

АНАЛІЗ ЕФЕКТИВНОСТІ ПОЛІНОМІАЛЬНИХ ОЦІНОК ПАРАМЕТРІВ ЛІНІЙНОЇ РЕГРЕСІЇ ПРИ СИМЕТРИЧНОМУ РОЗПОДІЛІ НЕГАУСОВИХ ПОМИЛОК

Заболотній С.В., Рудь М.П., Ткаченко О.М.

Черкаський державний технологічний університет 18006, Україна, м. Черкаси, бул. Шевченка 460

Використання регресійних моделей і методів оцінювання їх параметрів є одним з найпоширеніших підходів, що застосовуються при аналізі статистичних даних різноманітного походження. Необхідність використання цього математичного апарату виникає при вирішенні багатьох задач, пов'язаних із проектуванням і моделюванням систем технічної діагностики та неруйнівного контролю.

Відомо, що основною метою регресійного аналізу є встановлення причинно-наслідкових зв'язків у вигляді деякої детермінованої функції, що описує залежність між значеннями змінної, яку називають цільовою, та однією або декількома незалежними змінними (предикторами). Оскільки реальні дані зазвичай мають стохастичний характер, то обов'язковим компонентом регресійних моделей є випадкова помилка, для опису якої, як правило, застосовують гаусовий закон розподілу ймовірностей. Проте значна частина сучасних досліджень свідчить про те, що гаусова модель часто є лише зручною ідеалізацією, використання якої хоча і дозволяє істотно спрощувати розрахунки, проте може призвести до суттєвого погіршення їх точності.

Існують різні підходи до дій в умовах неадекватності гаусової моделі при описі регресійних помилок. Один з них базується на застосуванні робастних модифікацій методу найменших квадратів (МНК), застосування яких спрямовано на забезпечення стійкості щодо впливу екстремальних відхилень (викидів). Проте існує цілий ряд ситуацій, коли відмінність від гаусового розподілу породжена не «шкідливими» поодинокими викидами, а специфікою даних. У цьому випадку використовують підхід, заснований на перетвореннях (Боксу-Кокса, Джонсона та ін.), що нормалізують розподіл статистичних даних. Але в тих ситуаціях, коли основним критерієм є мінімізація дисперсії оцінок параметрів детермінованих складових регресійних моделей застосовується параметричний підхід, який базується на методі максимальної правдоподібності (ММП). Проте цей підхід характеризується суттєвою обчислювальною складністю і необхідністю попередньої специфікації (вибору) закону розподілу ймовірностей для моделі регресійних помилок [1].

Компромісним з точки зору пошуку балансу між складністю і точністю є підхід до статистичного опрацювання, який базується на описі негаусових випадкових величин статистиках вищих порядків (моментами або кумулянтах). У роботі [2] на основі використання методу максимізації поліному (ММПл) [3] запропонована адаптивна процедура знаходження оцінок параметрів регресії. Її застосування базується на уточненні МНК-оцінок шляхом урахування додаткової апостеріорної інформації про специфіку моделі помилок у вигляді оціночних значень кумулянтних коефіцієнтів вищих порядків регресійних МНК-залишків.

Метою даної роботи є порівняльний аналіз ефективності застосування запропонованого підходу для ситуації, коли детермінована складова регресійної

моделі задається лінійною залежністю, а розподіл регресійних помилок відрізняється від гаусового закону, але характеризується симетрією. Як критерій ефективності розглядається величина відношення дисперсії ММПл і МНК оцінок параметрів лінійної регресії.

Проведені теоретичні дослідження показали, що при симетричному розподілі регресійних помилок МНК-оцінки можна трактувати як окремий випадок ММПл-оцінок, що отримуються при використанні поліномів 1-го порядку. При використанні поліномів 2-ї степені дисперсії оцінок ММПл і МНК також співпадають. Проте при використанні поліномів 3-ї степені відносна точність ММПл-оцінок зростає. Коефіцієнт зменшення дисперсії залежить від ступеня «негаусовості» моделі, яка чисельно описується значеннями кумулянтних коефіцієнтів 4-го (коефіцієнт ексцесу) та 6-го порядків.

Отримані шляхом статистичного моделювання методом Монте-Карло експериментальні результати підтверджують ефективність застосування ММПл. Ці результати отримані при багаторазових випробувань ($M=10^4$) для різних типів (арксинусного, рівномірного, трапецієвидного, трикутного і Лапласа) симетричних розподілів помилок лінійної регресії та різних обсягах вибірових значень ($N=50, 100, 200$). При цьому, як тип розподілу моделі помилок, так і значення параметрів (кумулянтних коефіцієнтів), необхідних для знаходження ММПл-оцінок вважалися апіорно невідомими, а використовувалися їх апостеріорні оцінки. Сукупність отриманих результатів моделювання також свідчить, що із зростанням обсягу вибірових даних експериментальні оцінки коефіцієнтів зменшення дисперсії прагнуть до теоретично розрахованих значень, а розподіл ММПл-оцінок нормалізується, що дозволяє аналітично визначати їх довірчі інтервали.

Перелік використаних джерел:

1. Ryan, T. P. (2008). *Modern regression methods (Vol. 655)*. John Wiley & Sons, Inc.
2. Заболотній С. В., Ткаченко О.М. Поліноміальні адаптивні процедури регресійного аналізу із використанням моделей негаусових помилок на основі статистик вищих порядків // Тези доповідей IV Міжнародної науково-практичної конференції «Обчислювальний інтелект – 2017» (ComInt – 2017)– К: КНУ ім.Т.Шевченка, – 2017. С. 113-114.
3. Кунченко Ю.П., Лега Ю.Г. *Оценка параметров случайных величин методом максимизации полинома.* – К.: Наукова думка, 1991. – 180 с.