

Якщо значення R є близьким до 1, то взаємозв’язок між показником і факторами вважається щільним. Множинний коефіцієнт кореляції R є основною характеристикою тісноти взаємозв’язку між результативною ознакою та сукупністю факторних ознак. Зазначимо, що про коефіцієнт кореляції йдеться тоді, коли рівняння регресії є лінійною функцією. У разі нелінійної функції регресії вводять поняття кореляційного співвідношення, яке задається таким же рівнянням, але характеризує ступінь наближення рівняння регресії до даних спостереження.

У деяких випадках під час дослідження багатофакторних процесів доцільно попередньо дослідити ступінь зв’язку між окремими факторами попарно. Якщо всі попарні зв’язки наближаються в середньому до лінійних, то є всі підстави припускати, що і множинний зв'язок буде лінійним. Для визначення щільності зв’язку між двома з досліджуваних факторів (без врахування їх взаємодії з іншими змінними) застосовуються парні коефіцієнти кореляції. Методика розрахунку цих коефіцієнтів і їх інтерпретація є аналогічними до методики розрахунку лінійного коефіцієнта кореляції для випадку однофакторного зв’язку. Якщо відомі середньоквадратичні відхилення досліджуваних величин, то парні коефіцієнти кореляції задаються співвідношеннями:

$$r_{yxi}=\frac{\overline{x_i y}-\overline{x_i}\,\overline{y}}{s_{x_i}s_y}; \qquad r_{xixj}=\frac{\overline{x_i x_j}-\overline{x_i}\,\overline{x_j}}{s_{x_i}s_{x_j}}; \qquad i,j=1,2,\ldots,n.$$

Однак в реальних умовах всі величини, як правило, взаємозв’язані. Щільність такого зв’язку визначається частковими коефіцієнтами кореляції, які характеризують ступінь і вплив одного з аргументів на функцію за умови, що решта незалежних змінних залишаються постійними. Залежно від кількості змінних, вплив яких вилучається, часткові коефіцієнти кореляції можуть бути різного порядку: при вилученні впливу одної змінної отримуємо частковий коефіцієнт кореляції першого порядку; при вилученні впливу двох змінних – другого порядку і т.д. При цьому, як правило, парний коефіцієнт кореляції між функцією і аргументом не дорівнює відповідному частковому коефіцієнту.

Часткові коефіцієнти кореляції першого порядку між ознаками x_i та у у разі вилучення впливу ознаки x_j задаються співвідношеннями:

$$r_{yxi(xj)}=\frac{r_{yxi}-r_{yxj}r_{xixj}}{\sqrt{(1-r^2_{yxj})(1-r^2_{xixj})}}; \quad i,j=1,2,\ldots,n,$$

а у разі усунення результуючої ознаки:

$$r_{xixj(y)}=\frac{r_{xixj}-r_{yxi}r_{yxj}}{\sqrt{(1-r^2_{yxi})(1-r^2_{yxj})}}.$$

Часткові коефіцієнти кореляції другого порядку визначаються через часткові коефіцієнти кореляції першого порядку:

$$r_{yxi(xj,\,xk)}=\frac{r_{yxi(xj)}-r_{yxk(xj)}r_{xixk(xj)}}{\sqrt{(1-r^2_{yxk(xj)})(1-r^2_{xixk(xj)})}};\qquad i,j,k=1,2,\ldots,n.$$

Для обчислення коефіцієнтів частинної кореляції вищих порядків використовують рекурентну формулу:

$$r_{yxi}(x_i\ldots x_{i-1}x_{i+1}\ldots x_p)=\frac{r_{yxi}(x_1\ldots x_{i-1}x_{i+1}\ldots x_{p-1})-r_{yxp}(x_1\ldots x_{i-1}x_{i+1}\ldots x_{p-1})\,r_{xixp}(x_1\ldots x_{i-1}x_{i+1}\ldots x_{p-1})}{\sqrt{(1-r^2_{yxp}(x_1\ldots x_{i-1}x_{i+1}\ldots x_{p-1})(1-r^2_{xixp}(x_1\ldots x_{i-1}x_{i+1}\ldots x_{p-1})}},$$

де $r_{yxi}(x_i\ldots x_{i-1}x_{i+1}\ldots x_p)$ – коефіцієнти частинної кореляції (p-1) порядку (в дужках не міститься фактор x_i); $r_{yxi}(x_i\ldots x_{i-1}x_{i+1}\ldots x_{p-1})$, $r_{yxp}(x_i\ldots x_{i-1}x_{i+1}\ldots x_{p-1})$, $r_{xixp}(x_i\ldots x_{i-1}x_{i+1}\ldots x_p)$ – коефіцієнти частинної кореляції (p-2) порядку (в дужках не міститься фактор $x_{i)}$.

Перевірка статистичної значущості отриманих результатів.

– перевірка адекватності моделі загалом: перевіряємо початкову гіпотезу H_o: всі коефіцієнти рівняння множинної регресії (1) дорівнюють нулю:

$$a_i=0\qquad (i=1,2,\ldots,n)$$

проти альтернативної H_i існує хоча би один коефіцієнт a_i, відмінний від нуля. Перевірка здійснюється за допомогою статистики Фішера з n та (m-n-1) ступенями вільності:

$$F=\frac{\dfrac{\sum_{i=1}^{m}(\hat{y}_i-\bar{y})^2}{n}}{\dfrac{\sum_{i=1}^{m}(y_i-\hat{y}_i)^2}{m-n-1}}\qquad \text{або}\qquad F=\frac{R^2}{1-R^2}\,\frac{m-n-1}{n},$$

де n – кількість факторів,що увійшли в модель; m – загальна кількість спостережень; $\hat{y}_i$ – розрахункове значення залежної змінної при *i*-му спостереженні; y – середнє значення залежної змінної; y_i – значення залежної змінної при *i*-му спостереженні; R – коефіцієнт множинної кореляції.

За таблицями Фішера знаходиться критичне значення $F_{кр}$ з n та (m-n-1) ступенями вільності, задавши попередньо рівень довіри (1- α)100 %. Якщо $F>F_{кр}$, то це свідчить про адекватність побудованої моделі. Якщо модель не адекватна, то необхідно повернутися до етапу побудови моделі і, можливо, ввести додаткові фактори або перейти до нелінійної моделі.

– перевірка значущості коефіцієнтів рівняння регресії.

Для цього слід перевірити гіпотезу H_o: коефіцієнт $a_i=0$ проти альтернативної H_i: $a_i\neq 0$ для кожного коефіцієнта рівняння множинної регресії (1). Перевірка здійснюється за допомогою t-статистики, яка для параметрів багатофакторної регресії має вигляд:

$$t_i=\frac{a_i}{s^2_{ai}}\qquad \text{або}\qquad t_i=\frac{a_i}{\sqrt{s^2_{ai}c_{ii}}},$$

де s_{ai} – середньоквадратичне відхилення оцінки і-го параметра; c_{ii} – діагональні елементи матриці системи рівнянь (2).

Якщо значення t_i перевищує критичне значення, яке знаходиться за таблицями t-критерію Стьюдента, то відповідний параметр є статистично значимим і має істотний вплив на узагальнюючий показник.

– перевірка значущості коефіцієнта множинної кореляції R.

Перевіряється виконання нульової гіпотези H_o: R=0 за допомогою t-статистики:

$$t=\frac{R\sqrt{m-n-1}}{\sqrt{1-R^2}}.$$

Розрахункове значення статистики порівнюється з табличним $t_{табл}$ ($\alpha/2$; m-n-1), де α – вибраний рівень значущості, m-n-1 – число ступенів вільності. Якщо $|t|>t_{табл}$, то можна зробити висновок про достовірність коефіцієнта кореляції.

Для вибраного рівня значущості α і відповідного ступеня вільності k=m-n-1, інтервал надійності для множинного коефіцієнта кореляції має вигляд:

$$(R-\Delta R,\ R+\Delta R), \quad \text{де } \Delta R=t_{\alpha/2,k}\cdot\frac{1-R}{\sqrt{m}}.$$

Обчислення та інтерпретація параметрів регресійної залежності.

Знаючи рівняння регресії не можна встановити, який з факторів найбільше впливає на результативну ознаку, оскільки здебільшого коефіцієнти рівняння регресії мають різні розмірності, а тому є непорівняльними. На їх основі також не можна встановити, яка з факторних ознак має найбільші резерви для зміни результативного показника, тому що в коефіцієнтах регресії не враховано варіацію факторної ознаки.

З метою виявлення порівняльного зв'язку і впливу окремих факторів та тих резервів, що в них закладені, обчислюють часткові коефіцієнти еластичності, а також бета-коефіцієнти і дельта-коефіцієнти.

Відмінності в одиницях вимірювання факторів усувають використанням часткових коефіцієнтів еластичності, що задаються співвідношенням:

$$\varepsilon_i=\frac{\partial\hat{y}}{\partial x_i}\frac{\overline{x_i}}{\overline{y}}$$

або для лінійного рівняння множинної регресії $\hat{y}=\sum a_i x_i$

$$\varepsilon_i=a_i\frac{\overline{x_i}}{\overline{y}},$$

де a_i – коефіцієнт регресії при факторі x_i; x_i – середнє значення і-го параметра; y – середнє значення результативної ознаки.

Частковий коефіцієнт еластичності ε_i вказує, на скільки відсотків в середньому змінюється результативна ознака із зміною на 1% фактора x_i при фіксованому значенні інших параметрів.

Бета-коефіцієнт (стандартизований коефіцієнт регресії) використовується для визначення факторів, які мають найбільший резерв для покращення результативної ознаки з врахуванням відмінностей ступеня варіації факторів, закладених у рівняння множинної регресії. Він обчислюється за формулою

$$\beta_i=a_i\frac{s_{xi}}{s_y},$$

де σ_{xi} – середнє квадратичне відхилення *i*-го параметра; σ_y – середнє квадратичне відхилення результуючої ознаки; β – коефіцієнт показує, на яку частину середньоквадратичного відхилення змінюється результативна ознака при зміні відповідної факторної ознаки на значення її середньоквадратичного відхилення.

Дельта-коефіцієнт показує, яка частина вкладу досліджуваного фактору в сумарний вплив всіх відібраних факторів. Він задається співвідношенням:

$$\Delta_i=\frac{b_i r_i}{R^2},$$

де, $r_i = r_{yxi}$ – відповідний парний коефіцієнт кореляції; R^2 – коефіцієнт множинної детермінації.

Зазначимо, що збільшення кількості факторів, які вимагаються в модель множинної регресії, дозволяє встановити додаткові ресурси результуючої ознаки.

Визначення довірчих інтервалів для параметрів регресії.

Довірчий інтервал при рівні надійності (1-α) є інтервал з випадково визначеними межами, що з рівнем довіри (1-α) накриває істинне значення коефіцієнта рівняння регресії a_i і задається залежностями:

$$(a_i - t_{\alpha/2,k}\sigma^2_{ai};\ a_i + t_{\alpha/2,k}\sigma^2_{ai}),$$

де $t_{\alpha/2,k}$ – статистика Стьюдента з к = m – n – 1 ступенями свободи і рівні значущості α; σ^2_{ai} – середньоквадратичне відхилення оцінки параметра a_i.

За допомогою наведеної вище процедури багатофакторного регресійного аналізу можна здійснювати прогнозування зміни змодельованого процесу (явища) в результаті зміни одного чи більше його факторів.

Висновки

Наведена вище розрахункова процедура множинного кореляційно-регресійного аналізу дає змогу оцінити вплив кожного із чинників, що утворюють модель процесу чи явища, на результативну ознаку і спрогнозувати поведінку об'єкта на майбутнє. Надалі на цій основі, використовуючи результати робіт [7–8], планується дослідити вплив окремих чинників на коефіцієнти частоти травматизму працівників лісової галузі України та визначити рівень значимості кожного фактора. Отримані результати стануть базовими для розроблення методики оптимального планування охорони праці з метою зменшення до мінімуму професійних ризиків.

1. Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научних работников. – М.: ФИЗМАТ ЛИТ, 2006. – 816 с. 2. Елисеева И.И., Юзбашев М.М. Общая теорія статистики: Учебник. – М.: Финанси и статистика, 2004. – 656 с. 3. Андерсон Т. Введение в многомерний статистический аналіз. – М.: Физматиз, 1963. – 500 с. 4. Лук'яненко І.Г., Краснікова Л.І. Економетрика: Підручник. – К.: Товариство «Знання», КОО, 1998. – 494 с. 5. Социально-экономическая статистика: Учеб. для вузов / Под ред. проф. Банекатова. – М.: ЮНИТИ ДАНА. – 2002. – 703 с. 6. Лапач С.Н., Чубенко А.В., Бабич П.Н. Статистические методи в медико-биологических исследованиях с использованием Excel. – МОРИОН, 2001. – 408 с.7. Гогіташвілі Г.Г., Степанишин В.М., Тисовський Л.О. Аналіз статистичних даних щодо причин та наслідків виробничого травматизму працівників // Вісник Нац. ун-ту «Львівська політехніка». – 2011. – № 707. – С. 42–45. 8. Тисовський Л.О., Степанишин В.М. Регресійний аналіз причин виробничого травматизму працівників Держкомлісгоспу України. Лісове господарство, лісова, паперова і деревообробна промисловість: Міжвідом. наук.-техн. зб. – Львів: НЛТУ України, 2011. – Вип. 37.2. – С. 34–38.