**Stepik Course Part 3**

**Module 1 Обмеження при використанні лінійної регресії**

**Lecture 1 Лінійна регресія**

Методи порівняння (X - якісні, Y - кількісні):

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Кількість вибірок  (градація X) | | Дві вибірки | | Більше двох вибірок | |
|  | | Незалежні | Залежні | Незалежні | Залежні |
| Ознака Y | метричні | Параметричні методи порівняння | | | |
| t-Стюдента для незалежних вибірок | t-Стюдента для залежних вибірок | ANOVA | ANOVA з повторними вимірюваннями |
| рангові | Не параметричні методи порівняння | | | |
| U-Манна-Уітні, критерій серій | T-Вілкоксона, критерій знаків | H-Крускала-Уоллеса | x2-Фрідмана |

В реальності задачі порівняння **вирішуються методом** **регресії** (лінійної або логістичної).

На практиці більшість **обмежень**, накладених на лінійну регресію **НЕ виконуються**.

**Обмеження**, які накладаються **на лінійну регресію** (незалежна змінна Y і залежна X - кількісні):

* **Лінійний** взаємозв'язок змінних X і Y;
* **Незалежність** спостережень;
* **Незалежність** предикторів (перевірка на **мультиколінеарність**);
* **Гомоскедастичність**;
* **Нормальний** розподіл залишків;
* **Відсутність автокореляції** залишків;

**Lecture 2 Лінійність** **взаємозв'язку**

Для **перевірки лінійності** двох змінних, потрібно **нарисувати** точкову діаграму (**scatterplot**).

**Лінія регресії** – відображає максимально точно взаємозв’язок двох змінних, має формулу: ,

де b0 – intercept (де пряма пересікає вісь Y), b1 - slope (показує напрямок і кут нахилу (по знаку) до вісі X).

Якщо між двома ознаками **немає лінійного зв’язку**, то інтерпретація коефіцієнта b1 **не коректна** (b1 – показує на скільки зміниться значення залежної змінної, при одиничній зміні незалежної змінної, з врахуванням **лінійності** зв’язку).

Математичне обґрунтування b1:

, де

При одиничній зміні незалежної змінної (x), предиктор (y) зміниться на b1

При **відсутності лінійного зв’язку**, можна провести **трансформація змінних**, щоб зробити зв'язок більш лінійним.

**Трансформація Тюки** – трансформація незалежної, залежної або обох змінних, шляхом приведення змінної в степінь (від -∞ до +∞). При піднесенні до степенів зі значеннями нижче нуля, потрібно впевнитися, що не змінився знак залежності (при збільшенні x, росте y).

**Lecture 3 Логарифмічна трансформація змінних**

**Логарифмічна трансформація змінних** – трансформація незалежної, залежної або обох змінних, шляхом взяття натурального логарифма від них. На відмінку від піднесення до степеня, логарифмічна **трансформація дозволяє проінтерпретувати результати**. При задачі інтерпретації взаємозв’язку між змінними, а не передбачення, логарифмічна трансформація дає можливість **інтерпретувати природу взаємозв’язку**.

Математичне обґрунтування b1 при логарифмах змінних:

, де

Коефіцієнт b1 в моделі , показує наскільки процентів зміниться значення залежної змінної, при зміні предиктора на 1%.

Коефіцієнт b1 в моделі , показує, що при одиничній зміні змінної , змінна в середньому змінюється на відсотків.

Коефіцієнт b1 в моделі , показує, що зміна на 1% по в середньому призводить до зміни по змінній .

Часто при підозрі на не лінійність зв’язку між ознаками, не виконується вимога **нормальності розподілу залишків** –розподіл залишків буде значимо відрізнятись від нормально. Після трансформації Тюки або логарифмічної трансформації і приведення зв’язку до лінійного має виконуватись і вимога нормальності розподілу залишків.

**Трансформація Бокса-Кокса** (Box-Cox transformation) – метод трансформації даних. В регресії він використовується для трансформації залежної змінної у випадку, якщо у нас є ненормальний розподіл залишків і/або не лінійність взаємозв'язку, а також в разі гетероскедастичності:

, якщо

, якщо

Параметр підбирається для отримання найкращого результату.

**Lecture 4 Проблема гетероскедастичності**

**Гомоскедастичність** – незалежність дисперсії випадкових складових від номера спостереження (постійна мінливість залишків на всіх рівнях незалежної змінної).

Відповідність вимозі **гомоскедастичності** можна визначити, **нарисувавши графік залишків** (в python - seaborn.residplot).

Якщо побудувати регресію, де залежною змінною буде квадрат залишків моделі , а незалежною змінною буде предиктор , і в цій моделі виявиться високий і значимий R квадрат, це означає, що в даних є **гетероскедастичність**. Так працює **тест Уайта**.

Також існує **тест** **Бройша-Паґана** – в якості залежної змінно беруть залишки моделі , а незалежною змінною буде предиктор .

Для вирішення **проблеми гетероскедастичності**, варто **застосувати трансформацію даних** (квадрати або логарифми).

**Lecture 4 Мультиколінеарність**

**Мультиколінеарність** – наявність лінійного зв’язку між предикторами.

**Повна мультиколінеарність** – наявність лінійного зв’язку між предикторами з коефіцієнтом кореляції, який рівний 1 або -1. Може виникнути коли:

* Помилково зібрані **ті ж дані**, але у іншому виді;
* При створенні dummy variable, коли дві колонки показують одну і ту ж інформацію (наприклад колонка is\_man і колонка is\_woman, одну потрібно видалити);

Проблема мультиколінеарності виходить із **інтерпретації коефіцієнтів регресії**: одинична зміна незалежної ознаки приводить до зміни залежної ознаки на значення коефіцієнта біля незалежної, при умові, що **інші предиктори не міняються**. При високій кореляції між предикторами (мультиколінеарність) при зміні одного предиктора, інший також зміниться.

При мультиколінеарності **коефіцієнти** моделі можуть прийняти **не логічні значення** (**включаючи знак**) і виявитись **не значимими**, при чому сама **модель** (Adjusted R-squared) **буде значимою** (p-value).

Але при цьому **мультиколінеарність** **НЕ впливає на можливість передбачувати** дані моделлю.

Для перевірки на мультиколінеарність можна:

* побудувати **матрицю кореляцій**;
* розрахувати **VIF (variance inflation factor)** – **коефіцієнт інфляції дисперсії**, показує наскільки добре вибраний предиктор пояснюється іншими предиктороми;

, де – моделі регресії, де якості залежної змінної стоїть вибраний предиктор, а предиктори, які залишились, будуть незалежними змінними.

Якщо такий близький до 1, то це значить, що існує сильний лінійний зв'язок між даним предиктором і іншими незалежними змінними.

Квадратний корінь від показує, в скільки разів більша стандартна помилка даного коефіцієнту, у порівнянні із ситуацією, коли б даний предиктор був би абсолютно незалежний від інших предикторів.

Зазвичай коефіцієнт із потрібно видаляти з моделі.

Високе значення **НЕ впливає** на можливість **прогнозування** моделі, але **робить не інтерпретованими коефіцієнти** і гіпотези про взаємодію незалежної змінної з кожним окремим предиктором.

**Module 2 Змішані регресивні моделі**

**Lecture 5 Незалежності спостережень**

"Мудрість натовпу" - Френсіс Гальтон в 1906 році відвідав ринок, де проводилась лотерея для селян (800 чоловік), вони намагалися вгадати вагу бика, який важив 1198 фунтів. Ні один селянин не вгадав точного ваги, але якщо розрахувати середнє значення від їх прогнозів, то отримаємо 1197 фунтів.

Якщо б один і той же селянин кілька днів ходив на ринок, вгадував масу бика і вносив свій результат до загального, то в даних з’явилась би систематична помилка, порушилась вимога до незалежності спостережень.

Зазвичай в даних **порушується** вимога до **незалежності спостережень**, так як джерела даних в певній мірі містять зв’язки між собою.

**Джерела залежності** в даних:

1. **Повторні вимірювання** (наприклад одним і тим же пацієнтам спочатку дається плацебо, а потім ліки – для незалежності спостережень потрібно знайти окремі групи для плацебо і для ліків);
2. **Повторні проби** (наприклад одному і тому ж пацієнту даються ліки кілька разів);
3. **Кластеризація даних** (наприклад ліки тестуються на даних по пацієнтах платної безплатної клінік, варто очікувати схожих результатів від пацієнтів із клінік одного типу);

Проблеми, які виникають в результаті порушення вимоги до незалежності спостережень:

1. **Повторні вимірювання**:

* **штучне збільшення сили тесту** (**псевдореплікація**) (наприклад на 5 пацієнтах проведено по 20 тестувань препарату, по факту буде тільки 5 незалежних спостережень, так як ми багато дізнались як конкретний пацієнт реагує на ліки, але мало про те, як працюють ліки);
* **знижень чутливості тесту** (при побудові довірчих інтервалів і порівнянні середніх вибірок, при наявності повторних вимірювань, потрібно спочатку взяти середнє всередині незалежних спостережень (середнє по кожному пацієнту), а потім порівнювати отримані значення);

1. **Незбалансований дизайн** – на одному рівні незалежної змінної більше даних, ніж на іншому (наприклад давали пацієнту більше плацебо, ніж ліків; інший приклад – пошук інформації про рідкісні хвороби чи події, тоді інформації про здорових багато, а про хворих мало):

* **не коректні дані**;

1. **Кластеризація даних**:

* **не коректні дані**;

**Lecture 6 Змішані регресивні моделі**

**Ефект** – вплив незалежної змінної (предиктор) на залежну.

**Фіксований ефект** (головний) – вплив незалежної змінної, яка представляє найбільший інтерес (інформація, яка спеціально збирається або вимірюється – вплив ліків на пацієнта і т.д.).

**Випадковий ефект** – вплив незалежної змінної, яка НЕ представляє першочергового інтересу (інформація, яка в ДАНОМУ тесті не вважалась важливою, але можуть такими буде – ріст, вік, стать пацієнта і т.д.).

**Лінійна** модель з **фіксованими ефектами**:

, де – незалежні змінні (фіксовані ефекти), – коефіцієнти для фіксованих ефектів, які ми підбираємо (, і т.д.), – шум (частка дисперсії, яку ми не можемо пояснити).

**Лінійна** модель з **фіксованими** і **випадковими ефектами**:

, де – випадкові ефекти, – коефіцієнти для випадкових ефектів, які ми підбираємо (, і т.д.).

**Lecture 7 Статистична значимість, узагальнені моделі і випадкові ефекти**

Проблема оцінки p-значення предиктора в змішаній моделі виникає через те, що в модель включені випадкові ефекти, тому неможливо точно оцінити кількість степенів свободи для предикторів у моделі.

**Module 3 Введення в bootstrap**

**Lecture 8 Складаний ніж (jackknife)**

**Мат.очікування** () і **середнє** () не являються одним і тим же. Середнє дорівнює **оцінці мат.очікування** ().

, де – ймовірність події (певного значення), – діапазон значень, які може приймати , в діапазоні

, де – розмір вибірки.

**Jackknife** – метод виправлення зміщення певної оцінки (наприклад ми хочемо визначити максимальний розмір таргана, наївний оцінювач буде ловити тарганів (робити вибірки) і брати таргана максимального розміру, тому оцінка наївного оцінювача буде занижена. Реальний максимальний розмір таргана буде більшим. Щоб наївний оцінювач добре оцінив максимальний розмір, йому потрібна велика кількість вибірок).

Суть методу: для кожного елемента обчислюється середнє значення вибірки без врахування даного елемента, а потім - середнє всіх таких значень. Для вибірки з N елементів оцінка визначається шляхом обчислення середнього значення інших N-1 елементів.

**Lecture 8 Bootstrap**

**Bootstrap** – універсальний метод оцінки дисперсії і довірчих інтервалів. Не дає надійних результатів при занадто малих вибірках.

Приклад: Розрахувати 90% довірчий інтервал для медіани за допомогою basic bootstrap. У нас є вибірка, для вирішення поставленого завдання потрібен значення вибіркової медіани. Далі необхідно сформувати безліч бутсрапірованних вибірок з повтореннями із нашої вихідної вибірки. Тепер для кожної вибірки розрахуємо, наскільки медіана в даній вибірці відхилилася від значення, розрахованого в найпершому кроці. Залишилося тільки розрахувати 5 і 95 процентиль отриманого розподілу і добавити до них значення вибіркової медіани.