УДК 512.9:004.852

Ілик В. В., здобувач вищої освіти,
Хмелівський Ю. С., асистент кафедри інформаційних технологій,
Донецький національний університет імені Василя Стуса

ОЦІНКА ТОЧНОСТІ ЛІНІЙНОЇ РЕГРЕСІЇ В УМОВАХ МАЛИХ ВИБІРОК

Анотація. Лінійна регресія є основним методом статистичного аналізу, але в умовах малих вибірок її точність може знижуватися. Стаття розглядає методи покращення надійності моделей, як-от регуляризація, крос-валідація та баєсівські підходи з акцентом на прикладах із медичних, соціологічних і ринкових досліджень.

Ключові слова: лінійна регресія, малі вибірки, регуляризація, крос-валідація, баєсівські методи.

Вступ. Лінійна регресія є основним методом статистичного аналізу, що використовується для моделювання та вивчення зв’язків між залежною змінною (наприклад, результатом або відповіддю) і однією або кількома незалежними змінними (факторами, що впливають на результат). Це простий, але потужний інструмент, який дає змогу не лише описувати дані, а й робити прогнози [1].

Лінійна регресія передбачає, що залежність між змінними може бути описана лінійною функцією. Однак щоб забезпечити адекватність моделі, необхідно дотримуватись певних передумов, як-от лінійність, незалежність залишків, нормальність залишків та однорідність дисперсії. Наявність достатньої кількості даних є критично важливою, оскільки невеликий обсяг вибірки може призводити до ненадійних та нестабільних оцінок параметрів.

У малих вибірках статистичні оцінки стають менш надійними. Це пояснюється декількома факторами:

- Висока варіабельність. Невеликі вибірки можуть містити випадкові відхилення, що значно впливають на оцінки параметрів. Це може призвести до великої варіації в стандартних помилках, що ускладнює оцінку точності та довірчих інтервалів.
- Перенавчання (overfitting). За малих вибірок модель може виявити надмірну відповідність до даних, що не відображає реальної тенденції. Це призводить до ненадійних прогнозів на нових даних.
- Недостатність даних для перевірки. У малих вибірках важко провести перевірку моделей через обмежений обсяг даних. Це може ускладнити використання важливих методів, як-от крос-валідація.

Отже, проведення аналізу в умовах малих вибірок вимагає особливої уваги до обраної методології, а також до інтерпретації отриманих результатів [1].

З метою покращення точності оцінок у малих вибірках використовуються методики регуляризації, крос-валідації та зменшення розмірності.

Регуляризація – це метод, який дає змогу зменшити перенавчання шляхом додавання штрафу за складність моделі. Lasso-регресія (L1-регуляризація) та

Ridge-регресія (L2-регуляризація) є популярними техніками, які допомагають зменшити величину коефіцієнтів, що може зменшити вплив шуму на модель.

Крос-валідація полягає у розподілі даних на навчальну та тестову вибірки кілька разів, що дає змогу оцінити узагальнюючу здатність моделі. Зазвичай використовується k-fold крос-валідація, де дані розбиваються на k частин, а модель навчалася та тестувалася k разів [2].

Використання методів, як-от відбір змінних, допомагає зменшити кількість незалежних змінних до найбільш релевантних, що може підвищити точність оцінок. Основні підходи включають методи на основі кореляції або різні алгоритми відбору.

Застосування цих технік може допомогти покращити точність моделей в умовах малих вибірок, забезпечуючи більш стабільні та надійні результати [2].

Довірчі інтервали надають діапазон, у якому з певною ймовірністю (зазвичай 95 %) може знаходитися справжнє значення коефіцієнта. Цей показник дає змогу дослідникам оцінювати надійність своїх висновків і перевіряти статистичну значущість змінних у моделі.

Коефіцієнт детермінації показує, яка частина варіації залежної змінної може бути пояснена незалежними змінними в моделі. Значення R^2 коливається від 0 до 1, де 1 свідчить про те, що модель ідеально підходить до даних. Проте в умовах малих вибірок R^2 може бути оманливим, оскільки його можна підвищити шляхом додавання додаткових незалежних змінних, навіть якщо вони не мають значущого впливу. Тому важливо використовувати скоригований R^2, який враховує кількість незалежних змінних у моделі і є більш надійним показником якості, особливо в умовах малих вибірок.

Аналіз залишків є важливим етапом у перевірці припущень моделі. Залишки – це різниця між спостережуваними значеннями та значеннями, передбаченими моделлю. Аналіз залишків дає змогу перевірити лінійність, нормальність і однорідність дисперсії. Графіки залишків допомагають виявити наявність патернів або трендів, які можуть свідчити про проблеми в моделі. Наявність трендів або груп може свідчити про ненадійність моделі [3].

Баєсівські методи є потужним інструментом для аналізу даних, особливо в умовах малих вибірок. Цей підхід базується на теорії ймовірностей і дає змогу враховувати невизначеність в даних та моделях [3].

Однією з основних переваг байєсівського підходу є можливість моделювання невизначеності шляхом використання апріорних розподілів. Апріорні розподіли можуть базуватися на попередньому досвіді або інформації, що вже є. Це особливо корисно у випадках, коли дані обмежені. Вибір правильного апріорного розподілу може суттєво вплинути на результати. Наприклад, можна використовувати нормальний розподіл для оцінки коефіцієнтів, якщо є підстави вважати, що вони розподілені нормально.

Баєсівські методи дають змогу моделювати як лінійні, так і нелінійні зв’язки. Це може бути особливо корисно в умовах малих вибірок, де стандартна лінійна регресія може не відображати реальну залежність між змінними. Наприклад, можна використовувати методи, як-от регресія на основі дерев, які можуть адаптуватися до складних залежностей між змінними.

Баєсівські моделі дають змогу інтегрувати нові дані в наявну модель без її повного переобчислення. Це забезпечує гнучкість і дає змогу моделі залишатися актуальною в умовах, коли з'являються нові дані. Наприклад, у випадку, коли нові спостереження з'являються в часі, баєсівський підхід може швидко оновити оцінки коефіцієнтів без потреби в повторному навчанні моделі з нуля.

Також баєсівські методи мають потенціал для побудови складних ієрархічних моделей, які можуть моделювати різні рівні невизначеності в даних, що особливо корисно для досліджень у соціальних науках або медицині [4].

У медицині часто стикаються з проблемою обмежених даних, особливо під час досліджень рідкісних захворювань. У таких випадках важливо враховувати статистичну значущість результатів і можливість перенавчання моделей. Наприклад, у дослідженнях, що вивчають зв'язок між ліками та побічними ефектами, застосування байєсівських методів може допомогти інтегрувати інформацію з інших джерел та зменшити невизначеність в оцінках.

У соціології часто виникають ситуації, коли вибірка є малою через обмеження бюджету або доступності даних. У таких випадках важливо використовувати регуляризацію для уникнення перенавчання та скоригований R^2 для оцінки якості моделі. Наприклад, в дослідженнях, що вивчають соціально-економічні фактори, варто застосовувати техніки крос-валідації для перевірки надійності отриманих результатів.

У бізнесі, особливо в умовах стартапів, отримання великих обсягів даних може бути складним завданням. Наприклад, під час проведення опитувань для визначення споживчих вподобань важливо використовувати методи, як-от крос-валідація, для перевірки надійності результатів. Також може бути корисно використовувати альтернативні підходи, як-от побудова експериментів, щоб отримати додаткові дані [5].

Висновки. Для підвищення якості та точності моделей в умовах малих вибірок важливо використовувати низку підходів та технік. По-перше, необхідно регулярно проводити перевірки моделей за допомогою таких методів крос-валідації, щоб оцінювати їх ефективність. По-друге, застосування регуляризації допоможе зменшити ризик перенавчання. Баєсівські підходи також треба розглядати як ефективний інструмент для отримання надійних результатів в умовах невизначеності. Також аналіз залишків є важливим кроком, який дає змогу виявити потенційні проблеми в моделях і покращити їх точність. Для інвесторів чи дослідників це може бути особливо корисним під час обмежених даних, наприклад, в аналізі нерухомості чи медичних дослідженнях.

Список використаних джерел

1. Нельсон Д. Що таке лінійна регресія? 2021. URL: https://www.unite.ai/uk/what-is-linear-regression/ (дата звернення: 22.10.2024).

2. Cross-Validation Modeling: вебсайт. 2023. URL: https://itwiki.dev/data-science/ml-reference/ml-glossary/cross-validation-modeling (дата звернення: 22.10.2024).

3. Бондаренко Я. С., Рачко Д. О., Розливан А. О. Посібник до вивчення дисципліни «Імовірнісні графічні моделі». Частина 2. Навчання байєсівської мережі. Дніпро: Ліра, 2020. 40 с.