УДК 519.854.2

НЕКОРЕКТНІСТЬ ВИКОРИСТАННЯ МЕТОДІВ БАГАТОВИМІРНОГО РЕГРЕСІЙНОГО АНАЛІЗУ ДЛЯ ВИПАДКУ ОДНОВИМІРНОГО ПОЛІНОМІАЛЬНОГО АНАЛІЗУ

професор, професор, доктор технічних наук, Павлов О. А.

Коваленко Д. А.

Національний технічний університет України “Київський політехнічний інститут імені Ігоря Сікорського”, Україна, Київ

Розглядається задача одновимірного поліноміального аналізу та показується некоректність спрощеного алгоритму вирішення задачі за допомогою зведення задачі одновимірного поліноміального регресійного аналізу до багатовимірного лінійного регресійного аналізу. Наведена аргументація для зведення задачі до спрощеного варіанту, приклади коли така інтерпретація є коректною та методи уникнення наслідків спрощення. Наведені логічні висновки та експериментальні дані що показують що дане спрощення може призвести до неправильної інтерпретації результатів та ненадійних оцінок нелінійних членів регресії.

Ключові слова: одновимірна поліноміальна регресія, багатовимірна лінійна регресія, центрування, корреляція, регресійні моделі, незалежні випадкові величини.

Вступ. У статистиці, поліноміальна регресія це підвид регресійного аналізу у якому залежність між незалежною змінною X та залежною змінною У є нелінійним і моделюється через поліном степені n. Поліноміальна регресія використовується для опису процесів росту живих тканин, хімічних процесів, розподілу ізотопів та поширення епідемій. Поліноміальна регресія містить нелінійні

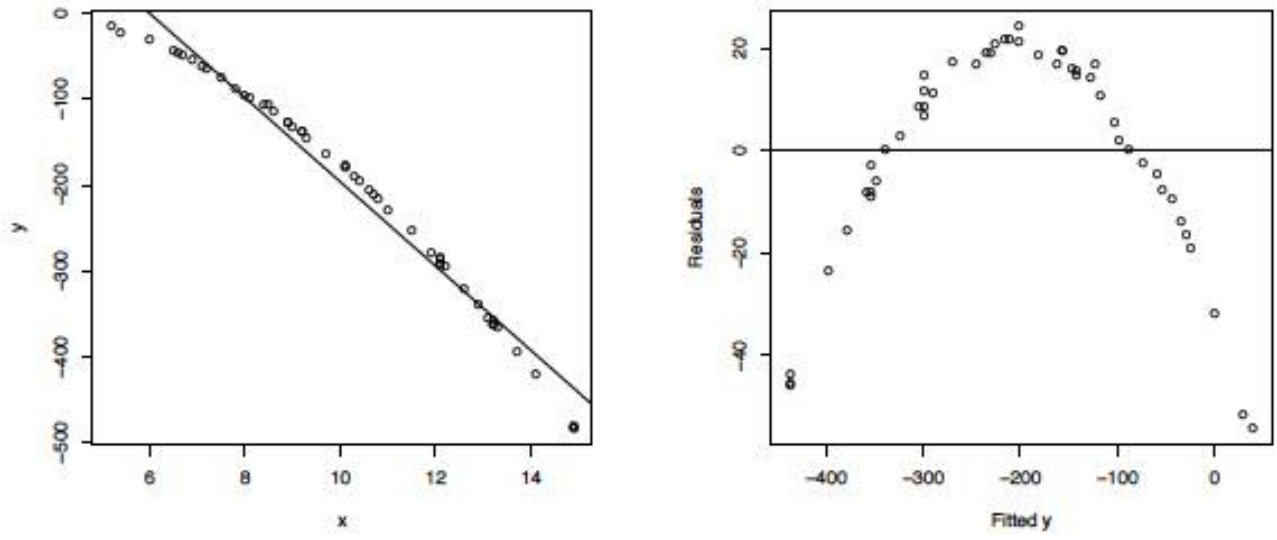
відношення, але її можна сформулювати у вигляді статистичної оцінки нелінійних параметрів. Це дозволяє використовувати метод багатовимірної лінійної регресії для знаходження коефіцієнтів при нелінійних змінних.

Мета та завдання статті. Метою статті є опис популярного алгоритму поліноміальної регресії за допомогою методів багатовимірного регресійного аналізу та доказ некоректності такого підходу у загальному випадку.

Постановка задачі.

Проблема відтворення невідомої залежності формулюється як класична задача прикладного регресійного аналізу: відтворення багатовимірної поліноміальної регресії по надлишковому опису і з довільно розподіленою похибкою. По результатам активних експериментів необхідно знайти невідомі коефіцієнти, частина з яких тотожно дорівнює нулю і невідома досліднику.

У будь-якій сфері діяльності людини перед дослідником невідомого явища або процесу постає проблема ефективного проведення експериментів та знаходження невідомої залежності (лінії



регресії) з високою точністю.

Рис 2.1 - Лінія регресії

Математична модель

Модель одновимірної поліноміальної регресії має вигляд:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + ... + \beta_m x_i^m + \varepsilon_i (i = 1, 2, ..., n)$$

може бути представлена у вигляді матриці X, вихідного вектора y, вектора параметрів в та вектора випадкових помилок ε. *i*-тий рядок X та y буде містити значення x та y для і-того значення із вхідних-вихідних даних. Модель може бути записана як система лінійних рівнянь:

$$\begin{bmatrix} y_1 \\ y_1 \\ y_1 \\ \text{M} \\ y_1 \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \text{L} & x_1^m \\ 1 & x_2 & x_2^2 & \text{L} & x_2^m \\ 1 & x_3 & x_3^2 & \text{L} & x_3^m \\ \text{M} & \text{M} & \text{M} & \text{O} & \text{M} \\ 1 & x_n & x_n^2 & \text{L} & x_n^m \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \text{M} \\ \beta_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \text{M} \\ \varepsilon_m \end{bmatrix} \quad (1)$$

Якщо використовувати матричну форму запису, отримуємо:

$$\vec{y} = X\vec{\beta} + \vec{\varepsilon}$$

Використовуючи звичайний метод найменших квадратів, знаходимо вектор оцінок коефіцієнтів поліноміальної регресії дорівнює:

$$\vec{\beta} = (X^T X)^{-1} X^T \vec{y}$$

Ми припускаємо, що обсяг вибірки більший за ступінь полінома, тоді, так як матриця X є матрицею Вандельмонда, вона є невиродженою. Отже, ми отримали результат за допомогою методу найменших квадратів, який є унікальним.

Порушення умови незалежності змінних

Звичайний метод найменших квадратів, що використовується для знаходження коренів рівняння гарантовано дає найкращі оцінки параметрів регресійної моделі при даних умовах:

1) Математичне очікування помилок дорівнює нулю;

2) Дисперсія помилок постійна;

3) Випадкові помилки є незалежними між собою
4) Випадкові помилки є незалежними від незалежних змінних X;
5) Модель є лінійною відносно параметрів
6) Відсутність мультиколінеарності – між незалежними змінними не повинно бути сильною залежності;
7) Випадкові помилки розподілені нормально.

Умови 1-5 та 7 виконуються для (1), перевіримо умову 6. Для цього побудуємо кореляційну матрицю для даних:

$$X = [20 \quad 33.96 \quad 17.58 \quad 64.4 \quad 55.74 \quad 64.85 \quad 11.12 \quad 28.99 \quad 96.25 \quad 2.85]$$

Таблиця 1.

Матриця Вандермонда

x^1	x^2	x^3	x^4	x^5
6.64E+01	4.41E+03	2.93E+05	.95E+07	.29E+09
4.83E+01	2.33E+03	1.13E+05	5.45E+06	2.63E+08
6.14E+01	3.77E+03	2.31E+05	1.42E+07	8.71E+08
3.38E+01	1.14E+03	3.85E+04	1.30E+06	4.40E+07
3.22E+01	1.04E+03	3.34E+04	1.08E+06	3.46E+07
3.82E+01	1.46E+03	5.56E+04	2.12E+06	8.10E+07
4.34E+01	1.89E+03	8.20E+04	3.56E+06	1.55E+08
4.29E+01	1.84E+03	7.88E+04	3.38E+06	1.45E+08
5.25E+01	2.76E+03	1.45E+05	7.60E+06	3.99E+08
9.31E+01	8.67E+03	8.07E+05	7.52E+07	7.00E+09

Таблиця 2.

Кореляційна матриця

	x^1	x^2	x^3	x^4	x^5
x^1	1	0.97	0.92	0.87	0.83
x^2	0.97	1	0.99	0.96	0.93
x^3	0.92	0.99	1	0.99	0.98

x^4	0.87	0.96	0.99	1	1
x^5	0.83	0.93	0.98	1	1

Як бачимо, кореляція між степенями у матриці Вандермонда досить високі і стовпчики цієї матриці не можуть бути використані як незалежні змінні при звичайному методі найменших квадратів. Розглянемо метод який використовуються для зменшення негативних ефектів зведення поліноміальної функції до лінійного виду.

Центрування

Для покращення результатів, використовують центрування вхідних даних:

$$X' = X - \bar{X}$$

Таблиця 3.

Кореляційна матриця після центрування

	x^1	x^2	x^3	x^4	x^5
x^1	1	0.18	0.93	0.2	0.84
x^2	0.18	1	0.23	0.96	0.28
x^3	0.93	0.23	1	0.29	0.98
x^4	0.2	0.96	0.29	1	0.35
x^5	0.84	0.28	0.98	0.35	1

Висновки. В даній статті було показано, що використання методів лінійного регресійного аналізу для задачі поліноміальної регресії не є коректним. Хоча лінія регресії отримана даним способом і оцінює дані досить точно, але знайдені коефіцієнти не можуть інтерпретуватися як реальні – адже кореляція між відповідними членами регресії є досить висока. Існують методи зменшення наслідків даного спрощення, а саме центрування та ортогональні поліноми, але вони не можуть гарантувати коректну роботу алгоритму

МНК у загальному випадку. Рекомендується використовувати більш досконалі алгоритми, такі як сплайни, регресійні дерева та інші.

Література:

1. Адлер Ю. П., Маркова Е. В., Грановский Ю. В. Планирование эксперимен- та при поиске оптимальных условий. – 2-е изд., перераб. и доп. – М.: Наука, 1976. – 280 с.

2. Айвазян С. А. Многомерный статистический анализ // Математическая энциклопедия / Гл. ред. И. М. Виноградов. – М., 1982. – Т.3. – Стб. 732- 738.

3. Аксенова Л. А. Новые полиномиальные подклассы труднорешаемой задачи «Минимизация суммарного взвешенного момента» для множества одного приоритета // Управляющие системы и машины, – 2002.– No6.– С.21-28

4. Андерсон Т. Введение в многомерный статистический анализ / Пер. с англ. Ю. Ф. Кичатова; Под ред. Б. В. Гнеденко. – М.: Физматтиз, 1963. – 500 с.

5. Веселов С. И., Шевченко В. Н. Об экспоненциальном росте коэффициентов агрегирующего уравнения: Тез. докл. 4 Феодосийской конф. по пробл. теорет. кибернетики. – Новосибирск: Ин-т математики СО АН СССР, 1977. – 53 с.

6. Гери М. Р., Джонсон Д. С. Вычислительные машины и труднорешаемые задачи. – М.: Мир, 1982. – 416 с.

7. Д. Худсон. Статистика для физиков. Москва, Мир, 1970.

8. Ершов А. А. Стабильные методы оценки параметров: (Обзор) // Автоматика и телемеханика. – 1978. – No 8. – С. 66-100.

9. Згуровский М. З., Павлов А. А. Иерархическое планирование в системах, имеющих сетевое представление технологических

процессов и ограни- ченные ресурсы, как задача принятия решений // Системні дослідження та інформаційні технології.– 2009.

References:

1. Adler Yu. P., Markova Ye. V., Granovskiĭ Yu. V. Planirovanie eksperimen- ta pri poiske optimalnykh uslovieĭ. – 2-e izd., pererab. i dop. – M.: Nauka, 1976. – 280 s.

2. Aĭvazyan S. A. Mnogomernyĭ statisticheskiĭ analiz // Matematicheskaya entsiklopediya / Gl. red. I. M. Vinogradov. – M., 1982. – T.Z. – Stb. 732- 738.

3. Aksenova L. A. Novye polinomialnye podklassy trudnoreshaemoĭ zadachi «Minimizatsiya summarnogo vzveshennogo momenta» dlya mnozhestva odnogo prioriteta // Upravlyayushchie sistemy i mashiny, – 2002.– No6.– S.21-28

4. Anderson T. Vvedenie v mnogomernyĭ statisticheskiĭ analiz / Per. s angl. Yu. F. Kichatova; Pod red. B. V. Gnedenko.–M.: Fizmattiz, 1963.– 500 s.

5. Veselov S. I., Shevchenko V. N. Ob eksponentsialnom roste koeffitsientov agregiruyushchego uravneniya: Tez. dokl. 4 Feodosiĭskoĭ konf. po probl. teoret. kibernetiki. – Novosibirsk: In-t matematiki SO AN SSSR, 1977. – 53 s.

6. Geri M. R., Dzhonson D. S. Vychislitelnye mashiny i trudnoreshaemye zadachi. – M.: Mir, 1982. – 416 s.

7. D. Khudson. Statistika dlya fizikov. Moskva, Mir, 1970.

8. Yershov A. A. Stabilnye metody otsenki parametrov: (Obzor) // Avtomatika i telemekhanika. – 1978. – No 8. – S. 66-100.

9. Zgurovskiĭ M. Z., Pavlov A. A. Ierarkhicheskoe planirovanie v sistemakh, imeyushchikh setevoe predstavlenie tekhnologicheskikh protsessov i ogranichennye resursy, kak zadacha prinyatiya reshenіĭ // Sistemni doslidzhennya ta informatsiĭni tekhnologiï.– 2009.