Київський національний університет технологій та дизайну

КРАСНИТСЬКИЙ С.М., МОСКОВЧУК.М.В.

ДОСЛІДЖЕННЯ ТА РОЗРОБКА МАТЕМАТИЧНОГО ЗАБЕЗПЕЧЕННЯ ДЛЯ ВВЕДЕННЯ ДОДАТКОВИХ ЗМІННИХ В ЛІНІЙНІ РЕГРЕСІЙНІ МОДЕЛІ

KRASNITSKIJ S.M., MOSKOVCHUK.M.V.

RESEARCH AND DEVELOPMENT OF MATHEMATICAL SOFTWARE FOR THE INTRODUCTION OF ADDITIONAL VARIABLES IN THE LINEAR REGRESSION MODEL

In developing mathematical models of physical phenomena and processes quite common situation where it is necessary to specify a set of independent variables that entered the model at baseline. In practice, this specification is often reduced to the solution of the question of whether or unreasonableness of the introduction of the model of a new group of additional variables. In the case of linear scalar models this situation is as follows. Let the initial assumption about the relationship between the dependent variable y and a set of independent variables x_0, x_1, …, x_{p-1} *is reduced to postulating dependence* $y = \beta_0 x_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} + \varepsilon$. *This step conjugate with increasing dimension matrix experiment that may well affect the accuracy of the estimates of coefficients. Another possibility is to use the algorithm gradual introduction of additional variables. In applying the algorithm specified dimension matrix with which we deal, does not increase if the number of additional variables that are added at each step is not greater than the previous number of independent variables. Moreover, it is possible to examine every step required to enter these variables. Standard statistical ensuring the availability of said algorithm is usually not expected. Therefore, development of appropriate computer software is quite urgent task.*

Вступ

При розробці математичних моделей реальних явищ і процесів досить часто зустрічається ситуація, коли треба уточнювати набір незалежних змінних, які введено в модель на початку дослідження.

На практиці таке уточнення часто зводиться до розв'язання питання про доцільність або недоцільність введення в модель нової групи додаткових змінних. У випадку лінійних скалярних моделей такий стан речей виглядає наступним чином.

Нехай початкове припущення про зв'язок між залежною змінною y і набором незалежних змінних x_0, x_1, …, x_{p-1} зводиться до постулювання залежності $y = \beta_0 x_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} + \varepsilon$,

Такий крок спряжений із збільшенням розмірності матриці експерименту, що цілком може вплинути на точність знаходження оцінок коефіцієнтів. Інша можливість полягає у застосуванні алгоритму поступового введення додаткових змінних. При застосуванні вказаного алгоритму розмірність матриць, з якими доводиться мати справу, не збільшується, якщо кількість додаткових змінних, що додаються на кожному кроці, не є більшою, ніж кількість попередніх незалежних змінних. До того ж, з'являється можливість на кожному кроці досліджувати потребу у введенні зазначених змінних. У стандартному статистичному забезпеченні наявність згаданого алгоритму, як правило, не передбачається. Тому розробка відповідного комп'ютерного програмного забезпечення є цілком актуальною задачею.

Постановка завдання

Створити спеціальне програмне забезпечення для проведення практичних та лабораторних занять з курсів «Ймовірнісні процеси і математична статистика», «Прикладна математика», «Прикладна антропологія», «Інтелектуальний аналіз даних», а також для використання в наукових підрозділах для дослідження таких проблем легкої промисловості як виявлення і оцінка степені впливу різних факторів на результати технологічного процесу.

Основна частина

Програмний продукт розроблений для роботи під керуванням операційної системи WINDOWS XP.

Розглянемо загальну лінійну модель регресійного аналізу: $Y = X\beta + E,$ де $Y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}$ — вектор спостережень (вектор залежних змінних),

$X = \begin{pmatrix} x_{10} x_{11} \dots x_{1,p-1} \\ x_{20} x_{21} \dots x_{2,p} \ 1 \\ \dots \ \dots \ \dots \ \dots \\ x_{n0} x_{n1} \dots x_{n,p-1} \end{pmatrix}$ — матриця експерименту (інакше, матриця плану, матриця незалежних змінних, матриця регресорів), $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_{p-1} \end{pmatrix} -$ **вектор невідомих коефіцієнтів,** $E = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{pmatrix}$ — вектор помилок.

Інколи виникає ситуація, коли модель підібрано і коефіцієнти оцінено, але виникли додаткові причини включити в модель нові незалежні змінні (регресори) $x_j, j = p, \dots, p+q.$

Припустимо, що вже після того, як підібрана модель регресії $MY = X\beta,\ DY = \sigma^2 I_n$, ми хочемо включити в неї додаткові регресори x_j, щоб модель з введенням цих факторів прийняла вигляд $G: MY = X\beta + Z\gamma = (X, Z)\begin{pmatrix} \beta \\ \gamma \end{pmatrix} = W\delta$. Тут позначено

$(X,Z) - W, \gamma - \begin{pmatrix} \beta_p \\ \beta_{p+1} \\ \dots \\ \beta_q \end{pmatrix}$ і $\begin{pmatrix} \beta \\ \gamma \end{pmatrix} = \delta,$ $X - \begin{pmatrix} x_{10} x_{11} \dots x_{1,p} \ 1 \\ x_{20} x_{21} \dots x_{2,p-1} \\ \dots \ \dots \ \dots \ \dots \\ x_{n0} x_{n1} \dots x_{n,p-1} \end{pmatrix},$

$Z = \begin{pmatrix} x_{1,p} x_{1,p+1} \dots x_{1,p+q} \\ x_{2,p} x_{2,p+1} \dots x_{2,p+q} \\ \dots \ \dots \ \dots \ \dots \\ x_{n,p} x_{n,p+1} \dots x_{n,p+q} \end{pmatrix}.$

Припускається, що стовпці матриці Z лінійно не залежать від стовпців матриці X , тобто матриця $W = (X, Z)$ розміру $n \times (q + p)$ має ранг $q + p$. Тоді маємо дві можливості знаходження оцінки найменших квадратів $\hat{\delta}_G$ вектора δ. Перша можливість (з використанням повної матриці експерименту):

$$\hat{\delta}_G = (W'W)^{-1}W'Y, D\hat{\delta}_G = \sigma^2(W'W)^{-1}.$$

Друга можливість (відповідає підходу з послідовним введенням нових даних): $\hat{\delta}_G = \begin{pmatrix} \hat{\beta}_G \\ \gamma_G \end{pmatrix}$, $\hat{\beta}_G = (X'X)^{-1}X'(Y - Z\gamma_G)$, $\gamma_G = (Z'RZ)^{-1}Z'RY$,

$$R = I_n - X(X'X)^{-1}X .$$

Саме ці дії належить виконати, якщо введення додаткових змінних використовує стандартне комп'ютерне забезпечення. Розроблене програмне забезпечення дозволяє будувати лінійні регресійні моделі даних у випадку, коли дані про змінні, що впливають на результат спостережень, поступають поступово, а попередні моделі вже побудовано. Зазначене програмне забезпечення виконує оцінки найменших квадратів параметрів моделі як за повністю сформованою матрицею експерименту, так і у випадку, коли розширення вказаних матриць виконується послідовно.

Висновки

Отже зазначене програмне забезпечення виконує оцінки найменших квадратів параметрів моделі як за повністю сформованою матрицею експерименту, так і у випадку, коли інформація про бажане включення в модель тих чи інших регресорних змінних поступає послідовно. Алгоритми послідовного оцінювання коефіцієнтів регресії, що застосовуються в останньому випадку, надають можливість уникнути нестійкого процесу обертання матриць великої розмірності, а також порівняти результати зазначених вище альтернативних методів оцінювання.

Література

1. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. — М. · С.-П. ·К.: 2007. — 911 с.
2. Себер Дж. Линейный регрессионный анализ. — М.: Мир, 1980. — 452 с.
3. Халафян А.А. STATISTICA Статистический анализ данных. — М.:БИНОМ, 2010. — 520 с.