

**НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»
Інститут прикладного системного аналізу**

ЛАБОРАТОРНА РОБОТА

з дисципліни «Моделювання Складних Систем»

**на тему: «Регресійний аналіз за неоднорідних та корельованих
спостережень»**

Виконали:

**Панасенко, Жутник, Наход, Задорожний,
Трач, Ткаченко, Дубик, Скоропадський,
Висоцький, Васильківський**

**Перевірила: проф., д.т.н. Чумаченко
О.І.**

МЕТА

Дослідити і оцінити регресійну модель для даної вибірки, використовуючи різні методи оцінки коефіцієнтів регресії та провести статистичний аналіз отриманих результатів.

ПОСТАНОВКА ЗАДАЧІ

1. Перевірити вибірку на однорідність (критерій Бартлета, Кохрена, модифікація Мкритерію Бартлета, граничний критерій неоднорідності дисперсій Харисона і Ман-Кейба, параметричний критерій Голдфільда і Квандта) 2. Оцінити коефіцієнти регресії за допомогою узагальненого методу найменших квадратів. 3. Оцінити коефіцієнти регресії за допомогою методу найменших квадратів. 4. Виконати статистичний аналіз результатів оцінювання. 5. Оцінити коефіцієнти регресії першого порядку. 6. Здійснити оцінювання узагальненої регресійної моделі.

ТЕОРЕТИЧНІ ВІДОМОСТІ

Критерій Бартлета - це статистичний тест, який використовується для перевірки гіпотези про однорідність дисперсій у групах даних. Цей критерій чутливий до нормального розподілу даних.

Його головна ідея полягає в порівнянні дисперсій між групами даних. Якщо дисперсії однорідні (тобто різниці між групами нестатистично значущі), то значення критерію буде незначним, що свідчить про прийняття нульової гіпотези про однорідність дисперсій. В іншому випадку, якщо дисперсії статистично відрізняються, значення критерію буде великим, що означатиме відхилення від нульової гіпотези.

Формально, для застосування критерію Бартлета ви можете використовувати наступні кроки:

1. Сформулювати нульову та альтернативну гіпотези. Нульова гіпотеза припускає, що дисперсії всіх груп даних однорідні, тоді як альтернативна гіпотеза стверджує, що дисперсії відрізняються хоча б у двох групах.
2. Обчислити значення критерію Бартлета.
3. Порівняти обчислене значення критерію з критичним значенням з відповідної таблиці для визначення статистичної значущості результату.

4. Зробити висновок про однорідність або неоднорідність дисперсій в групах даних на основі порівняння обчисленого значення критерію та його критичного значення.

Зокрема, для застосування критерію Бартлета важливо мати рівність кількості спостережень у різних групах та нормальний розподіл даних.

Критерій Кохрена - це статистичний тест, який використовується для перевірки однорідності дисперсій у групах даних. Він також відомий як критерій однорідності дисперсій.

Основна ідея критерію полягає у порівнянні максимальної та мінімальної дисперсій між групами. Якщо ці значення статистично не відрізняються, то припускається, що дисперсії у всіх групах однорідні.

Процедура застосування критерію Кохрена така:

1. Сформулювати нульову та альтернативну гіпотези. Нульова гіпотеза припускає, що дисперсії у всіх групах однорідні, альтернативна гіпотеза - що є хоча б одна відмінність між дисперсіями.
2. Обчислити максимальну та мінімальну дисперсії в групах.
3. Обчислити вибірккову дисперсію по кожній групі.
4. Обчислити статистику Кохрена шляхом ділення максимальної вибіркової дисперсії на суму всіх вибіркових дисперсій.
5. Порівняти обчислене значення статистики Кохрена з критичним значенням з таблиці для визначення статистичної значущості результату.
6. Зробити висновок щодо однорідності або неоднорідності дисперсій в групах на основі порівняння обчисленого значення статистики Кохрена та його критичного значення.

Критерій Кохрена є досить простим у застосуванні та може бути використаний для перевірки однорідності дисперсій у випадку, коли кількість груп не більше деякого певного обмеження, зазвичай до десяти. Модифікація М-критерію Бартлета - це адаптація критерію Бартлета для випадку, коли дані мають нормальний розподіл, але не мають однорідності дисперсій. Цей критерій використовується для порівняння дисперсій у групах даних, коли вибірки не мають однакових дисперсій.

Процедура застосування модифікації М-критерію Бартлета аналогічна до звичайного критерію Бартлета, але перед застосуванням необхідно виконати попередню нормалізацію даних. Після цього кроки такі:

1. Сформулювати нульову та альтернативну гіпотези. Нульова гіпотеза стверджує, що всі групи мають однакову дисперсію, альтернативна гіпотеза - що є хоча б одна відмінність між дисперсіями.
2. Виконати нормалізацію даних (якщо необхідно).
3. Обчислити вибірккові дисперсії по кожній групі.
4. Обчислити середню вибірккову дисперсію.
5. Обчислити статистику М-критерію Бартлета, яка базується на порівнянні середньої вибіркової дисперсії з максимальною вибірковою дисперсією.
6. Порівняти обчислене значення статистики М-критерію Бартлета з критичним значенням з таблиці для визначення статистичної значущості результату.
7. Зробити висновок щодо однорідності або неоднорідності дисперсій в групах на основі порівняння обчисленого значення статистики М-критерію Бартлета та його критичного значення.

Цей підхід дозволяє враховувати нормальний розподіл даних та вирішує проблему нерівномірності дисперсій у групах. Граничний критерій неоднорідності дисперсій Харісона і Манна-Кейба (Hartley's and Mann-Whitney's heteroscedasticity test) - це статистичний тест, який використовується для оцінки однорідності дисперсій у групах даних. Він оцінює максимальне відношення дисперсій між групами.

Процедура застосування цього критерію наступна:

1. Сформулювати нульову та альтернативну гіпотези. Нульова гіпотеза припускає, що всі групи мають однакову дисперсію, тоді як альтернативна гіпотеза стверджує, що є хоча б одна відмінність між дисперсіями.
2. Обчислити вибірккові дисперсії по кожній групі.
3. Знайти максимальне і мінімальне значення вибіркових дисперсій.
4. Обчислити відношення максимальної до мінімальної вибіркової дисперсії.
5. Знайти критичне значення граничного відношення дисперсій відповідно до обраного рівня значущості та кількості груп.
6. Порівняти обчислене відношення з критичним значенням.

7. Зробити висновок щодо однорідності або неоднорідності дисперсій в групах на основі порівняння обчисленого відношення з критичним значенням.

Цей критерій виявляється корисним в тих випадках, коли потрібно оцінити неоднорідність дисперсій у групах, особливо коли розподіл даних може бути важко інтерпретувати за допомогою інших методів. Параметричний критерій Голдфільда і Квандта - це статистичний тест, який використовується для перевірки гіпотези про однорідність дисперсій у групах даних. Цей критерій передбачає, що дані мають нормальний розподіл.

Процедура застосування критерію Голдфільда і Квандта наступна:

1. Сформулювати нульову та альтернативну гіпотези. Нульова гіпотеза припускає, що дисперсії в усіх групах однакові, тоді як альтернативна гіпотеза стверджує, що є хоча б одна відмінність між дисперсіями.
2. Обчислити вибіркові дисперсії по кожній групі.
3. Впорядкувати вибіркові дисперсії за зростанням.
4. Визначити кількість спостережень у кожній групі, включаючи всі групи.
5. Знайти критичне значення для потрібного рівня значущості та кількості груп у таблиці критичних значень для критерію Голдфільда і Квандта.
6. Порівняти обчислене критерієм значення тесту з критичним значенням.
7. Зробити висновок щодо однорідності або неоднорідності дисперсій в групах на основі порівняння обчисленого значення критерію з критичним.

Цей критерій є корисним у випадках, коли потрібно перевірити однорідність дисперсій у групах і дані відповідають умовам нормального розподілу. Він може бути особливо корисним, коли кількість груп велика і групи мають різний обсяг. Узагальнений метод найменших квадратів (УМНК) - це метод оцінювання параметрів регресійної моделі, який використовується у випадку, коли відомо, що похибки моделі не мають гауссівського розподілу. Цей метод є розширенням класичного методу найменших квадратів (МНК), який передбачає нормальний розподіл похибок.

Припустимо, що ми маємо регресійну модель у вигляді:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + \varepsilon_i$$

Основна ідея узагальненого методу найменших квадратів полягає в мінімізації суми квадратів вагованих відхилень між спостережуваними значеннями і значеннями, передбаченими моделлю. Ваги при цьому встановлюються відповідно до властивостей похибок.

Процедура оцінювання параметрів за допомогою УМНК включає наступні кроки:

1. Сформулювати регресійну модель і припущення про похибки моделі.
2. Обрати ваги для мінімізації функції втрат (зазвичай ваги вибираються залежно від природи похибок).
3. Мінімізувати суму квадратів вагованих відхилень між спостережуваними значеннями і значеннями, передбаченими моделлю.
4. Оцінити параметри моделі, що мінімізують цю суму.

Оцінки параметрів, отримані за допомогою УМНК, зазвичай є менш ефективними порівняно з оцінками, отриманими за допомогою МНК у випадку, коли похибки мають нормальний розподіл. Однак УМНК залишається корисним інструментом у випадках, коли це припущення про розподіл похибок не виконується. Метод найменших квадратів (МНК) - це стандартний метод оцінювання параметрів регресійної моделі, який полягає в мінімізації суми квадратів відхилень між спостережуваними значеннями залежної змінної і значеннями, передбаченими моделлю.

Оцінки, отримані за допомогою МНК, надають найкращі значення параметрів для регресійної моделі в тому сенсі, що вони мінімізують суму квадратів відхилень. Цей метод є широко використовуваним і дозволяє отримати найкращі оцінки параметрів, якщо припущення про нормальний розподіл похибок виконується.

Для виконання статистичного аналізу результатів оцінювання параметрів регресії за допомогою методу найменших квадратів, можна виконати наступні кроки:

1. **Перевірка припущень моделі:** Перш ніж переходити до аналізу результатів, слід перевірити припущення, які лежать в основі регресійної моделі, такі як нормальність розподілу похибок, лінійність зв'язку між змінними тощо. Для цього можна використовувати графічні методи (наприклад, Q-Q графіки, графіки залишків), а також статистичні тести (наприклад, тест Шапіро-Уїлка для нормальності розподілу похибок).

2. **Оцінка значущості параметрів моделі:** Для оцінки статистичної значущості кожного з параметрів регресії можна використовувати t-тест. Порівнюючи оцінки коефіцієнтів регресії з їх стандартними помилками, можна обчислити t-статистики та визначити статистичну значущість параметрів на певному рівні значущості.
3. **Оцінка адекватності моделі:** Для оцінки того, наскільки добре побудована модель пояснює варіацію в залежній змінній, можна використовувати різноманітні статистичні метрики, такі як коефіцієнт детермінації (R^2), кореляційний коефіцієнт Пірсона, а також F-тест на значущість моделі в цілому.
4. **Аналіз залишків моделі:** Перевірка залишків моделі є важливим етапом аналізу. Графіки залишків (наприклад, розподіл залишків, графік "залишків-прогнозованих значень") можуть допомогти виявити систематичні відхилення в моделі. Додатково можна застосувати тести на автокореляцію залишків, якщо це відповідає контексту дослідження.
5. **Перевірка мультиколінеарності:** Якщо в моделі присутня мультиколінеарність (високий рівень кореляції між незалежними змінними), це може спричинити проблеми при оцінці параметрів. Для виявлення мультиколінеарності можна використовувати статистичні тести, такі як визначення факторного числа, або обчислення кореляційної матриці між змінними.
6. **Інтерпретація результатів:** Нарешті, після виконання усіх статистичних аналізів, слід зробити висновки щодо статистичної значущості та адекватності моделі, а також провести інтерпретацію кожного з коефіцієнтів регресії в контексті дослідження.

Ці кроки допоможуть вам здійснити повний та об'єктивний статистичний аналіз результатів оцінювання параметрів регресії за допомогою методу найменших квадратів.

Після обчислення цих оцінок можна перевірити їх статистичну значимість за допомогою t-тесту. Також можна визначити коефіцієнт детермінації (R^2) для оцінки якості підгонки моделі до даних.

Узагальнена регресійна модель використовується, коли дані не відповідають класичним умовам простої лінійної регресії. Це може включати випадки, коли залежна змінна не має нормального розподілу, коли похибки мають

гетероскедастичну або асиметричну структуру, або коли залежність між змінними не лінійна.

Для оцінки узагальненої регресійної моделі можна використовувати методи, такі як метод найменших квадратів з вагами, робустні методи оцінювання (наприклад, регресія Хубера), або методи, які враховують структуру похибок (наприклад, гребнева регресія або ласо).

Основні кроки оцінювання узагальненої регресійної моделі:

1. **Вибір моделі:** Спочатку необхідно визначити структуру моделі, включаючи вибір залежних і незалежних змінних, а також вибір форми функціонального відношення між ними.
2. **Визначення функції ваг:** У випадках, коли похибки мають гетероскедастичну або асиметричну структуру, можна використовувати функції ваг, які дають більший вагу менш варіативним спостереженням.
3. **Оцінка параметрів моделі:** Після визначення моделі та функції ваг можна застосувати методи оцінювання, такі як МНК з вагами або робустні методи, для отримання оцінок параметрів моделі.
4. **Оцінка статистичної значимості параметрів:** Після отримання оцінок параметрів можна провести тести на їх статистичну значимість, такі як t-тест або тести Вальда.
5. **Оцінка адекватності моделі:** Для оцінки адекватності моделі можна використовувати різноманітні статистичні тести, такі як тест на гомоскедастичність похибок, тест на адекватність підгонки моделі тощо.
6. **Перевірка припущень:** Нарешті, необхідно перевірити, чи виконуються припущення, на яких ґрунтується модель, та вирішити, чи вони задовольняються.

Ці кроки допоможуть здійснити оцінювання узагальненої регресійної моделі та зробити висновки щодо її адекватності та значущості.

ОПИС ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

1. Вхідні дані

кл кроватьей	площа	км до центру	км до метро	поверх	вік	ціна
-----------------	-------	-----------------	----------------	--------	-----	------

1	26	1300	200	22	15	100000
---	----	------	-----	----	----	--------

2. Використане програмне забезпечення

R lang psych lsr

3. Результати

формула	тест Бартлета			Кохен
площа*км до центру*км до метро*повер х*вік*ціна	К квадрат	df	значення p	0.76
	2043	13	2.2E-16	

коефіцієнти		коефіцієнти	
intercept	кл кроват	intercept	кл кроват
6.28E+14	-2.95E+13	2.62E+13	5.87E+12

квадрат діаг		Харісон			
2.62E+13	5.87E+12	стат	зн П	парам	альт
		0.541	0.473	0.5	дуже мало

1. Перевірити вибірку на однорідність (критерій Бартлета, Кохрена, модифікація Мкритерію Бартлета, граничний критерій неоднорідності дисперсій Харисона і Ман-Кейба, параметричний критерій Голдфільда і Квандта)

вибірка є однорідною

2. Оцінити коефіцієнти регресії за допомогою узагальненого методу найменших квадратів.

Коефіцієнти є досить непоганими

3. Оцінити коефіцієнти регресії за допомогою методу найменших квадратів.

також доводить що модель досить непогана

4. Виконати статистичний аналіз результатів оцінювання.

Статистика теж досить непогана

5. Оцінити коефіцієнти регресії першого порядку.

Дуже малі але це очікувано

6. Здійснити оцінювання узагальненої регресійної моделі.

В цілому модель непогано себе показала

ВИСНОВКИ

Було поставлено мету. Виконано задачі. Вивчена теорія. Написаний код. Зроблена модель.