УДК 004.891.2

МЕТОДИКА ПОБУДОВИ РЕГРЕСІЙНИХ МОДЕЛЕЙ В УМОВАХ ЕФЕКТУ МУЛЬТИКОЛІНЕАРНОСТІ

Малкіна В. М.[1], д.т.н. *e-mail*: *vira.malkina@tsatu.edu.ua*
[1]*Таврійський державний агротехнологічний університет імені Дмитра Моторного*

Актуальність та постановка проблеми. Як відомо, модель з категорії Machine Learning, що встановлює зв'язок між одним або декількома незалежними предикторами і залежною континуальною ознакою має вигляд регресійної моделі. При побудові регресійних моделей, що описують різні процеси часто дослідники ігнорують ефект мультиколінеарності факторів. В цьому випадку інтерпретації оцінок параметрів регресійної моделі є неадекватними. Пропонується при проведенні регресійного аналізу проводити діагностику мультиколінеарності як один з етапів регресійного аналізу. Таким чином існує проблема розробки спеціальних методик побудови регресійних моделей в умовах мультиколінеарності.

Основні матеріали дослідження. У сучасних дослідженнях в Data Science досить часто застосовують методи кореляційно-регресійного аналізу. Розрахунок коефіцієнтів кореляції дозволяє виявити тісноту і напрямки зв'язку досліджуваних показників. Регресійний аналіз, який є продовженням кореляційного аналізу, дозволяє визначити аналітичне представлення зв'язку результуючої величини з факторними показниками. Використання регресійної моделі дозволяє сформувати важливі висновки про вплив кожного фактора на результуючі ознаки і оцінити ступінь цього впливу, прогнозувати результати управлінського впливу на зміну значень факторів даної моделі, аналізувати і визначати значення факторів для забезпечення оптимального значення результуючих ознак.

Як відомо, необхідною умовою для отримання незміщених ефективних оцінок параметрів регресійної моделі є умови теореми Гауса-Маркова [1].

Одна з умов теореми - вектори-фактори повинні бути лінійно незалежні. Проблемою при побудові адекватної регресійної моделі є наявність корельованих незалежних змінних, тобто, мультиколінеарності.

Ефект мультиколінеарності означає, що принаймні дві незалежні змінні, які впливають на предикату мають тісний кореляційний зв’язок. Питанню складнощів і негативному впливу мультиколінеарності на весь процес дослідження присвячено багато публікацій [3,4]. Як описано в літературі, основна проблема при прояві мультиколінеарностіє нестабільні і зміщені похибки параметрів регресійної моделі, що приводить до такого значення показника p-values, яке перевищую допустиме значення рівня значущості. Як результат, це призводить до неефективних оцінок, які не дають можливості адекватного аналізу процесу на основі такої регресійної моделі.

Ефект мультиколінеарності приводить до неадекватної інтерпретації оцінок впливу окремих змінних на предикату на основі регресійної моделі.

Таким чином, слід розглядати дві проблеми - виявлення мультиколінеарності і побудувати адекватної регресійної моделі при наявності мультиколінеарності.

Відомо багато способів виявлення мультиколінеарності. Одним з широко використовувані методів виявлення мультиколінеарності є метод на основі

показника variance inflation factor (VIF) [3], який показує, як збільшилася загальна дисперсія в порівнянні з дисперсією моделі однофакторної регресії

$$VIF = \frac{1}{1-R_j^2}, \quad (1)$$

де - R_j^2 коефіцієнт детермінації в регресії предикати від змінної $x_j, j = 1..n$.

Відомо [1], що, при значенні $VIF > 10$ присутній ефект мультиколінеарності.

Також, ознакою мультиколінеарності є значення коефіцієнта детермінації близького до 1, адекватність моделі за критерієм Фішера, і, одночасно, велика кількість не значимих, за критерієм Ст'юдента, параметрів моделі.

У випадку, коли наявність мультиколінеарності підтверджена, пропонуються наступні підходи до регресійного аналізу.

Перший підхід. Щоб оцінки параметрів регресійної моделі, в разі виявленої мультиколеніарності, були ефективними і надійними, пропонується використовувати методи регуляризації, які коригують відхилення від нормального розподілу залишків. Таким методом є метод LASSO (Least absolute shrink age and selection operator).У статті [6] на підставі запропонованої методики була побудована і проаналізована модель, яка описує вплив показників зберігання зерна (температура зернової маси, вологість зерна, температура повітря в зерносховище, температура холодоагенту і обсяг подачі повітря) на якісні характеристики зерна (клейковини).

Іншім підходом до усунення ефекту мультиколінеарності та побудови моделі регресії є видалення з розгляду корелюючих факторів.

З метою виявлення тих факторів, які слід залишити, пропонується побудувати регресійні моделі на різних наборах факторів та порівняти отримані моделі за набором критеріїв якості моделей.

По-перше, для порівняння якості моделей використовують критерій Акаіке AIC [2].

$$AIC = 2k + n\left(\ln\left(\frac{2\pi \cdot RSS}{n}\right) + 1\right), \quad (2)$$

де $RSS = \sum e_i^2$ - сума квадратів залишків регресії;

n - кількість спостережень;

k - кількість параметрів регресійної моделі.

По друге, для порівняння регресійних моделей з різною кількістю факторів використовують нормований коефіцієнт детермінації :

$$R^2_{adjusted} = 1 - (1 - R^2) \cdot \frac{n-1}{n-k}, \quad (3)$$

де - R^2- коефіцієнт детермінації регресійної моделі;

n – кількість спостережень;

k – кількість параметрів регресійної моделі.

Таким чином, в якості критеріїв для порівняння моделей пропонується наступні: кількість факторів, залучених в моделі, кількість незначущих факторів в моделі, кількість факторів, які мають значення показника $VIF > 10$ (тобто

кількість факторів, які породжують ефект мультиколінеарності), значення оцінки моделі за критерієм AIC (вважається, що, чим краще модель, тим менше значення критерію), значення нормованого коефіцієнта детермінації.

Запропонований підхід був реалізований при регресійному аналізі урожайності вишні від кліматичних умов вирощування в роботі [5].

Висновки. Часто при побудові регресійних моделей, які описують процеси вирощування, зберігання і переробки сільськогосподарської продукції не враховують ефект корельованості факторів. Несприятливий ефект мультиколінеарності негативно позначається на інтерпретації побудованої моделі, а саме при аналізі ступеня впливу кожного фактору окремо на досліджуваний показник. Ефект мультиколінеарності робить, практично неможливим адекватну інтерпретацію оцінки впливу кожного фактору на результуючу ознаку на підставі регресійній моделі. Запропоновано методику побудови і аналізу регресійних моделей, яка дозволяє побудувати ефективні оцінки параметрів регресії в умовах мультиколінеарності факторів на основі порівняння різних моделей за набором критеріїв.

Список використаних джерел:

1. Kutner M. H., Nachtsheim C., Neter J. Applied Linear Statistical Models (4thedn.) McGraw-Hill Education, 2004. 701 p.

2. Aiken LS, West SG (1991) Multipleregression: Testing and interpreting interaction. Newbury Park C. editor. SAGE Publication. Inc

3. Damodar N. Gujarati. Basic Econometrics. - 4. TheMcGraw-HillCompanies, 2004. 1002 p.

4. Gordon R. A. Issues in Multiple Regression. *The American Journal of Sociology*. 1968. №78. P. 592-616.

5. Малкіна В. М., Іванова І. Є., Сердюк М. Є., Крівонос І. А., Білоус Е. С. Регресійний аналіз залежності урожайності вишні від гідротермічних факторів в умовах мультиколінеарності. *Наукові горизонти*: збірник наукових праць. Житомир, 2019. Вип. 11 (84). С. 51-60.

6. Vira Malkina, Serhii Kiurchev, Valentina Verkholantseva, Viktor Dubik. Multicollinearity in the regression analysis of the wheat gluten indicator duringits storage. *Engineering for rural development Jelgava*. (Latvia University of Sciences and Techologies Faculty of Engineering 20 -22.05.2020). Latvia, 2020. P. 985 -990.