УДК 519.237.5

РОЗРОБКА КОМП'ЮТЕРНОЇ ПРОГРАМИ ДЛЯ РЕГРЕСІЙНОГО АНАЛІЗУ СТАТИСТИЧНИХ ДАНИХ З ПОХИБКАМИ ВИМІРЮВАНЬ

Студ. І.С. Хорунжа, гр.МгІТ-2-17,
Науковий керівник проф. С.М. Краснитський
Київський національний університет технологій та дизайну

Мета роботи − розробити комп'ютерну програму для регресійного аналізу статистичних даних з урахуванням можливості похибок вимірювань регресорів.

Завдання роботи полягає у розробці програмного забезпечення для оптимального в певному сенсі рішення поставленої задачі, зокрема створення програмного продукту для регресійного аналізу статистичних даних з похибками вимірювань, а також порівняння отриманих результатів та формулювання висновків про якість роботи розглянутих алгоритмів.

Об'єктом дослідження є математичні моделі регресійного типу.

Предметом дослідження є статистичні дані, з наявними похибки вимірювань.

Методи та засоби дослідження. При побудові алгоритму побудови моделей:
- елементи теорії ймовірностей;
- методи оцінювання параметрів регресійних залежностей;
- статистичні дані;
- використання середньо геометричної функціональної залежності;

Для програмної реалізації програми використовувалася середа Microsoft Visual FoxPro.

Наукова новизна та практичне значення отриманих результатів. У різних галузях науки, техніки, економіки доводиться зустрічатися із ситуаціями, коли дані, що використовуються для побудови математичних моделей досліджуваних процесів, напевне містять похибки вимірювань, якими не можна нехтувати. У даний час виконується чимало досліджень, що мають на меті подолати вказані недоліки у вихідних даних. Однією з відомих останніх робіт, що присвячена даній тематиці, є книга [2]. Проте доступних і зручних у використанні комп'ютерних програм для реалізації, ще не вистачає.

Результати дослідження. Автор даної роботи пропонує програмне забезпечення для реалізації методу роботи з даними з похибками, що базується на принципі, котрий має назву середньогеометричної функціональної залежності.

Роздивимось спочатку звичайну ситуацію, коли мі маємо відгук Y та один предиктор X, та при цьому постулюється функція відгуку

$$Y=\beta_0+\beta_1 X+\varepsilon \tag{1}$$

у вигляді рівняння прямої з випадковим доданком. В таких випадках припускається, що відгук Y містить помилку, тоді як X — ні. Але можливо, що змінна X також схильна до помилки? Нехай η_i- істинне значення змінної Y_i, а ξ_i – істинна величина X_i, i=1,2…., n. Тоді спостережувані значення Y_i та X_i виражаються співвідношеннями

$$Y_i=\eta_i+\varepsilon_i$$
$$X_i=\xi_i+\delta_i,$$

іе ε_i та δ_i є випадковими похибками, які додаються до η_i та ξ_i відповідно, а ε_i незалежна від ξ_i та δ_i; δ_i незалежна від η_i та ε_i. Постулована нами модель має вигляд

$$\eta_i=\beta_0+\beta_1\xi_i$$

Підставляючи відповідні вирази, отримаємо

$$Y_i = \beta_0 + \beta_1 \xi_i + \varepsilon_i^*,$$

де$\varepsilon_i^* = (\varepsilon_i - \beta_1 \delta_i)$

Бачимо, що у випадку, коли змінні X схильні до похибок, як і змінні Y, то регресійна задача помітно ускладнюється, навіть коли модель однофакторна. Тому завжди корисно по можливості так організувати експеримент, щоб відношення $\sigma_\delta^2/\sigma_\xi^2$ було малою величиною. Практично це означає, що розкид величини ξ мірою якої служить σ_ξ^2, повинно суттєво перевищувати розкид помилок, які містяться в змінних X. Те ж само й вірно для інших предикторів. Після того, як це забезпечено, помилками в змінних X можна знехтувати та використовувати звичайний метод найменших квадратів. Якщо цього зробити не можна, то при використанні звичайного аналізу можуть виникнути ускладнення.

У даному випадку в роботі [1] рекомендується використання замість класичних оцінок методу найменших квадратів в моделі (1) використовувати оцінки середньогеометричної функціональної залежності

$$\hat{\beta}_1 = (S_{YY}/S_{XX})^{1/2},$$
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

Тут і далі $\bar{X}, \bar{Y}$ — середні арифметичні спостережуваних значень X,Y відповідно,

$$S_{XY} = \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}),\, S_{XX} = \sum_{i=1}^{n} (X_i - \bar{X})^2, S_{YY} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2.$$

Оцінка $\hat{\beta}_1$ тут являє собою середнє геометричне величин

$$b_1 = S_{XY}/S_{XX}\ ,\ \alpha_1^{-1} = (S_{XY}/S_{XX})^{-1},$$

де b_1, α_1 — кутові коефіцієнти, що підібрані методом найменших квадратів для залежностей Y від X і X від Y:

$$\hat{Y} = b_0 + b_1 X, \hat{X} = \alpha_0 + \alpha_1 Y.$$

Таким чином, геометричне середнє $\hat{\beta}_1 = (b_1 \alpha_1^{-1})^{1/2}$ є компромісним значенням, що лежить між двома кутовими коефіцієнтами рівнянь залежності Y від X. Обґрунтування вказаного методу одержуємо, застосовуючи метод максимальної правдоподібності, який при відомому значенні λ приводить до виразу

$$\hat{\beta}_1 = \left[S_{YY} - \lambda S_{XX} + \{(S_{YY} - \lambda S_{XX})^2 + 4\lambda S_{XY}^2\}^{1/2}\right]/2S_{XY,}\ \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

Наведені вище оцінки $\hat{\beta}_1, \hat{\beta}_0$ одержуємо, поклавши в останніх двох рівностях

Висновки. Програмне забезпечення, що реалізує вищеописані кроки, дозволить раціоналізувати роботу дослідника. Знання сумісних характеристик поведінки моделей лінійного регресійного типу дозволить ефективно впровадити це на практиці, а саме – дасть змогу ефективно використовувати ці методи в різних галузях.

Ключові слов: *регресійний аналіз, регресійний метод, статистика.*

ЛІТЕРАТУРА

1. Дрейпер Н, Смит Г. Прикладной регрессионный анадиз — М-К,***D***: 2007. - 911 с.