Аналіз даних
Лабораторна робота №3: Регресійний аналіз

Данило Тавров

1 Завдання на роботу

1.1 Ідея роботи

У Лабораторній роботі 2 студенти намагалися застосувати методи статистичного виведення та встановити статистичну значущість тих спостережень, які було виявлено за результатами виконання Лабораторної роботи 1. На той момент студенти вимушені були одночасно аналізувати максимум дві змінні. У поточній роботі стоїть задача встановити залежність змінної, що становить інтерес, від декількох інших змінних.

Зокрема, у цій лабораторній роботі потрібно збудувати (лінійну) регресійну модель та оцінити її параметри. Студенти повинні детально проаналізувати модель та розглянути різні варіанти її специфікації.

1.1.1 Моделювання

Перше, що потрібно зробити — сформулювати дослідницьке питання, яке повинно мати причиново-наслідковий характер, тобто відповідь на яке повинна дати підстави стверджувати, чи існує вплив деякого фактора чи факторів на змінну, яка становить інтерес.

Відповідне дослідницьке питання повинно бути відображено у відповідній *структурній моделі*. Будуючи таку модель, студенти повинні:

- вибрати залежну та незалежні змінні та пояснити свій вибір;
- проаналізувати потребу застосування логаритмів до обраних змінних та зробити відповідні висновки;
- детально прокоментувати потенційну наявність зміщення від неврахованих змінних (omitted variable bias, OVB). Потрібно вказати, які невраховані змінні можуть вести до порушення умови про нульове умовне сподівання похибки;
- зазначити, як контрольні змінні було б *ідеально* мати в наявності, і які контрольні змінні є наявні *по факту*;
- сформулювати *гіпотези* щодо значень (або принаймні знаків) коефіцієнтів відповідної структурної моделі.

1.1.2 Перевірка моделі на стійкість

Для того, щоб показати тими обмеженими засобами, які в нас є, що оцінки коефіцієнтів моделі можуть викликати певну довіру, потрібно продемонструвати її стійкість (robustness) до потенційно некоректних варіантів специфікації. Для цього, окрім базової моделі та моделі «просто» з додатковими контрольними змінними, студенти повинні проаналізувати такі можливості:

- додавання поліномів вищих порядків відносно окремих регресорів. Додавання ступенів (квадратів, кубів тощо) регресорів варто аргументувати візуально, показуючи наявність нелінійного зв'язку між змінними;
- додавання факторів взаємодії між регресорами. Мотивація розгляду цих факторів повинна випливати з аналізу предметної області, тобто варто аргументувати можливість зміни впливу одного регресора залежно від значень іншого;

- взяття логаритмів окремих змінних та ступенів таких логаритмів;
- створення нових індикаторних (бінарних) змінних на основі неперервних (наприклад, розбити всіх людей на «молодих» і «старих» тощо).

Модель буде вважатися стійкою, якщо значення коефіцієнтів біля ключових регресорів не будуть сильно змінюватися від модифікації функціональної специфікації моделі. В іншому випадку можна зробити висновок, що в модель можна додати нові контрольні змінні для зменшення OVB.

Студенти не повинні забувати, що контрольну змінну, яка є категорійною (**особливо** якщо її значення закодовано числами!), потрібно включати в модель як декілька *бінарних* змінних, по одній змінній на кожну категорію (за виключенням деякої базової для уникнення повної мультиколінеарности).

1.1.3 Оцінювання моделей

Студенти повинні оцінити як базову модель, так і моделі з контрольними змінними та різними функціональними формами цих змінних. Оцінювання варто проводити методом OLS (або FE, якщо дані мають панельну природу).

Оцінюючи моделі, студенти повинні розглянути такі нюанси та прокоментувати їх у звіті та презентації:

- які стандартні похибки було використано і чому. **У жодному разі не можна використовувати гомоскедастичні похибки!** Можливим є варіант використання **кластеризованих** похибок, тому студенти повинні чітко зазначити, чому вони вирішили використати їх, і на якому рівні робили кластеризацію;
- яку інтерпретацію мають коефіцієнти біля ключових регресорів;
- чи є статистично значущі коефіцієнти біля ключових регресорів;
- чи є статистично значущі *групи коефіцієнтів*. Обов'язково варто тестувати на спільну значущість такі групи коефіцієнтів: усі ступені одного регресора вище 1, щоб перевірити, чи справді має місце нелінійність; регресор та всі фактори взаємодії з ним, щоб перевірити, чи справді значущим є ефект від цього регресора; усі бінарні змінні, утворені на основі однієї категорійної, щоб перевірити, чи справді значущим є ефект від *усієї* категорійної змінної; інші ситуації, коли доцільно тестувати групу коефіцієнтів, а не коефіцієнти окремо.

Доцільно проводити аналіз на **мультиколінеарність**. У нашому курсі ми не розглядаємо специфічних методів, тому достатньо просто навести кореляції між змінними та прокоментувати відповідні ризики.

1.1.4 Особливості датасетів

Студенти повинні бути свідомі того, що їхні датасети можуть мати одну або декілька з особливостей:

- мати панельну природу;
- містити бінарну залежну змінну;
- містити цензуровану залежну змінну;
- містити значну кількість пропущених даних, які неможливо просто взяти і викинути.

У цих випадках студенти повинні використовувати особливі методи, розглянуті на лекційних заняттях. Відмазки з притягнутими за вуха поясненнями, чому, наприклад, для панельних даних було використано звичайні методи перехресного аналізу — **не прийматимуться**.

1.2 Звітність щодо результатів роботи

Як і в усіх лабораторних роботах, студенти повинні надати викладачу:

- програмний код в R. Код повинно бути написано так, **щоб він успішно виконався на комп'ютері викладача**;
- звіт із лабораторної роботи у **форматі PDF** у **довільній формі** (див. нижче вимоги);
- презентацію до звіту у **форматі PDF** (див. нижче вимоги).

Презентацію результатів студенти повинні здійснити на одному з лабораторних занять або в інший час за домовленістю з викладачем. Тривалість презентації повинна складати **орієнтовно 15 хвилин**. Презентацію **буде записано** та **поширено серед студентів потоку**. На основі презентацій студентів **буде сформульовано окремі питання на заліковій контрольній роботі**.

1.3 Загальні вимоги до наповнення звіту

Звіт із лабораторної роботи повинен містити такі обов'язкові елементи:

- назву;
- вступ із мотивацією проведення дослідження. Потрібно чітко вказати, на які дослідницькі питання було намагання дати відповідь, зокрема, як вони пов'язані з питаннями Лабораторної роботи 1. Важливо вказати гіпотези щодо значень (або принаймні знаків) коефіцієнтів регресійної модели;
- результати проведеного аналізу. У цьому контексті основною є **таблиця** на кшталт наведених у лекціях та в завданні 4 МКР, де кожний стовпчик відповідає окремій оцінюваній моделі. Це повинна бути не «таблиця», яку видає R в консолі, а справжня таблиця, у коректний спосіб відформатована. Кожну розглянуту в таблиці модель потрібно прокоментувати відповідно до вимог, зазначених вище (чому саме такі змінні включено, чи є коефіцієнти та групи коефіцієнтів статистично значущі, яка їхня інтерпретація тощо);
- результати тестування гіпотез про статистичну значущість як окремих коефіцієнтів, так і груп коефіцієнтів (там, де це виправдано);
- висновки та огляд можливих обмежень проведеного аналізу (які можуть існувати причини сумніватися в причиново-наслідковій інтерпретації презентованих результатів);
- посилання на використані джерела.

1.4 Поради щодо оформлення презентацій

Захист кожної лабораторної роботи повинен супроводжуватися відповідною презентацією. Потрібно розуміти, що метою презентації є не детальний опис усіх кроків аналізу тощо, які можна прочитати у звіті чи навіть у коді, а саме презентація дослідницького питання та основних результатів аналізу. Приблизне наповнення презентації може бути таке:

- короткий опис у простих термінах дослідницького питання, у чому полягає його суть та цікавість, як воно змінилося з моменту захисту Лабораторної роботи 1;
- коротке нагадування про дані, які було використано для аналізу;
- опис результатів (таблиця з коефіцієнтами різних моделей, результати тестування гіпотез про статистичну значущість коефіцієнтів і груп коефіцієнтів тощо);
- відповідні висновки (чи було знайдено відповіді на питання, які саме, які додаткові питання виникли тощо).