Trần Minh Tâm

CÁC MÔ HÌNH KINH TẾ LƯỢNG HIỆN ĐẠI

TP. Hồ Chí Minh - 2025

Mục lục

Phần I

Tổng quan về kinh tế lượng và Python

Chương 1

Giới thiệu kinh tế lượng và vai trò của Python

1.1 Kinh tế lượng là gì?

Kinh tế lượng là lĩnh vực kết hợp giữa kinh tế học, thống kê và toán học để phân tích dữ liệu kinh tế. Kinh tế lượng đóng vai trò quan trọng trong nghiên cứu dữ liệu Wooldridge (2019). Phương pháp thực nghiệm trong kinh tế lượng được đề xuất trong Angrist and Pischke (2009).

1.2 Tại sao sử dụng Python trong kinh tế lượng?

Python ngày càng phổ biến trong nghiên cứu kinh tế lượng vì cú pháp dễ hiểu và hệ sinh thái phong phú.

1.3 Các thư viện quan trọng

Dưới đây là một số thư viện quan trọng trong Python cho kinh tế lượng:

- **NumPy**: Xử lý ma trận và tính toán số học.
- **Pandas**: Xử lý dữ liệu dạng bảng.
- **Statsmodels**: Thực hiện các mô hình kinh tế lượng.

1.4 Cài đặt môi trường làm việc

1.4.1 Cài đặt Python và Anaconda

Python là một ngôn ngữ lập trình mạnh mẽ và dễ dàng sử dụng trong kinh tế lượng **?**. Để cài đặt Python, ta nên sử dụng Anaconda, một phần phối chứa sẵn nhiều thư viện Python hữu ích cho phân tích dữ liệu và kinh tế lượng **?**.

Các bước cài đặt Anaconda

1. Truy cập trang chủ Anaconda: `https://www.anaconda.com/`

2. Tải và cài đặt phiên bản phù hợp với hệ điều hành.
3. Kiểm tra cài đặt bằng lệnh:

```
conda --version
python --version
```

1.4.2 Thiết lập môi trường làm việc bằng Jupyter Notebook

Jupyter Notebook là công cụ hữu ích cho phân tích dữ liệu, lập trình và trình bày kết quả khoa học **?**. **Cài đặt Jupyter Notebook** Sau khi cài đặt Anaconda, có thể chạy Jupyter Notebook bằng lệnh:

```
jupyter notebook
```

Hoặc cài đặt riêng:

```
pip install notebook
```

1.4.3 Hướng dẫn cài đặt thư viện

Các thư viện quan trọng trong kinh tế lượng bao gồm:

- **numpy**: Toán học ma trận.
- **pandas**: Xử lý dữ liệu dạng bảng.
- **statsmodels**: Hồi quy và phân tích dữ liệu.
- **scipy**: Các công cụ toán học.
- **matplotlib**: Vẽ đồ thị.
- **seaborn**: Trực quan hóa dữ liệu.

Chương 2

Xử lý dữ liệu trong kinh tế lượng

2.1 Tổng quan về dữ liệu

2.1.1 Khái niệm

2.1.2 Phân loại dữ liệu

2.1.3 Dữ liệu trong kinh tế lượng hiện đại

- **Dữ liệu chéo (Cross Sectional Data)**
- **Dữ liệu chuỗi thời gian (Time Series Data)**
- **Dữ liệu chéo gộp (Pooled Cross Sectional Data)**
- **Dữ liệu bảng (Panel Data)**
- **Dữ liệu không gian (Spatial Data)**
- **Dữ liệu tần số cao (High-Frequency Data)**
- **Dữ liệu văn bản (Text Data)**

2.2 Các phương pháp đo lường dữ liệu

2.2.1 Đo lường mức độ tập trung

a. Trung bình (Mean)

Định nghĩa: Trung bình (Mean) là một đại lượng đo lường xu hướng trung tâm của dữ liệu. Nó cho biết giá trị đại diện của một tập hợp dữ liệu bằng cách lấy tổng tất cả các giá trị chia cho số lượng phần tử.

Công thức tính trung bình:

- **Trung bình số học (Arithmetic Mean)**

  Trung bình số học của một tập dữ liệu gồm n quan sát $x_1, x_2, \ldots, x_n$ được tính bằng công thức:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2.1}$$

  Ví dụ: Tập dữ liệu: $\{10, 20, 30, 40, 50\}$

$$\bar{x} = \frac{10 + 20 + 30 + 40 + 50}{5} = 30 \tag{2.2}$$

- **Trung bình có trọng số (Weighted Mean)**

  Nếu mỗi giá trị x_i có trọng số tương ứng w_i, thì trung bình có trọng số được tính bằng:

$$\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i} \tag{2.3}$$

  Ví dụ: Điểm của sinh viên:

  – Toán (trọng số 4, điểm 8)

  – Lý (trọng số 3, điểm 7)

  – Hóa (trọng số 2, điểm 6)

$$\bar{x}_w = \frac{(4 \times 8) + (3 \times 7) + (2 \times 6)}{4 + 3 + 2} = \frac{32 + 21 + 12}{9} = \frac{65}{9} \approx 7.22 \tag{2.4}$$

- **Trung bình hình học (Geometric Mean)**

  Dùng khi dữ liệu có dạng tăng trưởng theo cấp số nhân:

$$GM = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}} = (x_1 \times x_2 \times \cdots \times x_n)^{\frac{1}{n}} \tag{2.5}$$

  Ví dụ: Tăng trưởng doanh thu qua 3 năm là 5%, 10%, 15%, trung bình hình học là:

$$GM = (1.05 \times 1.10 \times 1.15)^{\frac{1}{3}} \approx 1.096 \tag{2.6}$$

  Tức là mức tăng trung bình mỗi năm khoảng 9.6%.

- **Trung bình điều hòa (Harmonic Mean)**

  Dùng khi dữ liệu là tốc độ hoặc tỷ lệ:

$$HM = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \tag{2.7}$$

  Ví dụ: Nếu một ô tô đi 60 km/h trong 1 giờ và 40 km/h trong 1 giờ:

$$HM = \frac{2}{\frac{1}{60} + \frac{1}{40}} = \frac{2}{\frac{2}{120} + \frac{3}{120}} = \frac{2}{\frac{5}{120}} = \frac{240}{5} = 48 \text{ km/h} \tag{2.8}$$

b. Trung vị (Median)

Định nghĩa

Trung vị (Median) là giá trị nằm ở giữa một tập hợp dữ liệu đã được sắp xếp theo thứ tự tăng dần hoặc giảm dần. Nó chia tập dữ liệu thành hai phần bằng nhau: 50% giá trị nhỏ hơn trung vị và 50% giá trị lớn hơn trung vị.

-Ưu điểm:

- Ít bị ảnh hưởng bởi giá trị ngoại lai (outliers).
- Phù hợp khi dữ liệu có phân phối lệch.

Cách tính trung vị

*** Dữ liệu rời rạc**

- Nếu số lượng quan sát N là số lẻ:

$$\tilde{x} = X_{\frac{N+1}{2}}$$

- Nếu số lượng quan sát N là số chẵn:

$$\tilde{x} = \frac{X_{\frac{N}{2}} + X_{\frac{N}{2}+1}}{2}$$

*** Dữ liệu nhóm (có bảng tần số)**

Trung vị được tính theo công thức:

$$\tilde{x} = L + \frac{\frac{N}{2} - F}{f} \times h$$

Trong đó:

- L: Cận dưới của lớp chứa trung vị.
- N: Tổng số quan sát.
- F: Tần số tích lũy trước lớp chứa trung vị.
- f: Tần số của lớp chứa trung vị.
- h: Độ rộng lớp chứa trung vị.

*** Dữ liệu phân phối liên tục (sử dụng CDF)**

Trung vị là giá trị x sao cho:

$$F(\tilde{x}) = 0.5$$

Tức là điểm mà 50% dữ liệu nằm dưới nó trong hàm phân phối tích lũy.

Khi nào nên dùng trung vị thay vì trung bình?

- Khi dữ liệu có ngoại lai, trung vị ít bị ảnh hưởng hơn.
- Khi dữ liệu có phân phối lệch, trung vị thể hiện xu hướng trung tâm tốt hơn.
- Khi dữ liệu có dạng phân phối log-normal, chẳng hạn như bất động sản, thu nhập, giá cổ phiếu, v.v.

c. Mode

Định nghĩa: Mode (Yếu vị) là giá trị xuất hiện nhiều nhất trong một tập dữ liệu. Đây là một trong ba thước đo xu hướng trung tâm chính, bên cạnh Mean (trung bình) và Median (trung vị).

- Nếu một tập dữ liệu có một giá trị xuất hiện nhiều nhất, nó được gọi là **unimodal** (đơn mode).
- Nếu có hai giá trị cùng xuất hiện với tần suất cao nhất, tập dữ liệu được gọi là **bimodal** (hai mode).
- Nếu có nhiều hơn hai giá trị xuất hiện với tần suất cao nhất, tập dữ liệu được gọi là **multimodal** (đa mode).

*** Công thức xác định mode:** Mode không có công thức cố định như mean hay median. Nó đơn giản là giá trị có tần suất xuất hiện cao nhất trong tập dữ liệu.

Ví dụ:

- Dữ liệu: {2, 3, 5, 3, 3, 6, 7, 2, 2, 3}
- Mode = **3** (vì số 3 xuất hiện 4 lần, nhiều nhất trong tập dữ liệu).

*** Cách xác định mode trong phân bố tần suất** Với dữ liệu nhóm trong bảng tần suất, mode có thể được ước lượng bằng công thức:

$$\text{Mode} = L + \frac{(f_1 - f_0)}{(2f_1 - f_0 - f_2)} \times h \tag{2.9}$$

Trong đó:

- L là cận dưới của lớp có tần suất cao nhất (lớp modal),
- f_1 là tần suất của lớp modal,
- f_0 là tần suất của lớp trước lớp modal,
- f_2 là tần suất của lớp sau lớp modal,
- h là độ rộng của lớp.

Ví dụ: Nếu có bảng tần suất như sau:

Khoảng lớp	Tần suất
10 - 20	5
20 - 30	8
30 - 40	12
40 - 50	9
50 - 60	6

- Lớp có tần suất cao nhất là **30 - 40** với $f_1 = 12$, - $L = 30$, $f_0 = 8$, $f_2 = 9$, - $h = 10$.

Áp dụng công thức:

$$\text{Mode} = 30 + \frac{(12-8)}{(2\times 12 - 8 - 9)} \times 10 = 30 + \frac{4}{7} \times 10 = 30 + 5.71 = 35.71 \quad (2.10)$$

*** Cách xác định Mode bằng đạo hàm**

Mode của một phân bố liên tục là giá trị x sao cho hàm mật độ xác suất $f(x)$ đạt cực đại, tức là:

**** Bước 1: Lấy đạo hàm bậc nhất**

Lấy đạo hàm bậc nhất của $f(x)$ và giải phương trình:

$$f'(x) = 0 \quad (2.11)$$

Đây là điều kiện cần để tìm điểm cực trị.

**** Bước 2: Kiểm tra đạo hàm bậc hai**

- Nếu $f''(x) < 0$, thì x là điểm cực đại và là Mode.
- Nếu $f''(x) > 0$, thì x là điểm cực tiểu (không phải Mode).

=> Ví dụ: Mode của phân bố chuẩn

Xét phân bố chuẩn $N(\mu, \sigma^2)$ có hàm mật độ xác suất:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (2.12)$$

- Bước 1: Lấy đạo hàm

$$f'(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \cdot \left(-\frac{2(x-\mu)}{2\sigma^2}\right) \tag{2.13}$$

$$= f(x) \cdot \left(-\frac{x-\mu}{\sigma^2}\right) \tag{2.14}$$

Đặt $f'(x) = 0$, ta có $x - \mu = 0$ hay Mode $= \mu$.

- Bước 2: Kiểm tra đạo hàm bậc hai

$$f''(x) = f(x) \cdot \left(-\frac{1}{\sigma^2}\right) + f(x) \cdot \left(-\frac{x-\mu}{\sigma^2}\right)^2 \tag{2.15}$$

Tại $x = \mu$, ta có $f''(x) < 0$, suy ra đây là điểm cực đại.

**** Kết luận:**

Mode của phân bố chuẩn chính là trung bình μ.

Đặc điểm của mode:

- Mode có thể không tồn tại hoặc có nhiều hơn một giá trị trong tập dữ liệu.
- Mode có thể bị ảnh hưởng bởi sự thay đổi nhỏ trong tần suất của dữ liệu.
- Đối với dữ liệu định tính (categorical data), mode là thước đo trung tâm phù hợp nhất.

Ví dụ:

- Dữ liệu màu sắc yêu thích của 100 người: {Đỏ, Xanh, Xanh, Xanh, Đỏ, Đỏ, Đỏ, Xanh, Xanh, Đỏ, Đỏ, Xanh}
- Mode = **"Xanh"** (vì xuất hiện nhiều nhất).

Thuộc tính	**Mode**	**Mean (Trung bình)**	**Median (Trung vị)**
Định nghĩa	Giá trị xuất hiện nhiều nhất	Trung bình số học của tất cả giá trị	Giá trị chính giữa của tập dữ liệu
Khi nào dùng?	Khi dữ liệu có giá trị lặp lại hoặc là dữ liệu định tính	Khi dữ liệu phân bố đều, không bị lệch	Khi dữ liệu có giá trị ngoại lai
Bị ảnh hưởng bởi ngoại lai?	Không	Có	Ít bị ảnh hưởng

Bảng 2.1: So sánh Mode với Mean và Median

Ứng Dụng của Mode

- **Thống kê kinh doanh**: Xác định sản phẩm bán chạy nhất.
- **Giáo dục**: Xác định điểm số phổ biến nhất trong lớp học.
- **Tiếp thị**: Tìm màu sắc, kích cỡ hoặc mẫu mã sản phẩm được ưa chuộng nhất.

Mode là một thước đo quan trọng trong thống kê, giúp hiểu rõ hơn về xu hướng trung tâm của dữ liệu. Trong nhiều trường hợp, nó là công cụ hữu ích hơn mean và median, đặc biệt đối với dữ liệu phân loại hoặc dữ liệu có phân phối không chuẩn.

2.2.2 Đo lường mức độ phân tán

a. Tứ phân vị (Quartiles)

Tứ phân vị là các giá trị chia một tập dữ liệu đã được sắp xếp thành bốn phần bằng nhau. Các giá trị này giúp chúng ta hiểu rõ hơn về sự phân bố của dữ liệu.

*** Các loại tứ phân vị**

- **Tứ phân vị thứ nhất** (Q_1): Là phần tử nằm ở vị trí 25% của dữ liệu đã sắp xếp. Đây là trung vị của nửa dưới của dữ liệu.
- **Tứ phân vị thứ hai** (Q_2): Là trung vị (median) của toàn bộ dữ liệu, chia dữ liệu thành hai phần bằng nhau (50%).
- **Tứ phân vị thứ ba** (Q_3): Là phần tử nằm ở vị trí 75% của dữ liệu. Đây là trung vị của nửa trên của dữ liệu.

*** Cách tính tứ phân vị**

1. Sắp xếp dữ liệu theo thứ tự tăng dần.
2. Tìm Q_2 (trung vị của toàn bộ dữ liệu).
3. Tìm Q_1 (trung vị của nửa dưới) và Q_3 (trung vị của nửa trên).

Ví dụ

Xét dãy số:

$$2, 4, 7, 10, 12, 15, 18, 22, 25, 30$$

- Q_2 **(Median)**: Trung vị là giá trị nằm giữa dãy số. Ở đây có 10 số, trung vị là:

$$Q_2 = \frac{12 + 15}{2} = 13.5$$

- Q_1 **(Tứ phân vị thứ nhất)**: Trung vị của nửa dưới:

$$Q_1 = \frac{4+7}{2} = 5.5$$

- Q_3 **(Tứ phân vị thứ ba)**: Trung vị của nửa trên:

$$Q_3 = \frac{22+25}{2} = 23.5$$

*** Ý nghĩa của tứ phân vị**

- Giúp xác định vị trí trung tâm và mức độ phân tán của dữ liệu.
- Dùng để tính khoảng biến thiên liên tứ phân vị (IQR) nhằm đo lường độ phân tán.

b. Khoảng biến thiên liên tứ phân vị (IQR - Interquartile Range)

Khoảng biến thiên liên tứ phân vị (IQR) đo độ phân tán của 50% dữ liệu trung tâm bằng cách tính hiệu giữa tứ phân vị thứ ba và tứ phân vị thứ nhất.

Công thức tính IQR

$$IQR = Q_3 - Q_1 \tag{2.16}$$

Trong đó:

- Q_1 là tứ phân vị thứ nhất (25%).
- Q_3 là tứ phân vị thứ ba (75%).

Ví dụ

Từ ví dụ trước với:

$$Q_1 = 5.5,$$
$$Q_3 = 23.5$$

Ta có:

$$IQR = 23.5 - 5.5 = 18 \tag{2.17}$$

→ Khoảng 50% dữ liệu trung tâm nằm trong khoảng từ 5.5 đến 23.5.

*** Ý nghĩa của IQR**

- ✓ Không bị ảnh hưởng bởi ngoại lệ, vì chỉ xét khoảng giữa 50% dữ liệu.
- ✓ Giúp phát hiện giá trị ngoại lệ, dựa vào ngưỡng ngoài:

Giới hạn dưới và giới hạn trên

$$\text{Giới hạn dưới} = Q_1 - 1.5 \times IQR \tag{2.18}$$

$$\text{Giới hạn trên} = Q_3 + 1.5 \times IQR \tag{2.19}$$

Nếu một điểm dữ liệu nằm ngoài khoảng này, nó có thể là ngoại lệ.

*** Ví dụ về phát hiện ngoại lệ**

Với $Q_1 = 5.5$, $Q_3 = 23.5$, và $IQR = 18$:

$$\text{Giới hạn dưới} = 5.5 - 1.5 \times 18 = -21.5$$

$$\text{Giới hạn trên} = 23.5 + 1.5 \times 18 = 50.5$$

→ Nếu một giá trị nhỏ hơn -21.5 hoặc lớn hơn 50.5, nó có thể là ngoại lệ.

c. Phương sai

*** Định nghĩa**

Phương sai thể hiện mức độ chênh lệch của các giá trị dữ liệu so với giá trị trung bình. Nếu phương sai lớn, dữ liệu có mức độ phân tán cao; ngược lại, nếu phương sai nhỏ, các giá trị dữ liệu tập trung quanh giá trị trung bình.

Phương sai thường được ký hiệu là:

- σ^2 (sigma bình phương) cho tổng thể.
- s^2 cho mẫu thống kê.

*** Công thức tính phương sai**

**** Phương sai của tổng thể**

Khi có toàn bộ dữ liệu trong tổng thể, phương sai được tính theo công thức:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2 \tag{2.20}$$

Trong đó:

- σ^2 là phương sai của tổng thể.
- N là số lượng phần tử trong tổng thể.
- x_i là từng giá trị dữ liệu.
- μ là giá trị trung bình của tổng thể:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{2.21}$$

**** Phương sai của mẫu**

Khi chỉ có một mẫu từ tổng thể, phương sai được ước lượng bằng công thức:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{2.22}$$

Trong đó:

- s^2 là phương sai của mẫu.
- n là số lượng phần tử trong mẫu.
- x_i là từng giá trị trong mẫu.
- $\bar{x}$ là trung bình của mẫu:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.23}$$

Lưu ý rằng trong công thức phương sai mẫu, mẫu số là $n-1$ thay vì n để bù trừ độ chệch khi ước lượng phương sai của tổng thể từ mẫu nhỏ.

*** Ý nghĩa của phương sai**

- **Đo lường mức độ phân tán**: Nếu phương sai lớn, dữ liệu phân tán rộng; nếu phương sai nhỏ, dữ liệu tập trung gần giá trị trung bình.
- **Quan trọng trong thống kê và học máy**: Phương sai được sử dụng rộng rãi trong kiểm định giả thuyết, hồi quy tuyến tính, và các thuật toán học máy để đánh giá mức độ biến động của dữ liệu.
- **So sánh độ biến động giữa các tập dữ liệu**: Ví dụ, phương sai giá cổ phiếu cao cho thấy biến động lớn, trong khi phương sai nhiệt độ môi trường thấp cho thấy nhiệt độ ổn định.

d. Độ lệch chuẩn (Standard Deviation)

*** Định Nghĩa Độ lệch chuẩn**

Độ lệch chuẩn là một thước đo phản ánh mức độ phân tán của tập dữ liệu so với giá trị trung bình. Nếu độ lệch chuẩn lớn, dữ liệu có xu hướng phân tán rộng; nếu nhỏ, dữ liệu tập trung quanh giá trị trung bình.

*** Công Thức Tính Độ lệch chuẩn**

**** Độ lệch chuẩn của Tổng thể**

Công thức tính độ lệch chuẩn của tổng thể:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \tag{2.24}$$

Trong đó:

- σ là độ lệch chuẩn của tổng thể.
- N là số phần tử trong tổng thể.
- x_i là từng giá trị dữ liệu.
- μ là giá trị trung bình của tổng thể:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{2.25}$$

**** Độ lệch chuẩn của Mẫu**

Khi chỉ có một mẫu từ tổng thể, công thức tính độ lệch chuẩn mẫu là:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{2.26}$$

Trong đó:

- s là độ lệch chuẩn của mẫu.
- n là số phần tử trong mẫu.
- x_i là từng giá trị trong mẫu.
- $\bar{x}$ là trung bình của mẫu:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{2.27}$$

Lưu ý rằng trong công thức độ lệch chuẩn mẫu, mẫu số là $n-1$ thay vì n để bù trừ độ chệch khi ước lượng độ lệch chuẩn của tổng thể từ mẫu nhỏ.

*** Ý Nghĩa của Độ lệch chuẩn**

- **Đo lường mức độ phân tán:** Nếu độ lệch chuẩn lớn, dữ liệu phân tán rộng; nếu nhỏ, dữ liệu tập trung gần giá trị trung bình.
- **Quan trọng trong thống kê và học máy:** Độ lệch chuẩn được sử dụng trong kiểm định giả thuyết, hồi quy tuyến tính, và các thuật toán học máy.
- **So sánh độ biến động giữa các tập dữ liệu:** Ví dụ, độ lệch chuẩn giá cổ phiếu cao cho thấy biến động lớn, trong khi độ lệch chuẩn nhiệt độ môi trường thấp cho thấy nhiệt độ ổn định.

2.2.3 Đo lường hình dạng phân phối dữ liệu

Hình dạng phân phối mô tả cách dữ liệu được sắp xếp xung quanh giá trị trung tâm.

a. Độ lệch (Skewness)

Định nghĩa: Độ lệch đo lường mức độ đối xứng của phân phối dữ liệu.

Công thức:

$$\text{Skewness} = \frac{\sum(x_i - \bar{x})^3}{(n-1)s^3} \tag{2.28}$$

Diễn giải:

- Skewness = 0: Phân phối đối xứng.
- Skewness > 0: Phân phối lệch phải (đuôi dài bên phải).
- Skewness < 0: Phân phối lệch trái (đuôi dài bên trái).

Ví dụ: Thu nhập của dân số thường có phân phối lệch phải vì có ít người có thu nhập rất cao.

b. Độ nhọn (Kurtosis)

Định nghĩa: Độ nhọn đo mức độ "tập trung"của dữ liệu quanh trung bình so với phân phối chuẩn.

Công thức:

$$Kurtosis = \frac{\sum(x_i - \bar{x})^4}{(n-1)s^4} \tag{2.29}$$

Diễn giải:

- $Kurtosis = 3$: Phân phối chuẩn (mesokurtic).
- $Kurtosis > 3$: Phân phối có đỉnh nhọn (leptokurtic), có nhiều ngoại lai.
- $Kurtosis < 3$: Phân phối có đỉnh thấp, dẹt hơn (platykurtic).

Ví dụ: Giá cổ phiếu có thể có kurtosis cao vì có nhiều biến động lớn bất thường.

2.2.4 Đo lường mối quan hệ giữa các biến

Các thước đo này giúp đánh giá mối quan hệ giữa hai biến số

a. Hiệp phương sai (Covariance)

Định nghĩa: Hiệp phương sai đo mức độ thay đổi cùng nhau của hai biến số.

Công thức:

$$Cov(X,Y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1} \tag{2.30}$$

Diễn giải:

- Nếu $Cov(X,Y) > 0$, hai biến có xu hướng tăng hoặc giảm cùng nhau.

- Nếu $Cov(X, Y) < 0$, một biến tăng thì biến kia giảm.
- Nếu $Cov(X, Y) = 0$, hai biến không liên hệ tuyến tính với nhau.

Ví dụ: Hiệp phương sai giữa thu nhập và chi tiêu của một hộ gia đình thường là dương.

b. Hệ số tương quan Pearson (Pearson Correlation)

Định nghĩa: Đo lường mức độ tuyến tính của mối quan hệ giữa hai biến.

Công thức:

$$r = \frac{\text{Cov}(X, Y)}{s_X s_Y} \tag{2.31}$$

Diễn giải:

- $r = 1$: Mối quan hệ tuyến tính hoàn hảo dương.
- $r = -1$: Mối quan hệ tuyến tính hoàn hảo âm.
- $r = 0$: Không có tương quan tuyến tính.

Ví dụ: Tương quan giữa số giờ học và điểm thi thường dương, nhưng không phải lúc nào cũng là 1.

c. Hệ số tương quan Spearman (Spearman Correlation)

* Giới thiệu

Hệ số tương quan Spearman (**Spearman's rank correlation coefficient**), ký hiệu là ρ hoặc r_s, đo mức độ tương quan giữa hai tập hợp dữ liệu dựa trên **thứ hạng** thay vì giá trị thực tế. Nó được sử dụng khi dữ liệu không tuân theo phân phối chuẩn hoặc khi mối quan hệ giữa hai biến không hoàn toàn tuyến tính.

* Công thức tính

Hệ số Spearman được tính theo công thức:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{2.32}$$

Trong đó:

- d_i là hiệu giữa thứ hạng của từng cặp dữ liệu: $d_i = \text{rank}(x_i) - \text{rank}(y_i)$.
- n là số lượng quan sát (cặp dữ liệu).

X	Y	Rank(X)	Rank(Y)	$d_i = \text{Rank}(X) - \text{Rank}(Y)$
10	200	1	2	-1
20	180	2	1	1
30	220	3	4	-1
40	240	4	5	-1
50	210	5	3	2

Bảng 2.2: Ví dụ về tính hệ số Spearman

** Bảng tính toán ví dụ

Tính tổng bình phương sai lệch:

$$\sum d_i^2 = 1^2 + (-1)^2 + (-1)^2 + (-1)^2 + 2^2 = 1 + 1 + 1 + 1 + 4 = 8 \tag{2.33}$$

Thay vào công thức tính Spearman:

$$r_s = 1 - \frac{6(8)}{5(25-1)} = 1 - \frac{48}{120} = 1 - 0.4 = 0.6 \tag{2.34}$$

Kết quả $r_s = 0.6$ cho thấy mối quan hệ tương quan dương vừa phải giữa X và Y.

* So sánh với Pearson

Tiêu chí	Pearson (r)	Spearman (r_s)
Dữ liệu yêu cầu	Phân phối chuẩn	Không yêu cầu
Tính toán dựa trên	Giá trị thực	Thứ hạng
Đo lường quan hệ	Tuyến tính	Phi tuyến đơn điệu
Nhạy cảm với ngoại lệ	Có	Ít hơn

Bảng 2.3: So sánh hệ số Pearson và Spearman

* Khi nào nên sử dụng Spearman?

- Khi dữ liệu không tuân theo phân phối chuẩn.
- Khi dữ liệu có quan hệ phi tuyến nhưng đơn điệu (tăng hoặc giảm liên tục).
- Khi có nhiều ngoại lệ ảnh hưởng đến phân phối của dữ liệu.
- Khi làm việc với dữ liệu xếp hạng (ordinal data).

Hệ số tương quan Spearman là một công cụ hữu ích để đo lường mối quan hệ giữa hai biến trong trường hợp dữ liệu không tuyến tính hoặc không có phân phối chuẩn. Nó có tính ứng dụng cao trong phân tích dữ liệu xã hội, tài chính, và khoa học tự nhiên.

2.3 Xử lý dữ liệu trong kinh tế lượng

2.3.1 Định nghĩa bài toán

2.3.2 Thu thập dữ liệu

2.3.3 Xử lý dữ liệu

2.3.4 Kết luận

Chương 3

Luật phân bố xác suất

3.1 Giới thiệu

Xác suất là một công cụ quan trọng trong toán học và thống kê để mô tả sự không chắc chắn của các hiện tượng ngẫu nhiên. Trong chương này, chúng ta sẽ trình bày các luật phân bố xác suất, bao gồm phân bố rời rạc và liên tục, cùng với các định lý quan trọng.

3.2 Các Định Nghĩa Cơ Bản

3.2.1 Biến ngẫu nhiên

Định nghĩa: Một biến ngẫu nhiên là một hàm số ánh xạ từ không gian mẫu (tập hợp tất cả các kết quả có thể của một thí nghiệm ngẫu nhiên) vào tập số thực $\mathbb{R}$. Nói cách khác, biến ngẫu nhiên là một đại lượng số học có thể nhận các giá trị khác nhau do yếu tố ngẫu nhiên.

Ví dụ minh họa: Giả sử tung một con xúc xắc.

- Không gian mẫu: $S = \{1, 2, 3, 4, 5, 6\}$.
- Định nghĩa biến ngẫu nhiên X là "số chấm xuất hiện trên mặt ngửa của xúc xắc".
- Khi đó, X có thể nhận các giá trị $1, 2, 3, 4, 5, 6$, mỗi giá trị này tương ứng với một khả năng xảy ra.

3.2.2 Hàm phân bố xác suất (CDF - Cumulative Distribution Function)

Định nghĩa: Hàm phân bố xác suất của một biến ngẫu nhiên X được định nghĩa là:

$$F_X(x) = P(X \leq x), \forall x \in \mathbb{R}.$$

Hàm phân bố xác suất giúp mô tả cách xác suất được phân bố trên tập giá trị của biến ngẫu nhiên.

Ví dụ minh họa (Biến ngẫu nhiên rời rạc): Xét biến ngẫu nhiên X là số chấm trên mặt ngửa của một xúc xắc 6 mặt cân bằng. Khi đó, ta có:

$$F_X(x) = \begin{cases} 0, & x < 1 \\ \frac{1}{6}, & 1 \le x < 2 \\ \frac{2}{6}, & 2 \le x < 3 \\ \frac{3}{6}, & 3 \le x < 4 \\ \frac{4}{6}, & 4 \le x < 5 \\ \frac{5}{6}, & 5 \le x < 6 \\ 1, & x \ge 6 \end{cases}$$

3.2.3 Hàm mật độ xác suất (PDF - Probability Density Function)

Định nghĩa: Nếu X là một biến ngẫu nhiên liên tục, thì xác suất để X nằm trong khoảng $[a, b]$ được xác định bằng tích phân:

$$P(a \le X \le b) = \int_a^b f_X(x)\, dx.$$

Ví dụ minh họa: Xét biến ngẫu nhiên X có phân bố chuẩn (Gaussian) với kỳ vọng $\mu = 0$ và phương sai $\sigma^2 = 1$:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Xác suất X nằm trong khoảng từ -1 đến 1:

$$P(-1 \le X \le 1) = \int_{-1}^{1} f_X(x)\, dx \approx 0.6826.$$

3.2.4 Hàm khối xác suất (PMF - Probability Mass Function)

Định nghĩa: Nếu X là một biến ngẫu nhiên rời rạc, thì xác suất để X nhận giá trị x_i được xác định bằng hàm khối xác suất:

$$P(X = x_i) = p_X(x_i), \quad \sum_i p_X(x_i) = 1.$$

Ví dụ minh họa: Xét biến ngẫu nhiên X biểu diễn số lần xuất hiện mặt ngửa khi tung 2 đồng xu cân bằng. Khi đó, X có thể nhận các giá trị 0, 1, hoặc 2 với xác suất:

$$p_X(0) = P(X = 0) = \frac{1}{4}, \quad p_X(1) = P(X = 1) = \frac{2}{4}, \quad p_X(2) = P(X = 2) = \frac{1}{4}.$$

Tổng tất cả các xác suất:

$$\sum_i p_X(x_i) = \frac{1}{4} + \frac{2}{4} + \frac{1}{4} = 1.$$

Điều này xác nhận rằng tổng xác suất của tất cả giá trị có thể xảy ra bằng 1.

3.3 Luật Số Lớn

Luật số lớn (LLN) là một định lý cơ bản trong xác suất thống kê, mô tả xu hướng hội tụ của trung bình mẫu về giá trị kỳ vọng khi kích thước mẫu tăng. LLN đóng vai trò quan trọng trong thống kê và nhiều ứng dụng thực tế như kinh tế, khoa học dữ liệu.

Có hai dạng của định lý Luật số lớn:

3.3.1 Luật số lớn yếu (Weak Law of Large Numbers - WLLN)

Giả sử $X_1, X_2, \ldots, X_n$ là một dãy các biến ngẫu nhiên độc lập và cùng phân phối (i.i.d) với kỳ vọng hữu hạn $E[X_i] = \mu$. Khi đó, với mọi $\varepsilon > 0$, ta có:

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \varepsilon\right) = 0 \tag{3.1}$$

Điều này có nghĩa là khi kích thước mẫu n đủ lớn, xác suất để trung bình mẫu khác xa kỳ vọng thực tế sẽ tiến về 0.

Giải thích các ký hiệu:

- $X_1, X_2, \ldots, X_n$: Các biến ngẫu nhiên độc lập, cùng phân phối.
- $E[X_i]$: Kỳ vọng của biến ngẫu nhiên X_i, ký hiệu là μ.
- $\frac{1}{n}\sum_{i=1}^{n} X_i$: Trung bình mẫu.
- $P(A)$: Xác suất xảy ra của biến cố A.
- $\lim_{n\to\infty}$: Giới hạn khi kích thước mẫu tiến đến vô cùng.
- ε: Một số dương nhỏ tùy ý.

3.3.2 Luật số lớn mạnh (Strong Law of Large Numbers - SLLN)

Với cùng điều kiện như trên, Luật số lớn mạnh phát biểu rằng:

$$P\left(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mu\right) = 1 \tag{3.2}$$

Tức là trung bình mẫu sẽ hội tụ chắc chắn (almost surely) về kỳ vọng μ khi $n \to \infty$.

Giải thích các ký hiệu:

- Các ký hiệu tương tự như Luật số lớn yếu.
- $P(A) = 1$: Sự kiện A xảy ra với xác suất chắc chắn.
- $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mu$: Trung bình mẫu hội tụ về kỳ vọng μ khi $n \to \infty$.

3.3.3 Ví dụ minh họa

Giả sử ta có một đồng xu không cân bằng với xác suất xuất hiện mặt ngửa là $p = 0.6$. Gieo đồng xu n lần và tính xác suất trung bình của số lần xuất hiện mặt ngửa:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{3.3}$$

Theo Luật số lớn, khi n tăng, $\bar{X}_n$ sẽ hội tụ về $p = 0.6$.

Giải thích các ký hiệu:

- X_i: Biến ngẫu nhiên nhận giá trị 1 nếu lần gieo thứ i ra mặt ngửa, và 0 nếu ra mặt sấp.
- $\bar{X}_n$: Trung bình của các lần thử, tức là tỷ lệ số lần xuất hiện mặt ngửa trong n lần thử.
- Khi n càng lớn, $\bar{X}_n$ sẽ tiến gần về giá trị kỳ vọng $p = 0.6$, theo Luật số lớn.

Luật số lớn cho thấy khi thu thập nhiều dữ liệu hơn, giá trị trung bình của mẫu sẽ gần hơn với giá trị kỳ vọng thực tế. Đây là cơ sở lý thuyết quan trọng trong thống kê, tài chính, trí tuệ nhân tạo và nhiều lĩnh vực khác.

3.4 Các luật phân bố xác suất quan trọng

3.4.1 Phân bố nhị thức

Định nghĩa Phân bố nhị thức mô tả số lần xảy ra của một sự kiện trong một số lần thử độc lập, khi mỗi lần thử chỉ có hai kết quả: **thành công** hoặc **thất bại**.

Một biến ngẫu nhiên X tuân theo phân bố nhị thức với các tham số n (số lần thử) và p (xác suất thành công trong mỗi lần thử) nếu xác suất để X nhận giá trị k (tức là có đúng k lần thành công trong n phép thử) được tính theo công thức:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, 2, \ldots, n. \tag{3.4}$$

Trong đó:

- $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ là hệ số nhị thức (binomial coefficient).
- p^k là xác suất có đúng k lần thành công.
- $(1-p)^{n-k}$ là xác suất có $(n-k)$ lần thất bại.

Kỳ vọng và Phương sai

$$E(X) = np, \quad \text{Var}(X) = np(1-p). \tag{3.5}$$

Ví Dụ Giả sử một bài kiểm tra trắc nghiệm có 10 câu hỏi, mỗi câu có 4 đáp án nhưng chỉ có 1 đáp án đúng. Một học sinh chọn đáp án ngẫu nhiên cho mỗi câu. Gọi X là số câu trả lời đúng, thì X tuân theo phân bố nhị thức $B(10, 0.25)$ vì xác suất chọn đúng một đáp án là $p = 0.25$.

Ứng dụng thực tế Phân bố nhị thức có nhiều ứng dụng trong thực tế, bao gồm:

- Xác suất một sản phẩm bị lỗi khi lấy mẫu kiểm tra trong dây chuyền sản xuất.
- Dự đoán số lượng khách hàng tiềm năng sẽ mua sản phẩm sau khi quảng cáo.
- Xác suất thắng một trò chơi nếu người chơi có một tỷ lệ chiến thắng cố định.

3.4.2 Phân Bố Poisson

Định nghĩa : Phân bố Poisson là một phân bố xác suất rời rạc mô tả số lần xảy ra của một sự kiện trong một khoảng thời gian (hoặc không gian) nhất định khi các sự kiện đó xảy ra độc lập với nhau và có tỷ lệ trung bình không đổi.

Một biến ngẫu nhiên X tuân theo phân bố Poisson với tham số $\lambda > 0$ nếu nó có xác suất:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots \tag{3.6}$$

trong đó:

- λ là số lần xảy ra trung bình của sự kiện trong khoảng thời gian hoặc không gian xác định.
- $k!$ là giai thừa của k với quy ước $0! = 1$.
- $e \approx 2.718$ là hằng số Euler.

Ý nghĩa và ứng dụng Phân bố Poisson được sử dụng để mô tả số lần xảy ra của các sự kiện hiếm gặp trong một khoảng thời gian hoặc không gian cố định, chẳng hạn như:

- Số cuộc gọi đến tổng đài trong một giờ.
- Số lỗi xảy ra trong một hệ thống máy tính trong một ngày.
- Số tai nạn giao thông trên một đoạn đường trong một tuần.

- Số khách hàng đến một cửa hàng trong một khoảng thời gian nhất định.

Các đặc trưng của phân bố Poisson

- **Kỳ vọng (trung bình)**: $E(X) = \lambda$
- **Phương sai**: $\text{Var}(X) = \lambda$
- **Độ lệch chuẩn**: $\sigma = \sqrt{\lambda}$

Một số tính chất quan trọng:

- Phân bố Poisson có thể được sử dụng để xấp xỉ phân bố nhị thức $B(n, p)$ khi n lớn và p nhỏ sao cho $\lambda = np$.
- Nếu $X_1 \sim \text{Poisson}(\lambda_1)$ và $X_2 \sim \text{Poisson}(\lambda_2)$ độc lập, thì tổng của chúng cũng tuân theo phân bố Poisson:

$$X_1 + X_2 \sim \text{Poisson}(\lambda_1 + \lambda_2). \tag{3.7}$$

Ví dụ minh họa

- **Ví dụ 1: Số cuộc gọi đến tổng đài** Giả sử một tổng đài nhận trung bình 4 cuộc gọi mỗi phút. Hỏi xác suất để trong một phút có đúng 2 cuộc gọi đến là bao nhiêu?

  Áp dụng công thức phân bố Poisson với $\lambda = 4$, $k = 2$:

$$P(X = 2) = \frac{4^2 e^{-4}}{2!} = \frac{16e^{-4}}{2} \approx 0.1465. \tag{3.8}$$

  Vậy xác suất nhận đúng 2 cuộc gọi trong một phút là khoảng **14.65%**.

- **Ví dụ 2: Số lỗi phần mềm** Một phần mềm có trung bình 3 lỗi xảy ra mỗi ngày. Xác suất để hôm nay có **không có lỗi nào** là bao nhiêu?

  Dùng công thức với $\lambda = 3$, $k = 0$:

$$P(X = 0) = \frac{3^0 e^{-3}}{0!} = e^{-3} \approx 0.0498. \tag{3.9}$$

  Vậy xác suất không có lỗi nào trong ngày hôm nay là **4.98%**.

Mối liên hệ với các phân bố khác

- Khi $n \to \infty$, $p \to 0$ nhưng $np = \lambda$ cố định, phân bố nhị thức $B(n, p)$ xấp xỉ phân bố Poisson với tham số λ.
- Khi λ lớn, phân bố Poisson có thể được xấp xỉ bằng phân bố chuẩn:

$$X \approx N(\lambda, \lambda). \tag{3.10}$$

3.4.3 Phân Bố Chuẩn (Gauss)

Phân bố chuẩn, còn gọi là phân bố Gauss, là một trong những phân bố quan trọng nhất trong thống kê và xác suất. Nó được sử dụng rộng rãi trong nhiều lĩnh vực như tài chính, khoa học dữ liệu, kỹ thuật và kinh tế.
Định nghĩa Phân bố chuẩn có dạng hàm mật độ xác suất (PDF) như sau:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{3.11}$$

trong đó:

- μ là kỳ vọng (trung bình) của phân bố.
- σ là độ lệch chuẩn.
- σ^2 là phương sai.
- x là biến ngẫu nhiên tuân theo phân bố chuẩn.

Đặc điểm của Phân Bố Chuẩn
Phân bố chuẩn có một số đặc điểm quan trọng:

1. Đối xứng quanh giá trị trung bình μ.
2. Đường cong hình chuông với đỉnh tại $x = \mu$.
3. Tổng diện tích dưới đường cong bằng 1.
4. Khoảng $\mu \pm \sigma$ chứa khoảng 68.27% dữ liệu.
5. Khoảng $\mu \pm 2\sigma$ chứa khoảng 95.45% dữ liệu.
6. Khoảng $\mu \pm 3\sigma$ chứa khoảng 99.73% dữ liệu.

Phân bố chuẩn tắc
Phân bố chuẩn tắc (standard normal distribution) là trường hợp đặc biệt của phân bố chuẩn với:

- $\mu = 0$
- $\sigma = 1$

Trong trường hợp này, công thức phân bố chuẩn trở thành:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \tag{3.12}$$

Khi một biến ngẫu nhiên X tuân theo phân bố chuẩn với kỳ vọng μ và độ lệch chuẩn σ, ta có thể chuẩn hóa về phân bố chuẩn tắc bằng công thức:

$$Z = \frac{X - \mu}{\sigma} \tag{3.13}$$

Biến đổi này giúp ta dễ dàng tra cứu bảng phân bố chuẩn và tính toán xác suất.

Ứng dụng của Phân bố chuẩn

Phân bố chuẩn có rất nhiều ứng dụng trong thực tế:

- Kiểm định giả thuyết thống kê.
- Mô hình hóa dữ liệu thực tế trong nhiều lĩnh vực.
- Dùng trong kiểm soát chất lượng sản xuất.
- Ước lượng khoảng tin cậy trong thống kê.
- Dự báo và phân tích rủi ro trong tài chính.

3.5 Bậc tự do (Degrees of Freedom - DoF)

Trong thống kê, bậc tự do liên quan đến số lượng giá trị có thể thay đổi tự do trong một phép tính, thường xuất hiện trong kiểm định giả thuyết và phân bố xác suất.

3.5.1 Định nghĩa toán học của bậc tự do

Trong thống kê, **bậc tự do** của một phép tính là số lượng giá trị có thể thay đổi tự do mà không bị ràng buộc bởi các điều kiện hoặc mối quan hệ toán học khác.

Nếu có n quan sát nhưng một số quan sát bị ràng buộc bởi một hoặc nhiều điều kiện, thì bậc tự do là số lượng giá trị có thể thay đổi một cách độc lập.

Công thức tổng quát của bậc tự do trong thống kê:

$$df = n - k \tag{3.14}$$

trong đó:

- n là tổng số quan sát,
- k là số lượng tham số ước lượng từ dữ liệu.

Ví dụ: Nếu bạn có 5 số và biết trung bình của chúng, thì chỉ cần biết 4 số đầu tiên là có thể suy ra số thứ 5, nghĩa là chỉ có 4 bậc tự do.

3.5.2 Ý nghĩa trong ước lượng thống kê

Trong thống kê, khi tính toán các đặc trưng của mẫu (ví dụ: phương sai, độ lệch chuẩn), bậc tự do ảnh hưởng trực tiếp đến độ chính xác của ước lượng.

Phương sai mẫu s^2 Khi tính phương sai của mẫu, ta sử dụng công thức:

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} \tag{3.15}$$

Ở đây, $n-1$ là số bậc tự do, vì ta đã sử dụng một quan sát để tính giá trị trung bình $\bar{x}$, làm giảm số lượng giá trị có thể thay đổi độc lập.

Nếu dùng n thay vì $n-1$, ước lượng phương sai sẽ bị lệch (underestimate).

Ứng dụng thực tế: Khi tính phương sai của một tập dữ liệu nhỏ, việc sử dụng bậc tự do $n-1$ giúp tạo ra một ước lượng không thiên lệch cho phương sai tổng thể.

3.5.3 Bậc tự do trong kiểm định giả thuyết

Bậc tự do rất quan trọng trong các kiểm định thống kê như kiểm định t-test, kiểm định χ^2, và ANOVA.

*** Kiểm định t-test**

Kiểm định t-test được sử dụng để so sánh trung bình của hai nhóm.

Công thức bậc tự do trong kiểm định t-test một mẫu:

$$df = n - 1 \tag{3.16}$$

Trong kiểm định t-test hai mẫu độc lập:

$$df = n_1 + n_2 - 2 \tag{3.17}$$

trong đó n_1, n_2 là kích thước mẫu của hai nhóm.

Ứng dụng thực tế:

- So sánh điểm thi giữa hai lớp học.
- Đánh giá hiệu quả của một loại thuốc giữa hai nhóm bệnh nhân.

*** Kiểm định χ^2 (Kiểm định phù hợp và kiểm định độc lập)**

Kiểm định χ^2 giúp xác định sự khác biệt giữa các nhóm danh mục (categorical data).

Công thức bậc tự do trong bảng tần suất:

$$df = (r-1) \times (c-1) \tag{3.18}$$

trong đó r là số hàng, c là số cột.

Ứng dụng thực tế:

- Kiểm tra xem giới tính có ảnh hưởng đến sở thích mua sắm hay không.
- Đánh giá mối quan hệ giữa thói quen ăn uống và tình trạng sức khỏe.

*** Phân tích phương sai (ANOVA)**

Trong ANOVA, bậc tự do giúp xác định nguồn biến thiên giữa các nhóm và bên trong nhóm.

Công thức:

$$df_{between} = k - 1 \tag{3.19}$$

$$df_{within} = N - k \tag{3.20}$$

trong đó k là số nhóm và N là tổng số quan sát.

Ứng dụng thực tế:

- So sánh hiệu suất của ba phương pháp giảng dạy khác nhau.
- Đánh giá hiệu quả của ba chiến lược tiếp thị.

3.5.4 Bậc tự do trong hồi quy tuyến tính

Bậc tự do cũng quan trọng trong hồi quy tuyến tính vì nó ảnh hưởng đến chất lượng mô hình dự báo.

Trong mô hình hồi quy tuyến tính có dạng:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon \tag{3.21}$$

Bậc tự do được tính là:

$$df = n - (k + 1) \tag{3.22}$$

trong đó:

- n là số quan sát.
- k là số biến độc lập.

Ứng dụng thực tế

- Dự đoán giá bất động sản dựa trên diện tích, số phòng ngủ, và vị trí.
- Phân tích các yếu tố ảnh hưởng đến doanh thu doanh nghiệp.

3.5.5 Tác động của bậc tự do đến phân phối xác suất

Bậc tự do cũng ảnh hưởng đến hình dạng của một số phân phối xác suất như phân phối t-Student, phân phối χ^2, và phân phối F.

- Khi bậc tự do tăng, phân phối t-Student dần tiến gần đến phân phối chuẩn.
- Trong phân phối χ^2, bậc tự do ảnh hưởng đến mức độ phân tán của phân phối.
- Trong kiểm định F, bậc tự do ảnh hưởng đến xác suất từ chối giả thuyết không.

Ứng dụng thực tế

- Khi kiểm tra giả thuyết với số lượng mẫu nhỏ, ta sử dụng phân phối t-Student thay vì phân phối chuẩn.
- Trong kiểm định phương sai, số bậc tự do quyết định xác suất sai lầm loại I.

Chương 4

Các phương pháp phân tích dữ liệu bằng mô hình thống kê

Trong kinh tế lượng, hai phương pháp tiếp cận quan trọng cần đề cập là ước tính tham số và kiểm định giả thuyết, vì chúng là nền tảng để xây dựng và đánh giá các mô hình kinh tế lượng.

4.1 Phương pháp ước lượng tham số (Parameter Estimation)

4.1.1 Phương pháp bình quân nhỏ nhất (OLS - Ordinary Least Squares)

a. Giới thiệu về phương pháp OLS

Phương pháp bình phương nhỏ nhất (OLS) là một trong những phương pháp phổ biến nhất trong kinh tế lượng và thống kê để ước lượng các tham số của mô hình hồi quy tuyến tính. Mục tiêu của OLS là tìm ra các hệ số hồi quy sao cho tổng bình phương phần dư (sai số giữa giá trị thực tế và giá trị dự báo) là nhỏ nhất.

b. Mô hình hồi quy tuyến tính tổng quát

Giả sử mô hình hồi quy tuyến tính có dạng:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i \tag{4.1}$$

trong đó:

- Y_i là biến phụ thuộc (biến kết quả)
- X_{ij} là các biến độc lập (biến giải thích)
- $\beta_0, \beta_1, \ldots, \beta_k$ là các hệ số hồi quy cần ước lượng
- ε_i là sai số ngẫu nhiên

c. Nguyên lý của phương pháp OLS

Phương pháp OLS tìm kiếm các hệ số β bằng cách cực tiểu hóa tổng bình phương sai số:

$$S(\beta) = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - (\beta_0 + \beta_1 X_{i1} + ... + \beta_k X_{ik}))^2 \tag{4.2}$$

Để tìm các hệ số β, ta giải hệ phương trình bình thường (normal equations):

$$(X'X)\hat{\beta} = X'Y \tag{4.3}$$

trong đó:

- X là ma trận dữ liệu của các biến độc lập $(n \times k)$
- Y là vector của biến phụ thuộc $(n \times 1)$
- $\hat{\beta}$ là vector hệ số hồi quy $(k \times 1)$
- X' là ma trận chuyển vị của X

d. Các giả định của OLS

Phương pháp OLS hoạt động tốt khi các giả định sau được thỏa mãn:

1. **Tuyến tính**: Mô hình phải có dạng tuyến tính đối với các tham số.
2. **Kỳ vọng bằng 0 của sai số**: $E(\varepsilon_i) = 0$.
3. **Độc lập của sai số**: Sai số không có tương quan với nhau (không có tự tương quan).
4. **Phương sai đồng nhất**: Sai số có phương sai không đổi (không có hiện tượng phương sai thay đổi).
5. **Không có đa cộng tuyến hoàn hảo**: Các biến độc lập không được có tương quan tuyến tính hoàn hảo.
6. **Phân phối chuẩn của sai số (nếu mẫu nhỏ)**: Giả định này giúp kiểm định giả thuyết và tính khoảng tin cậy chính xác hơn.

e. Ước lượng và kiểm định ý nghĩa của hệ số hồi quy

Sau khi ước lượng các hệ số hồi quy bằng OLS, ta kiểm định ý nghĩa thống kê của chúng bằng kiểm định t (t-test). Giả thuyết kiểm định cho mỗi hệ số β_j:

- $H_0 : \beta_j = 0$ (hệ số không có ý nghĩa thống kê)
- $H_1 : \beta_j \neq 0$ (hệ số có ý nghĩa thống kê)

Chỉ số thống kê t được tính bằng:

$$t_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \tag{4.4}$$

trong đó $SE(\hat{\beta}_j)$ là sai số chuẩn của hệ số ước lượng β_j. Nếu giá trị p-value của kiểm định nhỏ hơn mức ý nghĩa α (thường là 0.05), ta bác bỏ H_0 và kết luận rằng hệ số có ý nghĩa thống kê.

f. Đánh giá chất lượng mô hình hồi quy

*** Hệ số xác định R^2**

Hệ số xác định R^2 đo lường mức độ giải thích của mô hình đối với biến phụ thuộc:

$$R^2 = 1 - \frac{SSR}{SST} = \frac{SSE}{SST} \tag{4.5}$$

trong đó:

- SSR là tổng bình phương sai số (Sum of Squared Residuals)
- SST là tổng bình phương tổng thể (Total Sum of Squares)
- SSE là tổng bình phương hồi quy (Sum of Squares for Regression)

*** Kiểm định F**

Kiểm định F đánh giá xem mô hình hồi quy có phù hợp hay không:

$$F = \frac{(SST - SSR)/k}{SSR/(n-k-1)} \tag{4.6}$$

Nếu p-value của kiểm định F nhỏ hơn mức ý nghĩa α, mô hình được xem là có ý nghĩa tổng thể.

Phương pháp bình phương nhỏ nhất (OLS) là một kỹ thuật phổ biến để ước lượng mô hình hồi quy tuyến tính. Khi các giả định của OLS được thỏa mãn, phương pháp này giúp chúng ta có được các ước lượng không chệch, hiệu quả và tối ưu. Tuy nhiên, nếu các giả định bị vi phạm, có thể cần đến các phương pháp hồi quy khác như hồi quy tổng quát (GLS), hồi quy Ridge hoặc hồi quy LASSO để khắc phục.

4.1.2 Phương pháp hợp lý tối đa (MLE - Maximum Likelihood Estimation)

a. Giới thiệu về phương pháp MLE

Phương pháp hợp lý tối đa (MLE - Maximum Likelihood Estimation) là một phương pháp thống kê dùng để ước lượng các tham số của một mô hình xác

suất dựa trên dữ liệu quan sát. Nguyên tắc cơ bản của MLE là tìm giá trị của các tham số sao cho xác suất quan sát được tập dữ liệu hiện có là lớn nhất.

Nếu mô hình xác suất có phân phối xác suất $f(Y \mid \theta)$, trong đó:

- $Y = (Y_1, Y_2, ..., Y_n)$ là tập dữ liệu quan sát
- θ là vector tham số cần ước lượng

Thì MLE sẽ tìm giá trị $\hat{\theta}$ sao cho hàm hợp lý $L(\theta)$ đạt cực đại:

$$\hat{\theta} = \arg\max_{\theta} L(\theta) \tag{4.7}$$

trong đó $L(\theta)$ là hàm hợp lý của dữ liệu:

$$L(\theta) = P(Y \mid \theta) = \prod_{i=1}^{n} f(Y_i \mid \theta) \tag{4.8}$$

b. Hàm hợp lý và hàm log-hợp lý

Do tích của nhiều xác suất nhỏ có thể dẫn đến vấn đề số học, ta thường sử dụng hàm log-hợp lý:

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(Y_i \mid \theta) \tag{4.9}$$

Bài toán tối đa hóa hàm hợp lý chuyển thành:

$$\hat{\theta} = \arg\max_{\theta} \ell(\theta) \tag{4.10}$$

c. Ví dụ: Ước lượng tham số trong phân phối chuẩn

Giả sử dữ liệu $Y_1, Y_2, ..., Y_n$ được lấy mẫu từ phân phối chuẩn:

$$Y_i \sim \mathcal{N}(\mu, \sigma^2) \tag{4.11}$$

Hàm mật độ xác suất của phân phối chuẩn là:

$$f(Y_i \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y_i - \mu)^2}{2\sigma^2}\right) \tag{4.12}$$

Hàm log-hợp lý:

$$\ell(\mu, \sigma^2) = \sum_{i=1}^{n} \left[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(Y_i - \mu)^2}{2\sigma^2}\right] \tag{4.13}$$

Tính đạo hàm theo μ:

$$\frac{\partial \ell}{\partial \mu} = \sum_{i=1}^{n} \frac{Y_i - \mu}{\sigma^2} \tag{4.14}$$

Giải phương trình $\frac{\partial \ell}{\partial \mu} = 0$ ta được:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} Y_i \tag{4.15}$$

Tương tự, ước lượng của σ^2 là:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{\mu})^2 \tag{4.16}$$

d. Các tính chất của ước lượng MLE

* Tính nhất quán

Ước lượng MLE là nhất quán, nghĩa là khi $n \to \infty$, giá trị ước lượng $\hat{\theta}$ hội tụ về giá trị thật θ_0.

* Tính không chệch và hiệu quả

Dưới các điều kiện thông thường, MLE gần như không chệch và đạt được giới hạn Cramér-Rao.

* Phân phối tiệm cận

Khi kích thước mẫu đủ lớn:

$$\hat{\theta} \sim \mathcal{N}(\theta_0, I(\theta_0)^{-1}) \tag{4.17}$$

trong đó $I(\theta)$ là ma trận thông tin Fisher:

$$I(\theta) = -E\left[\frac{\partial^2 \ell(\theta)}{\partial \theta^2}\right] \tag{4.18}$$

e. Kiểm định giả thuyết với MLE

Sau khi ước lượng θ, ta có thể kiểm định giả thuyết bằng kiểm định Wald:

$$z = \frac{\hat{\theta} - \theta_0}{\text{SE}(\hat{\theta})} \sim \mathcal{N}(0, 1) \tag{4.19}$$

trong đó $\text{SE}(\hat{\theta})$ là sai số chuẩn của $\hat{\theta}$.

Phương pháp hợp lý tối đa (MLE) là một phương pháp mạnh mẽ để ước lượng tham số của mô hình xác suất. MLE có nhiều tính chất quan trọng như tính nhất quán, hiệu quả và không chệch tiệm cận. Trong thực tế, MLE được

áp dụng rộng rãi trong thống kê, kinh tế lượng, machine learning và nhiều lĩnh vực khác.

4.1.3 Ước lượng Hậu nghiệm Tối đa (Maximum A Posteriori - MAP)

a. Giới thiệu

Ước lượng MAP là một phương pháp thống kê trong khuôn khổ Bayesian, được sử dụng để tìm tham số θ sao cho xác suất hậu nghiệm $P(\theta|D)$ đạt giá trị lớn nhất.

MAP là một mở rộng của Ước lượng hợp lý tối đa (MLE) bằng cách đưa thêm phân bố tiên nghiệm $P(\theta)$ vào mô hình, giúp kiểm soát nhiễu và tránh hiện tượng quá khớp.

b. Công thức toán học

Theo định lý Bayes, xác suất hậu nghiệm được tính bởi:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} \tag{4.20}$$

Trong đó:

- $P(\theta|D)$ là xác suất hậu nghiệm của tham số θ.
- $P(D|\theta)$ là hàm hợp lý (likelihood) – xác suất của dữ liệu quan sát được khi biết tham số θ.
- $P(\theta)$ là phân bố tiên nghiệm (prior) của θ.
- $P(D)$ là hàm bằng chứng:

$$P(D) = \int P(D|\theta)P(\theta)d\theta \tag{4.21}$$

Vì $P(D)$ là một hằng số, nên việc tối đa hóa $P(\theta|D)$ tương đương với tối đa hóa tử số:

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(D|\theta)P(\theta) \tag{4.22}$$

c. So sánh với MLE

Nếu phân bố tiên nghiệm $P(\theta)$ là đều (Uniform prior), ta có:

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(D|\theta) \tag{4.23}$$

Điều này chính là ước lượng MLE.

*** Ví dụ cụ thể**

=> MAP với phân phối Gaussian

Giả sử ta muốn ước lượng tham số θ từ dữ liệu $D = \{x_1, x_2, ..., x_n\}$ theo mô hình:

$$x_i \sim \mathcal{N}(\theta, \sigma^2)$$

với tiên nghiệm:

$$\theta \sim \mathcal{N}(\mu_0, \sigma_0^2)$$

**** Hàm hợp lý:**

$$P(D|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \theta)^2}{2\sigma^2}\right)$$

**** Phân bố tiên nghiệm:**

$$P(\theta) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\theta - \mu_0)^2}{2\sigma_0^2}\right)$$

**** Xác suất hậu nghiệm:**

$$P(\theta|D) \propto P(D|\theta)P(\theta)$$

Lấy log của biểu thức trên và tối đa hóa theo θ:

$$\log P(\theta|D) = -\sum_{i=1}^{n} \frac{(x_i - \theta)^2}{2\sigma^2} - \frac{(\theta - \mu_0)^2}{2\sigma_0^2} + C$$

Lấy đạo hàm theo θ và đặt bằng 0:

$$\sum_{i=1}^{n} \frac{x_i - \theta}{\sigma^2} + \frac{\mu_0 - \theta}{\sigma_0^2} = 0$$

Giải ra được:

$$\hat{\theta}_{MAP} = \frac{\frac{\mu_0}{\sigma_0^2} + \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}$$

Ta thấy rằng MAP là trung bình có trọng số giữa trung bình dữ liệu và giá trị tiên nghiệm.

- Nếu $\sigma_0^2 \to \infty$ (tức là không có prior), ta thu được MLE:

$$\hat{\theta}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Nếu $n \to \infty$, dữ liệu thống trị prior, nên MAP gần với MLE.

d. Ứng dụng của MAP

- Học máy (Machine Learning): MAP được dùng trong Hồi quy Bayesian và Phân loại Naïve Bayes.
- Xử lý ngôn ngữ tự nhiên (NLP): Smoothing trong mô hình Markov ẩn (HMM).
- Xử lý ảnh (Computer Vision): Khử nhiễu hình ảnh bằng cách thêm prior vào mô hình.
- Kinh tế lượng: MAP giúp giảm hiện tượng đa cộng tuyến trong hồi quy tuyến tính.

Ước lượng MAP giúp ước lượng tham số bằng cách kết hợp thông tin từ dữ liệu và thông tin tiên nghiệm. Nó mở rộng MLE bằng cách thêm prior vào mô hình, giúp ổn định ước lượng trong trường hợp dữ liệu ít hoặc có nhiễu cao.

4.1.4 Ước lượng Bayes đầy đủ (Bayesian Estimation)

a. Giới thiệu

Ước lượng Bayes đầy đủ là một phương pháp suy luận thống kê dựa trên lý thuyết Bayes, kết hợp thông tin từ dữ liệu quan sát với thông tin tiên nghiệm (prior) để đưa ra ước lượng xác suất của tham số cần ước lượng.

b. Công thức Bayes cho ước lượng tham số

Giả sử ta có dữ liệu quan sát $D = \{x_1, x_2, \ldots, x_n\}$ và cần ước lượng tham số θ. Theo định lý Bayes, xác suất hậu nghiệm của tham số θ được tính bằng:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} \tag{4.24}$$

trong đó:

- $P(D|\theta)$ là **hàm hợp lý (likelihood)**, thể hiện xác suất quan sát dữ liệu D khi biết tham số θ.
- $P(\theta)$ là **phân bố tiên nghiệm (prior distribution)** của θ.
- $P(D)$ là **bằng chứng (evidence)**, được tính bằng:

$$P(D) = \int P(D|\theta)P(\theta)d\theta \tag{4.25}$$

c. Ước lượng Bayes đầy đủ và kỳ vọng hậu nghiệm

Ước lượng Bayes đầy đủ của tham số θ thường được lấy là **kỳ vọng hậu nghiệm (posterior mean)**:

$$\hat{\theta}_{\text{Bayes}} = E[\theta|D] = \int \theta P(\theta|D) d\theta \tag{4.26}$$

Ngoài ra, có thể chọn **trung vị hậu nghiệm** hoặc **chế độ hậu nghiệm (MAP - Maximum A Posteriori)**:

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} P(\theta|D) \tag{4.27}$$

Nếu prior là phân bố không thông tin (non-informative prior) hoặc khi kích thước mẫu lớn, thì ước lượng Bayes thường hội tụ về Ước lượng hợp lý tối đa (MLE).

* Ví dụ: Ước lượng Bayes với tham số của phân phối Gaussian

Giả sử dữ liệu $D = \{x_1, x_2, ..., x_n\}$ được lấy mẫu từ phân phối Gaussian:

$$x_i \sim \mathcal{N}(\theta, \sigma^2) \tag{4.28}$$

với phương sai σ^2 đã biết. Ta muốn ước lượng tham số θ theo phương pháp Bayes.

* Bước 1: Chọn phân bố tiên nghiệm

Giả sử ta chọn prior của θ là một phân phối Gaussian:

$$\theta \sim \mathcal{N}(\mu_0, \sigma_0^2) \tag{4.29}$$

* Bước 2: Tính xác suất hậu nghiệm

Hàm hợp lý của dữ liệu là:

$$P(D|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x_i - \theta)^2}{2\sigma^2}\right) \tag{4.30}$$

Theo định lý Bayes, xác suất hậu nghiệm của θ là:

$$P(\theta|D) \propto P(D|\theta)P(\theta) \tag{4.31}$$

Lấy log của xác suất hậu nghiệm và tối đa hóa theo θ, ta thu được:

$$\hat{\theta}_{\text{Bayes}} = \frac{\frac{\mu_0}{\sigma_0^2} + \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}} \tag{4.32}$$

**** Nhận xét:**

- $\hat{\theta}_{\text{Bayes}}$ là trung bình có trọng số giữa giá trị tiên nghiệm μ_0 và trung bình mẫu của dữ liệu.
- Nếu prior không có thông tin (tức là $\sigma_0^2 \to \infty$), thì $\hat{\theta}_{\text{Bayes}}$ hội tụ về MLE:

$$\hat{\theta}_{\text{MLE}} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{4.33}$$

d. So sánh với các phương pháp khác

Tiêu chí	MAP	Bayes (Kỳ vọng hậu nghiệm)	MLE
Cách chọn	$\arg\max P(\theta\|D)$	$E[\theta\|D]$	$\arg\max P(D\|\theta)$
Ảnh hưởng của prior	Có	Có (mạnh hơn MAP)	Không
Khi $n \to \infty$	Hội tụ về MLE	Hội tụ về MLE	Chính là MLE

Bảng 4.1: So sánh các phương pháp ước lượng

e. Ứng dụng thực tế

- **Học máy**: Naive Bayes, Gaussian Process Regression.
- **Khoa học dữ liệu**: Phân tích dữ liệu với bất định cao.
- **Y học**: Ước lượng hiệu quả của thuốc.
- **Kinh tế lượng**: Dự báo tài chính bằng mô hình Bayes.

Ước lượng Bayes đầy đủ kết hợp thông tin tiên nghiệm và dữ liệu quan sát để đưa ra kết quả chính xác hơn so với phương pháp MLE. Tuy nhiên, tính toán có thể phức tạp và cần sử dụng phương pháp xấp xỉ như MCMC.

4.2 Kiểm định giả thuyết thống kê (Hypothesis Testing)

4.2.1 Giá trị p (p-value)

a. Định nghĩa p-value

Trong thống kê, p-value (giá trị p) là xác suất thu được kết quả ít nhất cực đoan như quan sát thực tế, giả sử rằng giả thuyết không (null hypothesis, H_0) là đúng.

Công thức toán học: p-value được định nghĩa là:

$$p = P(T \geq T_{obs}|H_0) \tag{4.34}$$

trong đó:

- T là thống kê kiểm định,
- T_{obs} là giá trị thống kê kiểm định tính toán từ mẫu dữ liệu,
- H_0 là giả thuyết không.

b. Cách tính p-value

Cách tính p-value phụ thuộc vào loại kiểm định:

- **Kiểm định một phía:** p-value được tính bằng xác suất của phần đuôi của phân phối thống kê kiểm định vượt quá giá trị quan sát được.

$$p = P(T \geq T_{obs}|H_0) \quad \text{hoặc} \quad p = P(T \leq T_{obs}|H_0) \tag{4.35}$$

- **Kiểm định hai phía:** p-value là tổng của hai xác suất đuôi của phân phối thống kê kiểm định:

$$p = 2\min\{P(T \geq T_{obs}|H_0), P(T \leq T_{obs}|H_0)\} \tag{4.36}$$

c. Ý nghĩa của p-value

- Nếu p-value nhỏ hơn mức ý nghĩa α (thường là 0.05 hoặc 0.01), bác bỏ giả thuyết không H_0.
- Nếu p-value lớn hơn mức ý nghĩa α, không đủ bằng chứng để bác bỏ H_0.

*** Ví dụ minh họa**

Giả sử chúng ta thực hiện kiểm định giả thuyết với thống kê kiểm định tuân theo phân phối chuẩn chuẩn hóa $N(0,1)$ và giá trị quan sát được là $T_{obs} = 2.1$. Khi đó:

$$p = P(Z \geq 2.1) = 1 - \Phi(2.1) \approx 0.0179 \tag{4.37}$$

Nếu chọn mức ý nghĩa $\alpha = 0.05$, ta bác bỏ giả thuyết không.

4.2.2 Kiểm định giả thuyết về hệ số hồi quy/Kiểm định t (t-test)

a. Giới thiệu về kiểm định t trong hồi quy

Trong mô hình hồi quy tuyến tính tổng quát:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i \tag{4.38}$$

Phương pháp bình phương nhỏ nhất (OLS) được sử dụng để ước lượng các hệ số hồi quy $\hat{\beta}_j$. Sau khi ước lượng, ta cần kiểm định xem các hệ số này có ý nghĩa thống kê hay không. Kiểm định t (t-test) được sử dụng để đánh giá xem một hệ số hồi quy β_j có khác 0 một cách có ý nghĩa thống kê hay không.

b. Xây dựng giả thuyết kiểm định

Với mỗi hệ số hồi quy β_j, ta có giả thuyết kiểm định:

- **Giả thuyết không (H_0)**: Hệ số hồi quy không có ý nghĩa thống kê.

$$H_0 : \beta_j = 0 \tag{4.39}$$

- **Giả thuyết đối (H_1)**: Hệ số hồi quy có ý nghĩa thống kê.

$$H_1 : \beta_j \neq 0 \tag{4.40}$$

c. Thống kê kiểm định t

Thống kê kiểm định t được tính theo công thức:

$$t_j = \frac{\hat{\beta}_j - 0}{SE(\hat{\beta}_j)} \tag{4.41}$$

trong đó:

- $\hat{\beta}_j$ là ước lượng của hệ số hồi quy β_j.
- $SE(\hat{\beta}_j)$ là sai số chuẩn của $\hat{\beta}_j$, được tính bởi:

$$SE(\hat{\beta}_j) = \sqrt{\sigma^2 (X'X)^{-1}_{jj}} \tag{4.42}$$

với σ^2 là phương sai của sai số ngẫu nhiên, ước lượng bởi:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{n - k - 1} \tag{4.43}$$

trong đó:

  - $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$ là phần dư của mô hình hồi quy.
  - n là số quan sát.
  - k là số biến độc lập trong mô hình (không tính hằng số).

d. Phân phối của thống kê t

Thống kê kiểm định t tuân theo phân phối **Student t** với $n-k-1$ **bậc tự do**.

- Nếu kích thước mẫu n lớn ($n > 30$), phân phối t gần với phân phối chuẩn $N(0, 1)$.
- Nếu n nhỏ, ta phải sử dụng bảng phân phối t để tìm giá trị tới hạn.

e. Quy tắc ra quyết định

Với mức ý nghĩa α (thường chọn $\alpha = 0.05$ hoặc $\alpha = 0.01$), ta xác định giá trị tới hạn $t_{\alpha/2,n-k-1}$ từ bảng phân phối t.

- Nếu $|t_j| > t_{\alpha/2,n-k-1}$, ta bác bỏ giả thuyết $H_0 \Rightarrow$ Kết luận rằng β_j có ý nghĩa thống kê.
- Nếu $|t_j| \leq t_{\alpha/2,n-k-1}$, ta không đủ cơ sở bác bỏ $H_0 \Rightarrow$ Kết luận rằng không có đủ bằng chứng để khẳng định β_j khác 0.

Ngoài ra, ta cũng có thể sử dụng **p-value**:

- Nếu **p-value** $< \alpha \Rightarrow$ bác bỏ H_0, hệ số có ý nghĩa.
- Nếu **p-value** $\geq \alpha \Rightarrow$ không bác bỏ H_0, hệ số không có ý nghĩa.

*** Ví dụ minh họa**

Giả sử ta có mô hình hồi quy:

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon \tag{4.44}$$

với ước lượng OLS thu được:

$$\hat{\beta}_1 = 2.5, \quad SE(\hat{\beta}_1) = 0.8 \tag{4.45}$$

Số quan sát $n = 25$, số biến $k = 1$, nên bậc tự do $df = 25 - 1 - 1 = 23$.
Thống kê t:

$$t_1 = \frac{2.5}{0.8} = 3.125 \tag{4.46}$$

Tra bảng phân phối t với $df = 23$ và $\alpha = 0.05$, ta có giá trị tới hạn:

$$t_{0.025,23} \approx 2.069 \tag{4.47}$$

Vì $3.125 > 2.069$, ta bác bỏ $H_0 \Rightarrow$ Kết luận rằng β_1 có ý nghĩa thống kê ở mức $\alpha = 0.05$.

Kiểm định t giúp đánh giá mức độ ảnh hưởng của từng biến độc lập trong mô hình hồi quy tuyến tính. Khi sử dụng kiểm định này, cần đảm bảo các giả định của OLS được thỏa mãn để đảm bảo tính chính xác của kết quả kiểm định.

4.2.3 Kiểm định F

a. Giới thiệu về kiểm định F

Kiểm định F được sử dụng để kiểm tra xem toàn bộ mô hình hồi quy có ý nghĩa thống kê hay không. Cụ thể, nó kiểm định giả thuyết rằng tất cả các hệ số hồi quy (ngoại trừ hằng số) đều bằng 0.
Giả thuyết kiểm định:

- H_0: Các hệ số hồi quy không có ý nghĩa thống kê, tức là $\beta_1 = \beta_2 = \cdots = \beta_k = 0$.
- H_1: Ít nhất một trong các hệ số hồi quy khác 0.

Nếu bác bỏ H_0, ta kết luận rằng ít nhất một biến độc lập có ảnh hưởng đáng kể đến biến phụ thuộc.

b. Công thức kiểm định F

Thống kê F được tính bằng công thức:

$$F = \frac{\left(\frac{SST-SSR}{k}\right)}{\left(\frac{SSR}{n-k-1}\right)} \tag{4.48}$$

Trong đó:

- SST (Total Sum of Squares) là tổng bình phương tổng thể:

$$SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \tag{4.49}$$

- SSR (Sum of Squared Residuals) là tổng bình phương phần dư:

$$SSR = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{4.50}$$

- SSE (Sum of Squares for Regression) là tổng bình phương hồi quy:

$$SSE = SST - SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \tag{4.51}$$

Mô hình có:

- n là số quan sát.
- k là số biến độc lập.
- $n - k - 1$ là bậc tự do của phần dư.

c. Phân phối của thống kê F

Thống kê F tuân theo phân phối F với hai bậc tự do:

- $df_1 = k$ (số biến độc lập).
- $df_2 = n - k - 1$ (số quan sát trừ đi số tham số cần ước lượng).

Nếu giá trị **p-value** nhỏ hơn mức ý nghĩa α (thường là 0.05), ta bác bỏ giả thuyết H_0 và kết luận rằng mô hình có ý nghĩa thống kê.

d. Ý nghĩa của kiểm định F

- Nếu giá trị F lớn và p-value nhỏ, mô hình có ý nghĩa tổng thể.
- Nếu giá trị F nhỏ và p-value lớn, mô hình không có ý nghĩa, tức là biến độc lập không giải thích được biến phụ thuộc.

4.2.4 Kiểm định hiện tượng phương sai sai số thay đổi (heteroskedasticity)

a. Giới thiệu

Trong mô hình hồi quy tuyến tính, giả định phương sai của sai số là không đổi (homoskedasticity). Tuy nhiên, nếu phương sai của sai số thay đổi theo biến độc lập, ta có hiện tượng phương sai sai số thay đổi (heteroskedasticity). Điều này có thể dẫn đến ước lượng không hiệu quả trong hồi quy OLS và ảnh hưởng đến kiểm định giả thuyết.

b. Mô hình toán học

Xét mô hình hồi quy tuyến tính:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i, \tag{4.52}$$

với giả định:

$$E(\varepsilon_i) = 0, \quad \text{Var}(\varepsilon_i) = \sigma_i^2. \tag{4.53}$$

Nếu σ_i^2 không phải là hằng số mà phụ thuộc vào một hoặc nhiều biến độc lập, ta có heteroskedasticity:

$$\sigma_i^2 = g(x_{i1}, x_{i2}, \ldots, x_{ik}). \tag{4.54}$$

c. Các kiểm định phương sai sai số thay đổi

*** Kiểm định Breusch-Pagan**

Kiểm định Breusch-Pagan kiểm tra xem phương sai của sai số có phụ thuộc vào biến độc lập không. Cụ thể, hồi quy phần dư bình phương $\hat{\varepsilon}_i^2$ theo biến độc lập:

$$\hat{\varepsilon}_i^2 = \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \cdots + \gamma_k x_{ik} + u_i. \tag{4.55}$$

Sau đó, sử dụng thống kê kiểm định LM:

$$LM = nR^2 \sim \chi_k^2, \tag{4.56}$$

trong đó R^2 là hệ số xác định từ hồi quy trên. Nếu giá trị LM lớn hơn giá trị tới hạn, bác bỏ giả thuyết không có heteroskedasticity.

*** Kiểm định White**

Kiểm định White kiểm tra heteroskedasticity mà không giả định cấu trúc cụ thể của phương sai sai số. Hồi quy phần dư bình phương theo tất cả các biến độc lập và bình phương của chúng:

$$\hat{\varepsilon}_i^2 = \gamma_0 + \sum_{j=1}^{k} \gamma_j x_{ij} + \sum_{j=1}^{k} \sum_{m=j}^{k} \gamma_{jm} x_{ij} x_{im} + u_i. \tag{4.57}$$

Thống kê kiểm định tương tự như Breusch-Pagan:

$$LM = nR^2 \sim \chi_m^2, \tag{4.58}$$

trong đó m là số bậc tự do.

Hiện tượng phương sai sai số thay đổi làm ảnh hưởng đến tính hiệu quả của ước lượng OLS. Các kiểm định như Breusch-Pagan và White giúp phát hiện heteroskedasticity để điều chỉnh mô hình phù hợp.

4.2.5 Kiểm định tự tương quan

a. Giới thiệu

Tự tương quan (autocorrelation) là hiện tượng khi các sai số trong mô hình hồi quy không độc lập với nhau. Điều này thường xuất hiện trong dữ liệu chuỗi thời gian hoặc khi có yếu tố hệ thống chưa được mô hình hóa đúng.

b. Mô hình hồi quy tuyến tính

Giả sử mô hình hồi quy tuyến tính:

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_k x_{kt} + \varepsilon_t, \tag{4.59}$$

trong đó ε_t là sai số ngẫu nhiên.

Nếu tồn tại tự tương quan, ta có:

$$Cov(\varepsilon_t, \varepsilon_{t-1}) \neq 0. \tag{4.60}$$

c. Kiểm định Durbin-Watson

Một trong những kiểm định phổ biến để phát hiện tự tương quan bậc nhất là kiểm định Durbin-Watson (DW). Thống kê kiểm định DW được tính như sau:

$$DW = \frac{\sum_{t=2}^{n} (\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{t=1}^{n} \varepsilon_t^2}. \tag{4.61}$$

Giá trị DW nằm trong khoảng $[0, 4]$ và được diễn giải như sau: - $DW \approx 2$: Không có tự tương quan bậc nhất. - $DW < 2$: Có tự tương quan dương. - $DW > 2$: Có tự tương quan âm.

d. Kiểm định Breusch-Godfrey

Kiểm định Durbin-Watson chỉ áp dụng cho tự tương quan bậc nhất. Để kiểm định tự tương quan bậc cao hơn, ta sử dụng kiểm định Breusch-Godfrey:

1. Hồi quy mô hình gốc và lấy phần dư $\hat{\varepsilon}_t$. 2. Hồi quy phần dư theo chính nó:

$$\hat{\varepsilon}_t = \rho_1 \hat{\varepsilon}_{t-1} + \rho_2 \hat{\varepsilon}_{t-2} + \cdots + \rho_p \hat{\varepsilon}_{t-p} + \eta_t. \tag{4.62}$$

3. Kiểm định giả thuyết:

- H_0: Không có tự tương quan bậc p.
- H_1: Có tự tương quan bậc p.

Kiểm định dựa trên thống kê LM:

$$LM = nR^2 \sim \chi^2_p, \tag{4.63}$$

trong đó R^2 là hệ số xác định từ hồi quy phụ.

Nếu phát hiện tự tương quan, có thể sử dụng các phương pháp như mô hình sai số tổng quát (GLS) hoặc ước lượng Newey-West để hiệu chỉnh phương sai của ước lượng.

Phần II

Mô hình hồi quy tuyến tính

Chương 5

Mô hình hồi quy tuyến tinh đơn giản (Simple Linear Regression)

5.1 Giới thiệu

Mô hình hồi quy tuyến tính đơn giản là một trong những mô hình cơ bản nhất trong kinh tế lượng, được sử dụng để mô tả mối quan hệ giữa một biến phụ thuộc (Y) và một biến độc lập (X).

5.2 Phương trình tổng quát

Phương trình hồi quy tuyến tính đơn giản có dạng:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, 2, ..., n \tag{5.1}$$

Trong đó:

- Y_i: Giá trị của biến phụ thuộc tại quan sát thứ i.
- X_i: Giá trị của biến độc lập tại quan sát thứ i.
- β_0: Hệ số chặn (intercept).
- β_1: Hệ số hồi quy (slope coefficient).
- ε_i: Sai số ngẫu nhiên, phản ánh các yếu tố không quan sát được.

5.3 Giả định của mô hình hồi quy tuyến tính

Để ước lượng chính xác, mô hình cần thỏa mãn các giả định sau:

1. **Tính tuyến tính**: Mối quan hệ giữa X và Y là tuyến tính.
2. **Độc lập của sai số**: Các sai số ε_i là độc lập với nhau.
3. **Phân phối chuẩn của sai số**: $\varepsilon_i \sim N(0, \sigma^2)$.

4. **Phương sai không đổi** (Homoskedasticity): $Var(\varepsilon_i) = \sigma^2$, không phụ thuộc vào X.
5. **Không có đa cộng tuyến**: Không có sự tương quan hoàn hảo giữa các biến độc lập (trong trường hợp mô hình mở rộng nhiều biến).

5.4 Ước lượng tham số bằng phương pháp bình phương nhỏ nhất (OLS)

Mục tiêu của OLS là tìm các giá trị $\hat{\beta}_0$ và $\hat{\beta}_1$ sao cho tổng bình phương sai số (RSS) nhỏ nhất:

$$RSS = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2 \tag{5.2}$$

Để tìm cực tiểu của RSS, ta giải hệ phương trình đạo hàm:

$$\frac{\partial RSS}{\partial \beta_0} = -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i) = 0 \tag{5.3}$$

$$\frac{\partial RSS}{\partial \beta_1} = -2\sum_{i=1}^{n} X_i(Y_i - \beta_0 - \beta_1 X_i) = 0 \tag{5.4}$$

Giải hệ phương trình trên, ta có các ước lượng:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{5.5}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \tag{5.6}$$

Trong đó $\bar{X}$ và $\bar{Y}$ lần lượt là trung bình của X và Y.

5.5 Tính chất của ước lượng OLS

Dưới các giả định của mô hình, các ước lượng OLS có các tính chất:

- Không chệch: $E(\hat{\beta}_0) = \beta_0$, $E(\hat{\beta}_1) = \beta_1$.
- Hiệu quả: Có phương sai nhỏ nhất trong lớp các ước lượng tuyến tính không chệch (BLUE - Best Linear Unbiased Estimator).
- Phân phối của $\hat{\beta}_1$ là chuẩn: $\hat{\beta}_1 \sim N(\beta_1, \sigma^2 / \sum(X_i - \bar{X})^2)$.

5.6 Đánh giá độ phù hợp của mô hình

Hệ số xác định (R^2) đo lường mức độ giải thích của biến độc lập đối với biến phụ thuộc:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \tag{5.7}$$

R^2 nằm trong khoảng $[0, 1]$, giá trị càng cao mô hình càng phù hợp.

5.7 Kết luận

Mô hình hồi quy tuyến tính đơn giản là công cụ quan trọng trong phân tích kinh tế lượng. Bằng cách sử dụng phương pháp OLS, ta có thể ước lượng các tham số và đánh giá độ phù hợp của mô hình một cách chính xác.

Chương 6

Mô hình hồi quy tuyến tính với biến tiên lượng phân nhóm

6.1 Giới thiệu

Mô hình hồi quy tuyến tính với biến tiên lượng phân nhóm (Grouped Predictor Linear Regression Model) được sử dụng khi một biến giải thích có thể được chia thành các nhóm khác nhau, mỗi nhóm có thể có ảnh hưởng khác nhau đến biến phụ thuộc.

6.2 Mô hình toán học

Giả sử mô hình hồi quy tổng quát có dạng:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i, \tag{6.1}$$

trong đó:

- Y_i là biến phụ thuộc (kết quả quan sát được).
- X_{ij} là biến tiên lượng (predictor variables), với $j = 1, 2, \ldots, p$.
- $\beta_0, \beta_1, \ldots, \beta_p$ là các tham số cần ước lượng.
- ε_i là nhiễu ngẫu nhiên có phân phối $\mathcal{N}(0, \sigma^2)$.

Trong trường hợp biến tiên lượng có thể phân nhóm, ta sử dụng biến giả (dummy variables) để đại diện cho các nhóm:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_1 G_{i1} + \gamma_2 G_{i2} + \cdots + \gamma_k G_{ik} + \varepsilon_i, \tag{6.2}$$

trong đó:

- G_{ij} là biến giả, nhận giá trị 1 nếu quan sát thuộc nhóm j và 0 nếu không.
- γ_j thể hiện tác động của nhóm j lên biến phụ thuộc.

6.3 Ước lượng tham số

Các tham số của mô hình được ước lượng bằng phương pháp bình phương nhỏ nhất (OLS). Hệ số hồi quy $\hat{\beta}$ được tính bằng công thức:

$$\hat{\beta} = (X^T X)^{-1} X^T Y, \tag{6.3}$$

trong đó:

- X là ma trận thiết kế chứa các biến tiên lượng và biến giả,
- Y là vector quan sát của biến phụ thuộc.

6.4 Kiểm định ý nghĩa

Để kiểm tra xem nhóm có ảnh hưởng đến biến phụ thuộc hay không, ta thực hiện kiểm định giả thuyết:

$$H_0 : \gamma_1 = \gamma_2 = \cdots = \gamma_k = 0. \tag{6.4}$$

Sử dụng kiểm định F:

$$F = \frac{(RSS_r - RSS_{ur})/k}{RSS_{ur}/(n-p-k)}, \tag{6.5}$$

trong đó:

- RSS_r là tổng bình phương phần dư của mô hình bị ràng buộc (không có biến phân nhóm),
- RSS_{ur} là tổng bình phương phần dư của mô hình đầy đủ,
- n là số quan sát, p là số biến tiên lượng không phân nhóm.

Nếu giá trị F đủ lớn, ta bác bỏ H_0, kết luận rằng nhóm có ảnh hưởng đến biến phụ thuộc.

6.5 Ứng dụng thực tiễn

Mô hình này được áp dụng trong nhiều lĩnh vực:

- Phân tích mức lương theo nhóm ngành nghề.
- Đánh giá hiệu quả của các chiến lược marketing theo khu vực địa lý.
- Dự báo nhu cầu tiêu dùng dựa trên phân nhóm thu nhập.

Chương 7

Mô hình hồi quy đa biến

7.1 Định nghĩa Mô hình hồi quy đa biến

Mô hình hồi quy đa biến mở rộng từ mô hình hồi quy tuyến tính đơn giản bằng cách sử dụng nhiều biến giải thích (predictors) thay vì chỉ một. Công thức tổng quát của mô hình có dạng:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \tag{7.1}$$

trong đó:

- Y là biến phụ thuộc (response variable).
- $X_1, X_2, ..., X_p$ là các biến độc lập (predictor variables).
- β_0 là hằng số chặn (intercept).
- $\beta_1, \beta_2, ..., \beta_p$ là các hệ số hồi quy (regression coefficients).
- ε là sai số ngẫu nhiên (random error).

Dưới dạng ma trận, mô hình có thể viết là:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{7.2}$$

7.2 Ước lượng tham số bằng phương pháp bình phương tối thiểu (OLS)

Mục tiêu của hồi quy là ước lượng các hệ số $\boldsymbol{\beta}$ sao cho sai số bình phương tổng (SSE) nhỏ nhất:

$$SSE = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \tag{7.3}$$

Dùng phương pháp bình phương tối thiểu (OLS), ước lượng của $\boldsymbol{\beta}$ là:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{7.4}$$

7.3 Đánh giá mô hình

7.3.1 Độ phù hợp của mô hình - Hệ số xác định R^2

Hệ số xác định R^2 đo lường mức độ giải thích của mô hình đối với phương sai của biến phụ thuộc:

$$R^2 = 1 - \frac{SSE}{SST} \tag{7.5}$$

trong đó:

- $SST = \sum(Y_i - \bar{Y})^2$ là tổng phương sai tổng cộng (Total Sum of Squares).
- $SSE = \sum(Y_i - \hat{Y})^2$ là tổng phương sai sai số (Residual Sum of Squares).

Nếu R^2 càng gần 1, mô hình càng giải thích tốt phương sai của Y.

7.3.2 Kiểm định ý nghĩa của từng hệ số hồi quy

Để kiểm tra xem một biến X_j có ảnh hưởng thống kê đến Y hay không, ta kiểm định giả thuyết:

$$\begin{aligned} H_0 &: \beta_j = 0 \quad \text{(không có tác động)} \\ H_1 &: \beta_j \neq 0 \quad \text{(có tác động)} \end{aligned}$$

Dùng kiểm định t:

$$t_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \tag{7.6}$$

với $SE(\hat{\beta}_j)$ là sai số chuẩn của ước lượng hệ số. Giá trị t_j này được so sánh với phân phối t để ra quyết định.

7.3.3 Kiểm định tổng thể mô hình (Kiểm định F)

Kiểm định giả thuyết:

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = \cdots = \beta_p = 0 \quad \text{(không có biến nào có tác động)} \\ H_1 &: \text{ít nhất một hệ số } \beta_j \neq 0 \end{aligned}$$

Dùng kiểm định tổng thể F:

$$F = \frac{SST - SSE}{SSE} \times \frac{n - p - 1}{p} \tag{7.7}$$

so sánh với phân phối $F_{p,n-p-1}$ để quyết định giữ hay bác bỏ H_0.

7.4 Giả định của mô hình hồi quy đa biến

1. Tuyến tính: Y có quan hệ tuyến tính với các biến độc lập.
2. Không có đa cộng tuyến: Các biến độc lập không có quan hệ tuyến tính quá mạnh (có thể kiểm tra bằng hệ số VIF).
3. Phân phối chuẩn của sai số: $\varepsilon \sim N(0, \sigma^2)$ (có thể kiểm tra bằng biểu đồ Q-Q plot).
4. Phương sai không đổi: $\text{Var}(\varepsilon) = \sigma^2$, có thể kiểm tra bằng Breusch-Pagan test.
5. Không có tự tương quan: Các sai số không được phụ thuộc nhau (có thể kiểm tra bằng Durbin-Watson test).

7.5 Ứng dụng thực tế

Mô hình hồi quy đa biến được áp dụng trong:

- Kinh tế học: Dự báo tăng trưởng GDP dựa trên các yếu tố như lãi suất, tỷ lệ thất nghiệp, đầu tư.
- Tài chính: Định giá cổ phiếu dựa trên các biến như thu nhập công ty, lãi suất thị trường.
- Marketing: Dự đoán doanh số bán hàng dựa trên giá sản phẩm, ngân sách quảng cáo, mùa vụ.
- Y tế: Dự báo nguy cơ mắc bệnh dựa trên yếu tố như chỉ số BMI, huyết áp, độ tuổi.

7.6 Mở rộng mô hình

- Hồi quy phi tuyến: Dùng biến đa thức hoặc hàm logarit.
- Hồi quy Ridge & Lasso: Để xử lý đa cộng tuyến bằng cách thêm điều chuẩn.
- Hồi quy logistic: Dùng cho bài toán phân loại (biến phụ thuộc là nhị phân).

Chương 8

Mô hình hồi quy đa thức

8.1 Giới thiệu

Mô hình hồi quy đa thức là một mở rộng của mô hình hồi quy tuyến tính, trong đó quan hệ giữa biến phụ thuộc Y và biến độc lập X được mô tả bằng một đa thức bậc p thay vì một đường thẳng.

8.2 Mô hình toán học

Mô hình hồi quy đa thức bậc p có dạng tổng quát như sau:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_p X^p + \varepsilon, \tag{8.1}$$

trong đó:

- Y: Biến phụ thuộc,
- X: Biến độc lập,
- $\beta_0, \beta_1, \ldots, \beta_p$: Các tham số hồi quy cần ước lượng,
- ε: Sai số ngẫu nhiên, giả định $\varepsilon \sim N(0, \sigma^2)$.

8.3 Ước lượng tham số

Các tham số $\beta_0, \beta_1, \ldots, \beta_p$ có thể được ước lượng bằng phương pháp bình phương tối thiểu (OLS). Ta xây dựng hàm mất mát:

$$L(\beta_0, \beta_1, \ldots, \beta_p) = \sum_{i=1}^{n} \left(Y_i - (\beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_p X_i^p) \right)^2. \tag{8.2}$$

Để tìm nghiệm tối ưu, ta giải hệ phương trình đạo hàm bậc nhất:

$$\frac{\partial L}{\partial \beta_j} = 0, \quad j = 0, 1, \ldots, p. \tag{8.3}$$

8.4 Đánh giá mô hình

8.4.1 Hệ số xác định R^2

Hệ số xác định R^2 được tính như sau:

$$R^2 = 1 - \frac{SSE}{SST}, \tag{8.4}$$

trong đó:

- $SST = \sum(Y_i - \bar{Y})^2$ là tổng phương sai tổng,
- $SSE = \sum(Y_i - \hat{Y}_i)^2$ là tổng phương sai sai số.

8.4.2 Kiểm định ý nghĩa của mô hình

Kiểm định tổng thể mô hình sử dụng kiểm định F với giả thuyết:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0, \quad \text{(không có tác động)}$$
$$H_1 : \text{Ít nhất một } \beta_j \neq 0, \quad \text{(có tác động)}$$

Thống kê kiểm định F được tính như sau:

$$F = \frac{SST - SSE}{p} \div \frac{SSE}{n - p - 1}. \tag{8.5}$$

So sánh với phân phối $F_{p,n-p-1}$ để quyết định giữ hay bác bỏ H_0.

8.5 Kiểm tra giả định

- Kiểm tra phân phối của sai số: $\varepsilon \sim N(0, \sigma^2)$ (có thể kiểm tra bằng biểu đồ Q-Q plot).
- Kiểm tra phương sai không đổi: $\text{Var}(\varepsilon) = \sigma^2$, có thể kiểm tra bằng Breusch-Pagan test.

Chương 9

Mô hình hồi quy vững chắc (Robust Regression)

9.1 Giới thiệu

Hồi quy vững chắc (Robust Regression) là một phương pháp ước lượng mô hình hồi quy tuyến tính khi dữ liệu chứa các ngoại lệ (outliers) hoặc vi phạm giả định phân phối chuẩn của phần dư.

9.2 Hàm mất mát và phương pháp ước lượng

Hồi quy vững chắc sử dụng một hàm mất mát (loss function) thay thế cho phương pháp bình phương tối thiểu thông thường (OLS). Tổng quát, bài toán tối ưu có dạng:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \rho(y_i - X_i\beta) \tag{9.1}$$

trong đó $\rho(\cdot)$ là một hàm mất mát vững chắc.

9.3 Các phương pháp hồi quy vững chắc

9.3.1 Hồi quy Huber

Hồi quy Huber sử dụng một hàm mất mát kết hợp giữa bình phương lỗi nhỏ và giá trị tuyệt đối cho lỗi lớn:

$$\rho_H(r) = \begin{cases} r^2/2, & \text{nếu } |r| \leq c \\ c(|r| - c/2), & \text{nếu } |r| > c \end{cases} \tag{9.2}$$

trong đó c là một ngưỡng xác định.

9.3.2 Hồi quy Tukey

Hồi quy Tukey sử dụng một hàm mất mát chặn giá trị của các ngoại lệ lớn:

$$\rho_T(r) = \begin{cases} c^2 \left(1 - \left(1 - \frac{r^2}{c^2}\right)^3\right)/6, & \text{nếu } |r| \leq c \\ c^2/6, & \text{nếu } |r| > c \end{cases} \tag{9.3}$$

9.4 Thuật toán IRLS

Hồi quy vững chắc thường được ước lượng bằng thuật toán Bình phương tối thiểu có trọng số lặp lại (Iteratively Reweighted Least Squares - IRLS):

1. Khởi tạo $\beta^{(0)}$ từ hồi quy OLS.
2. Tại bước lặp thứ t, tính trọng số $w_i^{(t)}$ từ đạo hàm của hàm mất mát:

$$w_i^{(t)} = \frac{\psi(r_i^{(t)})}{r_i^{(t)}} \tag{9.4}$$

3. Giải bài toán hồi quy có trọng số:

$$\beta^{(t+1)} = \arg\min_{\beta} \sum_{i=1}^{n} w_i^{(t)} (y_i - X_i\beta)^2 \tag{9.5}$$

4. Lặp lại cho đến khi $\beta^{(t)}$ hội tụ.

9.5 Kết luận

Hồi quy vững chắc cung cấp một giải pháp hiệu quả để giảm thiểu tác động của ngoại lệ trong mô hình hồi quy tuyến tính. Phương pháp này giúp cải thiện tính ổn định của ước lượng khi dữ liệu vi phạm giả định chuẩn hoặc có giá trị ngoại lai.

Chương 10

Mô hình hồi quy đa biến đa thức (Multivariate Polynomial Regression)

10.1 Giới thiệu

Mô hình hồi quy đa biến đa thức (Multivariate Polynomial Regression) là một mở rộng của hồi quy tuyến tính đa biến, trong đó ta đưa vào các bậc cao hơn của biến độc lập để mô hình hóa quan hệ phi tuyến giữa biến phụ thuộc và các biến độc lập.

10.2 Mô hình Toán học

Giả sử ta có một tập dữ liệu gồm n quan sát với p biến độc lập $X_1, X_2, \ldots, X_p$, mô hình hồi quy đa biến đa thức bậc d có dạng tổng quát như sau:

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \sum_{j=1}^{p} \sum_{k=j}^{p} \beta_{jk} X_j X_k + \sum_{j=1}^{p} \sum_{k=j}^{p} \sum_{l=k}^{p} \beta_{jkl} X_j X_k X_l + \cdots + \varepsilon \quad (10.1)$$

trong đó:

- Y là biến phụ thuộc,
- X_j là các biến độc lập,
- $\beta_0, \beta_j, \beta_{jk}, \beta_{jkl}, \ldots$ là các hệ số hồi quy cần ước lượng,
- $\varepsilon \sim N(0, \sigma^2)$ là sai số ngẫu nhiên.

10.3 Ma trận Thiết kế

Để biểu diễn mô hình dưới dạng ma trận, ta có thể định nghĩa ma trận thiết kế X như sau:

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} & x_{11}^2 & x_{12}^2 & \dots & x_{1p}^2 \\ 1 & x_{21} & x_{22} & \dots & x_{2p} & x_{21}^2 & x_{22}^2 & \dots & x_{2p}^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} & x_{n1}^2 & x_{n2}^2 & \dots & x_{np}^2 \end{bmatrix} \tag{10.2}$$

Khi đó, mô hình có thể viết gọn lại dưới dạng:

$$Y = X\beta + \varepsilon \tag{10.3}$$

với:

- $Y = (Y_1, Y_2, \dots, Y_n)^T$ là vector các giá trị quan sát,
- β là vector hệ số hồi quy,
- ε là vector sai số ngẫu nhiên.

10.4 Ước lượng tham số

Hệ số hồi quy β có thể được ước lượng bằng phương pháp bình phương nhỏ nhất (OLS - Ordinary Least Squares):

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{10.4}$$

Điều này đòi hỏi $X^T X$ khả nghịch, nếu không, ta có thể sử dụng hồi quy Ridge hoặc các phương pháp điều chuẩn khác.

10.5 Đánh giá Mô hình

Các tiêu chí đánh giá mô hình bao gồm:

- Hệ số xác định R^2: đo lường mức độ giải thích của mô hình đối với biến phụ thuộc.
- Kiểm định F: kiểm tra ý nghĩa tổng thể của mô hình.
- Kiểm định t: kiểm tra ý nghĩa của từng hệ số hồi quy.
- Phân tích phần dư: kiểm tra giả định về sai số.

10.6 Kết luận

Mô hình hồi quy đa biến đa thức là một công cụ hữu ích để mô hình hóa các quan hệ phi tuyến giữa biến phụ thuộc và các biến độc lập. Tuy nhiên, cần chú ý đến vấn đề đa cộng tuyến khi sử dụng các bậc cao của biến độc lập.

Phần III

Mô hình hồi quy phi tuyến

Chương 11

Mô hình hồi quy hàm mũ (Exponential Regression)

11.1 Giới thiệu

Mô hình hồi quy hàm mũ được sử dụng để mô tả mối quan hệ giữa biến độc lập x và biến phụ thuộc y trong dạng hàm mũ:

$$y = ae^{bx} + \epsilon, \tag{11.1}$$

trong đó:

- y là biến phụ thuộc,
- x là biến độc lập,
- a, b là các tham số cần ước lượng,
- ϵ là sai số ngẫu nhiên.

11.2 Biến đổi tuyến tính

Để ước lượng tham số a và b, ta thực hiện phép biến đổi logarit tự nhiên hai vế:

$$\ln y = \ln a + bx + \epsilon'. \tag{11.2}$$

Gọi $Y' = \ln y$ và $A = \ln a$, ta có mô hình hồi quy tuyến tính:

$$Y' = A + bx + \epsilon'. \tag{11.3}$$

Khi đó, các tham số A và b có thể được ước lượng bằng phương pháp bình phương tối thiểu (OLS):

$$b = \frac{n\sum x_i \ln y_i - \sum x_i \sum \ln y_i}{n\sum x_i^2 - (\sum x_i)^2}, \tag{11.4}$$

$$A = \frac{\sum \ln y_i - b\sum x_i}{n}. \tag{11.5}$$

Sau khi tìm được A, ta suy ra a bằng:

$$a = e^A. \tag{11.6}$$

11.3 Đánh giá mô hình

Mô hình có thể được đánh giá bằng hệ số xác định R^2:

$$R^2 = 1 - \frac{SSE}{SST}, \tag{11.7}$$

trong đó:

- $SST = \sum(Y_i' - \bar{Y}')^2$ là tổng phương sai tổng cộng,
- $SSE = \sum(Y_i' - \hat{Y}'_i)^2$ là tổng phương sai sai số.

11.4 Kết luận

Mô hình hồi quy hàm mũ thích hợp khi dữ liệu thể hiện sự tăng trưởng hoặc suy giảm theo cấp số nhân. Sau khi ước lượng tham số, mô hình có thể được sử dụng để dự báo giá trị tương lai của y dựa trên giá trị của x.

Chương 12

Mô hình hồi quy logarit (Logarithmic Regression)

12.1 Giới thiệu

Mô hình hồi quy logarit được sử dụng khi dữ liệu có xu hướng thay đổi nhanh ban đầu và sau đó chậm dần theo thời gian hoặc theo biến giải thích. Mô hình này phù hợp cho các hiện tượng giảm tốc hoặc có sự bão hòa.

12.2 Định nghĩa mô hình

Mô hình hồi quy logarit có dạng tổng quát:

$$Y = \beta_0 + \beta_1 \ln(X) + \varepsilon \tag{12.1}$$

trong đó:

- Y là biến phụ thuộc.
- X là biến độc lập.
- β_0, β_1 là các tham số hồi quy.
- ε là sai số ngẫu nhiên với giả thiết $\varepsilon \sim N(0, \sigma^2)$.

12.3 Ước lượng tham số

Các tham số β_0 và β_1 có thể được ước lượng bằng phương pháp bình phương nhỏ nhất (OLS). Hàm mất mát cần tối thiểu hóa là:

$$L(\beta_0, \beta_1) = \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 \ln(X_i)\right)^2 \tag{12.2}$$

Điều kiện tối ưu được tìm bằng cách giải hệ phương trình đạo hàm riêng:

$$\frac{\partial L}{\partial \beta_0} = -2\sum_{i=1}^{n}\left(Y_i - \beta_0 - \beta_1 \ln(X_i)\right) = 0 \tag{12.3}$$

$$\frac{\partial L}{\partial \beta_1} = -2\sum_{i=1}^{n}\ln(X_i)\left(Y_i - \beta_0 - \beta_1 \ln(X_i)\right) = 0 \tag{12.4}$$

Giải hệ phương trình này ta có công thức ước lượng:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\left(\ln(X_i) - \overline{\ln(X)}\right)(Y_i - \bar{Y})}{\sum_{i=1}^{n}\left(\ln(X_i) - \overline{\ln(X)}\right)^2} \tag{12.5}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \overline{\ln(X)} \tag{12.6}$$

trong đó:

- $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ là trung bình của Y.
- $\overline{\ln(X)} = \frac{1}{n}\sum_{i=1}^{n}\ln(X_i)$ là trung bình của $\ln(X)$.

12.4 Kiểm định mô hình

12.4.1 Hệ số xác định (R^2)

Mức độ phù hợp của mô hình được đánh giá bằng hệ số xác định:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \tag{12.7}$$

12.4.2 Kiểm định ý nghĩa hệ số hồi quy

Ta kiểm định giả thuyết:

$$H_0 : \beta_1 = 0 \tag{12.8}$$

$$H_1 : \beta_1 \neq 0 \tag{12.9}$$

Với thống kê kiểm định:

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \tag{12.10}$$

trong đó $SE(\hat{\beta}_1)$ là sai số chuẩn của $\hat{\beta}_1$. Giá trị này được so sánh với phân phối t để quyết định giữ hay bác bỏ H_0.

Chương 13

Mô hình hồi quy hàm Cobb-Douglas (Cobb-Douglas Regression)

13.1 Giới thiệu

Mô hình hồi quy hàm Cobb-Douglas là một dạng hồi quy phi tuyến thường được sử dụng trong kinh tế học để mô hình hóa mối quan hệ giữa sản lượng và các yếu tố đầu vào như vốn và lao động.

13.2 Biểu diễn toán học

Mô hình Cobb-Douglas có dạng tổng quát:

$$Y = AX_1^{\beta_1} X_2^{\beta_2} \dots X_k^{\beta_k} e^{\varepsilon} \tag{13.1}$$

trong đó:

- Y là biến phụ thuộc (sản lượng, đầu ra,...),
- $X_1, X_2, \dots, X_k$ là các biến độc lập (các yếu tố đầu vào như vốn, lao động,...),
- A là hằng số,
- $\beta_1, \beta_2, \dots, \beta_k$ là các tham số cần ước lượng,
- e^{ε} là thành phần sai số ngẫu nhiên.

13.3 Tuyến tính hóa mô hình

Do mô hình Cobb-Douglas có dạng phi tuyến, ta lấy log hai vế của phương trình để biến đổi về dạng hồi quy tuyến tính:

$$\ln Y = \ln A + \beta_1 \ln X_1 + \beta_2 \ln X_2 + \dots + \beta_k \ln X_k + \varepsilon \tag{13.2}$$

Khi đó, ta có thể viết lại dưới dạng:

$$Y^* = \beta_0 + \beta_1 X_1^* + \beta_2 X_2^* + \cdots + \beta_k X_k^* + \varepsilon \tag{13.3}$$

trong đó:

- $Y^* = \ln Y$, $X_i^* = \ln X_i$,
- $\beta_0 = \ln A$,
- Hồi quy tuyến tính có thể được ước lượng bằng phương pháp bình phương nhỏ nhất (OLS).

13.4 Ước lượng tham số

Sử dụng phương pháp bình phương nhỏ nhất (OLS), các tham số $\beta_0, \beta_1, \ldots, \beta_k$ được ước lượng bằng cách giải:

$$\hat{\beta} = (X^T X)^{-1} X^T Y^* \tag{13.4}$$

13.5 Kiểm định mô hình

Sau khi ước lượng, ta cần kiểm định ý nghĩa của các hệ số hồi quy thông qua kiểm định t:

$$t_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \tag{13.5}$$

trong đó $SE(\hat{\beta}_j)$ là sai số chuẩn của $\hat{\beta}_j$. Giá trị t_j được so sánh với phân phối t để quyết định giữ hay bác bỏ giả thuyết $H_0 : \beta_j = 0$.

Ngoài ra, kiểm định tổng thể mô hình sử dụng kiểm định F:

$$F = \frac{SST - SSE}{k} \Big/ \frac{SSE}{n - k - 1} \tag{13.6}$$

trong đó SST là tổng phương sai tổng và SSE là tổng phương sai sai số.

13.6 Ứng dụng thực tế

Mô hình Cobb-Douglas thường được sử dụng trong:

- Phân tích sản xuất trong kinh tế học,
- Xây dựng mô hình tăng trưởng,
- Đánh giá tác động của vốn và lao động đến GDP.

Chương 14

Mô hình hồi quy Logistic (Logistic Regression)

14.1 Giới thiệu

Mô hình hồi quy Logistic là một phương pháp thống kê được sử dụng để mô hình hóa xác suất của một biến phụ thuộc nhị phân dựa vào một hoặc nhiều biến độc lập.

14.2 Công thức tổng quát

Giả sử ta có một tập dữ liệu với n quan sát, mỗi quan sát bao gồm một biến phản hồi nhị phân $Y_i \in \{0, 1\}$ và một vector các biến độc lập $X_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$. Hồi quy logistic mô hình hóa xác suất có điều kiện của Y_i như sau:

$$P(Y_i = 1|X_i) = \pi_i = \frac{e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}}}{1 + e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}}}. \tag{14.1}$$

Lấy logit (hàm log-odds) của xác suất, ta có phương trình tuyến tính:

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}. \tag{14.2}$$

14.3 Ước lượng tham số

Các hệ số hồi quy $\beta_0, \beta_1, \ldots, \beta_p$ được ước lượng bằng phương pháp hợp lý cực đại (Maximum Likelihood Estimation - MLE). Hàm hợp lý cho toàn bộ mẫu là:

$$L(\beta) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}. \tag{14.3}$$

Lấy log của hàm hợp lý, ta có log-likelihood:

$$\ell(\beta) = \sum_{i=1}^{n} \left[y_i \log(\pi_i) + (1-y_i)\log(1-\pi_i)\right]. \tag{14.4}$$

Ước lượng $\hat{\beta}$ được tìm bằng cách giải phương trình đạo hàm bậc nhất của $\ell(\beta)$:

$$\frac{\partial \ell(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} (y_i - \pi_i) x_{ij} = 0, \quad \forall j \in \{0, 1, \ldots, p\}. \tag{14.5}$$

14.4 Kiểm định ý nghĩa mô hình

Để kiểm định ý nghĩa của các hệ số hồi quy, ta sử dụng kiểm định Wald hoặc kiểm định LR:

- **Kiểm định Wald**: Xét giả thuyết $H_0 : \beta_j = 0$ so với $H_1 : \beta_j \neq 0$. Thống kê kiểm định:

$$z_j = \frac{\hat{\beta}_j}{\mathrm{SE}(\hat{\beta}_j)} \sim N(0, 1). \tag{14.6}$$

- **Kiểm định Likelihood Ratio (LR)**: So sánh log-likelihood của mô hình đầy đủ và mô hình rút gọn:

$$G = -2 \left[\ell(\beta^{(0)}) - \ell(\hat{\beta}) \right] \sim \chi^2_p. \tag{14.7}$$

14.5 Kết luận

Mô hình hồi quy logistic là một công cụ mạnh mẽ để phân loại nhị phân, thường được sử dụng trong thống kê, học máy và các lĩnh vực khác như y tế, tài chính và khoa học xã hội.

Chương 15

Mô hình hồi quy Probit (Probit Regression)

15.1 Giới thiệu

Mô hình hồi quy Probit là một mô hình hồi quy sử dụng hàm phân phối chuẩn tích lũy để mô tả xác suất của một biến nhị phân phụ thuộc vào một hoặc nhiều biến độc lập.

15.2 Mô hình toán học

Giả sử rằng ta có một biến phản hồi nhị phân Y với giá trị $Y \in \{0, 1\}$, mô hình Probit được định nghĩa như sau:

$$P(Y = 1|X) = \Phi(X\beta) \tag{15.1}$$

trong đó:

- $\Phi(\cdot)$ là hàm phân phối tích lũy (CDF) của phân phối chuẩn chuẩn $N(0, 1)$,
- X là vector các biến độc lập,
- β là vector hệ số hồi quy.

15.3 Mô hình dạng tiềm ẩn

Mô hình Probit có thể được xem xét dưới dạng mô hình biến tiềm ẩn:

$$Y^* = X\beta + \varepsilon, \quad \varepsilon \sim N(0, 1) \tag{15.2}$$

Với:

$$Y = \begin{cases} 1, & \text{nếu } Y^* > 0 \\ 0, & \text{nếu } Y^* \leq 0 \end{cases} \tag{15.3}$$

15.4 Ước lượng tham số

Các tham số của mô hình Probit được ước lượng bằng phương pháp hợp lý cực đại (Maximum Likelihood Estimation - MLE). Hàm hợp lý được cho bởi:

$$L(\beta) = \prod_{i=1}^{n} \Phi(X_i\beta)^{Y_i}(1 - \Phi(X_i\beta))^{(1-Y_i)} \tag{15.4}$$

Và log-hàm hợp lý là:

$$\ell(\beta) = \sum_{i=1}^{n} [Y_i \log \Phi(X_i\beta) + (1 - Y_i) \log(1 - \Phi(X_i\beta))] \tag{15.5}$$

Ước lượng β được tìm bằng cách cực đại hóa hàm log-hợp lý này.

15.5 Suy diễn thống kê

Để kiểm định ý nghĩa thống kê của các hệ số hồi quy β_j, ta có thể sử dụng kiểm định Wald:

$$z_j = \frac{\hat{\beta}_j}{\text{SE}(\hat{\beta}_j)} \sim N(0, 1) \tag{15.6}$$

Hoặc kiểm định tỷ số hợp lý (Likelihood Ratio Test, LRT).

15.6 Kết luận

Mô hình hồi quy Probit phù hợp cho các bài toán phân loại nhị phân và có nhiều ứng dụng trong kinh tế lượng, y học, và khoa học xã hội.

Chương 16

Mô hình hồi quy Tobit (Tobit Regression)

16.1 Giới thiệu

Mô hình hồi quy Tobit được phát triển bởi James Tobin (1958) để xử lý dữ liệu bị kiểm duyệt (censored data). Đây là một mô hình thống kê mở rộng từ hồi quy tuyến tính nhưng có tính đến sự kiểm duyệt của biến phụ thuộc.

16.2 Mô hình toán học

Giả sử chúng ta có mô hình hồi quy tuyến tính chuẩn:

$$y_i^* = x_i\beta + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2) \tag{16.1}$$

trong đó:

- y_i^* là biến tiềm ẩn (latent variable), có thể nhận giá trị bất kỳ.
- x_i là vector các biến giải thích (các yếu tố ảnh hưởng).
- β là vector tham số cần ước lượng.
- ε_i là sai số ngẫu nhiên, giả sử phân phối chuẩn với trung bình 0 và phương sai σ^2.

Tuy nhiên, trong thực tế, giá trị quan sát được y_i có thể bị kiểm duyệt như sau:

$$y_i = \begin{cases} y_i^*, & \text{nếu } y_i^* > L, \\ L, & \text{nếu } y_i^* \leq L. \end{cases} \tag{16.2}$$

Trong đó L là ngưỡng kiểm duyệt (ví dụ như mức thu nhập tối thiểu hoặc mức tiêu dùng tối thiểu).

16.3 Ước lượng tham số

Do y_i bị kiểm duyệt, phương pháp hồi quy OLS thông thường không thể áp dụng. Thay vào đó, phương pháp hợp lý tối đa (Maximum Likelihood Estimation - MLE) được sử dụng để ước lượng tham số.

Hàm hợp lý của mô hình Tobit được viết dưới dạng:

$$L(\beta, \sigma^2) = \prod_{y_i > L} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - x_i\beta)^2}{2\sigma^2}\right) \prod_{y_i = L} \Phi\left(\frac{L - x_i\beta}{\sigma}\right), \tag{16.3}$$

trong đó $\Phi(\cdot)$ là hàm phân phối tích lũy (CDF) của phân phối chuẩn.

Ước lượng hợp lý tối đa (MLE) của β và σ^2 được tính bằng cách cực đại hóa hàm log-hợp lý:

$$\log L(\beta, \sigma^2) = \sum_{y_i > L} \left[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y_i - x_i\beta)^2}{2\sigma^2}\right] + \sum_{y_i = L} \log \Phi\left(\frac{L - x_i\beta}{\sigma}\right). \tag{16.4}$$

Các tham số β và σ^2 thường được ước lượng bằng phương pháp Newton-Raphson hoặc các kỹ thuật tối ưu hóa số học khác.

16.4 Ứng dụng của mô hình Tobit

Mô hình Tobit thường được sử dụng trong các lĩnh vực:

- Kinh tế: Phân tích chi tiêu hộ gia đình với giá trị chi tiêu tối thiểu.
- Tài chính: Phân tích mức đầu tư khi có ràng buộc vốn.
- Khoa học xã hội: Đánh giá mức độ hài lòng khi dữ liệu bị giới hạn.

16.5 Kết luận

Mô hình hồi quy Tobit là một công cụ mạnh mẽ để phân tích dữ liệu bị kiểm duyệt, mở rộng khả năng của mô hình hồi quy tuyến tính truyền thống bằng cách xử lý các giá trị bị giới hạn. Phương pháp MLE được sử dụng để ước lượng tham số trong mô hình này.

Chương 17

Mô hình hồi quy Poisson (Poisson Regression)

17.1 Giới thiệu

Mô hình hồi quy Poisson được sử dụng để mô hình hóa dữ liệu đếm, trong đó biến phụ thuộc (biến phản hồi) Y là một biến rời rạc nhận các giá trị nguyên không âm. Mô hình này giả định rằng số lần xuất hiện của sự kiện tuân theo phân phối Poisson.

17.2 Định nghĩa mô hình

Cho biến phản hồi Y_i tuân theo phân phối Poisson:

$$Y_i \sim \text{Poisson}(\lambda_i) \tag{17.1}$$

trong đó $\lambda_i > 0$ là giá trị kỳ vọng có điều kiện của Y_i.

Hàm phân phối xác suất (PMF) của biến Poisson là:

$$P(Y_i = k|X_i) = \frac{e^{-\lambda_i}\lambda_i^k}{k!}, \quad k = 0, 1, 2, \ldots \tag{17.2}$$

17.3 Hàm liên kết và mô hình tuyến tính

Hàm kỳ vọng của Y_i được mô hình hóa thông qua một hàm liên kết log:

$$\log(\lambda_i) = X_i\beta \tag{17.3}$$

trong đó:

- X_i là vector hàng của các biến độc lập (biến giải thích) cho quan sát thứ i.
- β là vector hệ số hồi quy.

Từ đó, ta có:

$$\lambda_i = e^{X_i\beta} \tag{17.4}$$

17.4 Ước lượng tham số

Các tham số β của mô hình được ước lượng bằng phương pháp hợp lý tối đa (MLE). Hàm hợp lý của mô hình là:

$$L(\beta) = \prod_{i=1}^{n} \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} \tag{17.5}$$

Hàm log-hợp lý:

$$\ell(\beta) = \sum_{i=1}^{n} (y_i \log \lambda_i - \lambda_i - \log y_i!) \tag{17.6}$$

Thay $\lambda_i = e^{X_i\beta}$ vào:

$$\ell(\beta) = \sum_{i=1}^{n} \left(y_i X_i \beta - e^{X_i\beta} - \log y_i! \right) \tag{17.7}$$

Hệ số β được tìm bằng cách giải phương trình:

$$\frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^{n} X_i (y_i - e^{X_i\beta}) = 0 \tag{17.8}$$

Phương trình này thường được giải bằng các phương pháp số như Newton-Raphson.

17.5 Kiểm định ý nghĩa của mô hình

17.5.1 Kiểm định Wald

Ta kiểm định giả thuyết:

$$H_0 : \beta_j = 0 \quad \text{v.s.} \quad H_1 : \beta_j \neq 0 \tag{17.9}$$

Dùng thống kê kiểm định Wald:

$$W_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \sim N(0, 1) \tag{17.10}$$

17.5.2 Kiểm định độ phù hợp

Ta có thể sử dụng thống kê Deviance hoặc kiểm định Pearson để kiểm tra độ phù hợp của mô hình.

Phần IV

Hồi quy sống còn (Survival Regression)

Chương 18

Mô hình hồi quy Cox (Cox Proportional Hazards Model)

18.1 Giới thiệu

Mô hình hồi quy Cox (Cox Proportional Hazards Model) là một phương pháp phân tích sống sót (survival analysis) phổ biến được giới thiệu bởi David Cox vào năm 1972. Mô hình này được sử dụng để phân tích tác động của các biến giải thích đối với thời gian sống sót của một đối tượng.

18.2 Cấu trúc mô hình

Mô hình Cox không giả định một dạng cụ thể cho hàm nguy cơ (hazard function) $h(t)$, nhưng giả định rằng tỷ lệ nguy cơ giữa các cá thể là không đổi theo thời gian. Hàm nguy cơ có dạng:

$$h(t|X) = h_0(t)e^{\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p}, \tag{18.1}$$

trong đó:

- $h(t|X)$: Hàm nguy cơ tại thời điểm t khi có biến dự báo X.
- $h_0(t)$: Hàm nguy cơ cơ bản (baseline hazard function), không phụ thuộc vào biến giải thích.
- $X_1, X_2, \ldots, X_p$: Các biến dự báo (covariates).
- $\beta_1, \beta_2, \ldots, \beta_p$: Các hệ số hồi quy cần ước lượng.

18.3 Ước lượng tham số

Các hệ số β được ước lượng thông qua phương pháp hợp lý từng phần (partial likelihood). Hàm hợp lý từng phần được biểu diễn dưới dạng:

$$L(\beta) = \prod_{i=1}^{n} \frac{e^{\beta^T X_i}}{\sum_{j \in R(t_i)} e^{\beta^T X_j}}, \tag{18.2}$$

trong đó:

- t_i là thời gian sự kiện xảy ra cho cá thể thứ i.
- $R(t_i)$ là tập hợp các cá thể còn sống ngay trước thời điểm t_i.

Hàm log-likelihood tương ứng là:

$$\ell(\beta) = \sum_{i=1}^{n} \left(\beta^T X_i - \log \sum_{j \in R(t_i)} e^{\beta^T X_j} \right). \tag{18.3}$$

Ước lượng cực đại hợp lý (Maximum Likelihood Estimation - MLE) được sử dụng để tìm các hệ số β tối ưu bằng cách giải phương trình đạo hàm của hàm log-likelihood:

$$\frac{\partial \ell(\beta)}{\partial \beta} = 0. \tag{18.4}$$

18.4 Giả định của mô hình Cox

Mô hình Cox dựa trên các giả định sau:

- Tỷ lệ nguy cơ (hazard ratio) giữa các cá thể là không đổi theo thời gian.
- Các biến giải thích có tác động tuyến tính lên log tỷ lệ nguy cơ.
- Không có sự vi phạm nghiêm trọng về phương sai và tự tương quan.

18.5 Kiểm định giả thuyết và đánh giá mô hình

18.5.1 Kiểm định ý nghĩa của các hệ số

Giá trị β được kiểm định bằng kiểm định Wald hoặc kiểm định tỷ số hợp lý (Likelihood Ratio Test - LRT) với giả thuyết:

$$H_0 : \beta_i = 0 \quad \text{(biến không ảnh hưởng đến thời gian sống sót)}. \tag{18.5}$$

18.5.2 Kiểm tra giả định tỷ lệ nguy cơ

Giả định tỷ lệ nguy cơ có thể được kiểm tra bằng:

- Phương pháp đồ thị Schoenfeld residuals.
- Kiểm định thống kê dựa trên phần dư Schoenfeld.

18.6 Ứng dụng thực tế

Mô hình Cox được sử dụng rộng rãi trong các lĩnh vực:

- Y học: Ước lượng thời gian sống sót của bệnh nhân với các yếu tố nguy cơ.
- Tài chính: Dự báo thời gian vỡ nợ của doanh nghiệp.
- Kinh tế: Phân tích thời gian tồn tại của doanh nghiệp trên thị trường.

18.7 Kết luận

Mô hình hồi quy Cox là một công cụ mạnh mẽ trong phân tích sống sót, cho phép nghiên cứu tác động của nhiều yếu tố lên thời gian xảy ra sự kiện mà không cần giả định dạng cụ thể của hàm nguy cơ cơ bản.

Chương 19

Mô hình Weibull (Weibull Regression Model)

19.1 Giới thiệu

Mô hình hồi quy Weibull là một phương pháp thống kê dùng để mô hình hóa thời gian đến một sự kiện (thời gian sống hoặc thời gian hỏng hóc) với phân phối Weibull. Đây là một trong những mô hình phổ biến trong phân tích sống sót và độ tin cậy.

19.2 Phân phối Weibull

Phân phối Weibull có hàm mật độ xác suất (PDF) được định nghĩa như sau:

$$f(t;\lambda,k) = \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1} \exp\left(-\left(\frac{t}{\lambda}\right)^{k}\right), \quad t > 0, \tag{19.1}$$

trong đó:

- t là biến ngẫu nhiên thời gian sống,
- $\lambda > 0$ là tham số tỷ lệ (scale parameter),
- $k > 0$ là tham số hình dạng (shape parameter).

Hàm phân phối tích lũy (CDF) của phân phối Weibull là:

$$F(t) = 1 - \exp\left(-\left(\frac{t}{\lambda}\right)^{k}\right). \tag{19.2}$$

19.3 Mô hình hồi quy Weibull

Mô hình hồi quy Weibull mở rộng phân phối Weibull bằng cách đưa vào một tập hợp các biến dự báo $\mathbf{x} = (x_1, x_2, ..., x_p)$. Mô hình được định nghĩa bởi:

$$\lambda_i = \exp(-\mathbf{x}_i^T \beta), \tag{19.3}$$

trong đó:

- λ_i là tham số tỷ lệ cho quan sát thứ i,
- $\mathbf{x}_i$ là vector các biến dự báo,
- β là vector hệ số hồi quy.

Hàm mật độ xác suất có dạng:

$$f(t_i|\mathbf{x}_i) = k \exp(-\mathbf{x}_i^T \beta) t_i^{k-1} \exp\left(-t_i^k \exp(-k\mathbf{x}_i^T \beta)\right). \tag{19.4}$$

Hàm log-likelihood của mô hình Weibull có dạng:

$$\mathcal{L}(\beta, k) = \sum_{i=1}^{n} \left[d_i \left(\log k + (k-1) \log t_i - \mathbf{x}_i^T \beta \right) - t_i^k \exp(-k\mathbf{x}_i^T \beta) \right], \tag{19.5}$$

trong đó d_i là biến chỉ báo (1 nếu sự kiện xảy ra, 0 nếu bị kiểm duyệt).

19.4 Ước lượng tham số

Các tham số β và k có thể được ước lượng bằng cách tối đa hóa hàm log-likelihood bằng phương pháp Newton-Raphson hoặc các thuật toán tối ưu số khác.

19.5 Ứng dụng thực tế

Mô hình hồi quy Weibull thường được sử dụng trong các lĩnh vực như:

- Phân tích độ tin cậy trong kỹ thuật,
- Phân tích sống sót trong y học,
- Mô hình hóa thời gian đến sự kiện trong kinh tế và tài chính.

19.6 Kết luận

Mô hình hồi quy Weibull là một công cụ mạnh mẽ để phân tích thời gian sống sót với phân phối Weibull. Với tính linh hoạt trong điều chỉnh hình dạng phân phối, mô hình này giúp phản ánh tốt hơn bản chất của dữ liệu so với các mô hình hồi quy tuyến tính thông thường.

Chương 20

Mô hình hồi quy Log-logistic (Log-logistic Regression Model)

20.1 Giới thiệu

Mô hình hồi quy Log-logistic là một dạng mô hình phân tích sống sót hoặc phân tích thời gian, tương tự như mô hình Weibull nhưng có khả năng xử lý tốt hơn đối với dữ liệu có đuôi phân phối dày. Mô hình này được sử dụng rộng rãi trong các lĩnh vực như kinh tế, y học, và kỹ thuật.

20.2 Định nghĩa mô hình

Giả sử biến thời gian sống hoặc thời gian đến sự kiện T tuân theo phân phối Log-logistic với hàm mật độ xác suất (PDF) được cho bởi:

$$f(T) = \frac{(\lambda\gamma)(T/\lambda)^{\gamma-1}}{(1+(T/\lambda)^{\gamma})^2}, \quad T > 0, \tag{20.1}$$

trong đó:

- $\lambda > 0$ là tham số tỷ lệ (scale parameter),
- $\gamma > 0$ là tham số hình dạng (shape parameter).

Hàm phân phối tích lũy (CDF) của mô hình Log-logistic được viết dưới dạng:

$$F(T) = \frac{1}{1+(T/\lambda)^{-\gamma}}, \quad T > 0. \tag{20.2}$$

Hàm nguy cơ (hazard function) được xác định bởi:

$$h(T) = \frac{(\lambda\gamma)(T/\lambda)^{\gamma-1}}{(1+(T/\lambda)^{\gamma})T}. \tag{20.3}$$

20.3 Hồi quy Log-logistic

Để mô hình hóa tác động của các biến dự báo $X = (X_1, X_2, \ldots, X_p)$ lên biến phụ thuộc T, ta sử dụng dạng hồi quy Log-logistic:

$$\log(T) = X\beta + \varepsilon, \tag{20.4}$$

trong đó:

- $X\beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$ là tổ hợp tuyến tính của các biến dự báo,
- ε là nhiễu có phân phối Logistic chuẩn với trung bình 0 và phương sai $\pi^2/3$.

Từ đây, ta có hàm phân phối tích lũy có điều kiện:

$$P(T \leq t|X) = \frac{1}{1 + \exp(-(\log t - X\beta)/\sigma)}, \tag{20.5}$$

trong đó σ là tham số tỷ lệ.

20.4 Ước lượng tham số

Tham số của mô hình có thể được ước lượng bằng phương pháp hợp lý tối đa (Maximum Likelihood Estimation - MLE). Hàm hợp lý có dạng:

$$L(\beta, \sigma) = \prod_{i=1}^{n} f(T_i|X_i; \beta, \sigma), \tag{20.6}$$

trong đó T_i là thời gian quan sát của cá nhân thứ i. Tương đương, log-hàm hợp lý là:

$$\ell(\beta, \sigma) = \sum_{i=1}^{n} \left[\log f(T_i|X_i; \beta, \sigma)\right]. \tag{20.7}$$

Giải phương trình đạo hàm của log-hàm hợp lý sẽ cho giá trị ước lượng $\hat{\beta}$ và $\hat{\sigma}$.

20.5 Kiểm định mô hình

20.5.1 Kiểm định Wald

Kiểm định Wald được sử dụng để kiểm tra ý nghĩa thống kê của từng hệ số hồi quy:

$$W_j = \frac{\hat{\beta}_j}{\mathrm{SE}(\hat{\beta}_j)}, \tag{20.8}$$

trong đó $\mathrm{SE}(\hat{\beta}_j)$ là sai số chuẩn của ước lượng $\hat{\beta}_j$. Nếu $|W_j|$ lớn hơn một giá trị tới hạn, ta bác bỏ giả thuyết $H_0 : \beta_j = 0$.

20.5.2 Kiểm định tỉ số hợp lý (Likelihood Ratio Test)

Kiểm định này so sánh hai mô hình lồng nhau:

$$LR = -2(\ell_{\text{restricted}} - \ell_{\text{full}}), \tag{20.9}$$

trong đó $\ell_{\text{restricted}}$ là log-hàm hợp lý của mô hình ràng buộc (không có biến giải thích), còn ℓ_{full} là log-hàm hợp lý của mô hình đầy đủ. Giá trị LR được so sánh với phân phối χ^2.

20.5.3 Kiểm định hệ số hồi quy chung (Wald Test tổng quát)

Ta kiểm định giả thuyết tổng quát:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0. \tag{20.10}$$

Giá trị kiểm định Wald tổng quát có dạng:

$$W = \beta' \left(\text{Var}(\hat{\beta})\right)^{-1} \beta \sim \chi^2_p. \tag{20.11}$$

20.6 Kết luận

Mô hình hồi quy Log-logistic là một công cụ mạnh mẽ trong phân tích dữ liệu sống sót và thời gian đến sự kiện. Với khả năng xử lý các đuôi phân phối dài, nó thích hợp cho nhiều ứng dụng thực tế như kinh tế, y học, và kỹ thuật.

Chương 21

Mô hình hồi quy Gamma (Gamma Regression Model)

21.1 Giới thiệu

Mô hình hồi quy Gamma là một dạng mô hình hồi quy được sử dụng khi biến phụ thuộc (Y) là một biến dương và có độ lệch về phía phải. Mô hình này phù hợp với dữ liệu có phân phối Gamma.

21.2 Định nghĩa mô hình

Giả sử ta có một tập dữ liệu gồm n quan sát, với biến phụ thuộc Y_i và vector các biến độc lập $\mathbf{X}_i = (X_{i1}, X_{i2}, ..., X_{ip})$. Mô hình hồi quy Gamma được định nghĩa như sau:

$$Y_i \sim \text{Gamma}(\alpha, \theta_i) \tag{21.1}$$

trong đó:

- Y_i là biến phản hồi dương.
- α là tham số hình dạng (shape parameter) của phân phối Gamma.
- θ_i là tham số tỷ lệ (scale parameter), phụ thuộc vào các biến dự báo X_i thông qua hàm liên kết.

21.3 Hàm liên kết

Trong mô hình hồi quy Gamma, kỳ vọng có điều kiện của Y_i được biểu diễn dưới dạng:

$$\mathbb{E}[Y_i|\mathbf{X}_i] = \mu_i = g^{-1}(\eta_i) \tag{21.2}$$

trong đó:

- $\eta_i = \mathbf{X}_i^T \boldsymbol{\beta}$ là hàm tuyến tính của các biến độc lập.

- $g(\mu_i)$ là một hàm liên kết để đảm bảo tính dương của μ_i.

Một số hàm liên kết phổ biến được sử dụng trong hồi quy Gamma:

- Hàm log: $g(\mu) = \log(\mu)$, tức là $\mu_i = \exp(\eta_i)$.
- Hàm nghịch đảo: $g(\mu) = 1/\mu$, tức là $\mu_i = 1/\eta_i$.
- Hàm căn bậc hai: $g(\mu) = \sqrt{\mu}$.

21.4 Ước lượng tham số

Các tham số $\boldsymbol{\beta}$ trong mô hình được ước lượng bằng phương pháp hợp lý tối đa (Maximum Likelihood Estimation - MLE). Hàm mật độ xác suất của một quan sát Y_i theo phân phối Gamma có dạng:

$$f(Y_i|\alpha, \theta_i) = \frac{1}{\Gamma(\alpha)\theta_i^\alpha} Y_i^{\alpha-1} e^{-Y_i/\theta_i} \tag{21.3}$$

Hàm log-likelihood của mẫu có dạng:

$$\ell(\boldsymbol{\beta}, \alpha) = \sum_{i=1}^{n} \left[\alpha \log Y_i - \log \Gamma(\alpha) - \alpha \log \theta_i - \frac{Y_i}{\theta_i} \right] \tag{21.4}$$

Để tìm ước lượng hợp lý tối đa (MLE), ta giải hệ phương trình:

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = 0, \quad \frac{\partial \ell}{\partial \alpha} = 0. \tag{21.5}$$

Do phương trình không có nghiệm tường minh, ta sử dụng các thuật toán tối ưu hóa như phương pháp Newton-Raphson hoặc Fisher Scoring để tìm nghiệm.

21.5 Kiểm định mô hình

Sau khi ước lượng mô hình, ta cần kiểm tra tính phù hợp của mô hình:

- Kiểm tra phương sai: Phương sai của Y có dạng:

$$\text{Var}(Y_i) = \alpha \theta_i^2. \tag{21.6}$$

  Nếu phương sai không tỷ lệ với bình phương trung bình, mô hình có thể không phù hợp.

- Kiểm tra phân phối dư lượng: Vẽ biểu đồ Q-Q plot của dư lượng Pearson để kiểm tra giả định phân phối Gamma.
- Kiểm tra giả thuyết: Dùng kiểm định Wald hoặc kiểm định tỉ số hợp lý để đánh giá ý nghĩa của các tham số.

21.6 Kết luận

Mô hình hồi quy Gamma là một công cụ hữu ích khi phân tích dữ liệu với biến phản hồi dương. Việc lựa chọn hàm liên kết phù hợp và kiểm định mô hình là rất quan trọng để đảm bảo tính chính xác của kết quả phân tích.

Chương 22

Mô hình hồi quy hỗn hợp (Frailty Models)

22.1 Giới thiệu

Mô hình hồi quy hỗn hợp (Frailty Models) là một phần mở rộng của mô hình hồi quy sinh tồn truyền thống, chẳng hạn như mô hình Cox, nhằm xử lý sự phụ thuộc giữa các quan sát do có các yếu tố không quan sát được (frailty).

22.2 Mô hình toán học

Giả sử thời gian sống T_i của cá nhân i tuân theo phân phối nguy cơ $h_i(t)$, mô hình Cox chuẩn được viết như:

$$h_i(t|X_i) = h_0(t)e^{X_i\beta}, \tag{22.1}$$

trong đó:

- $h_0(t)$ là hàm nguy cơ cơ sở,
- X_i là vector các biến tiên lượng,
- β là vector hệ số hồi quy.

Tuy nhiên, mô hình Cox giả định rằng tất cả các cá nhân có cùng một hàm nguy cơ khi đã tính đến các biến tiên lượng. Trong thực tế, có thể có những yếu tố không quan sát được ảnh hưởng đến nguy cơ, được mô hình hóa bằng một biến ngẫu nhiên Z_i:

$$h_i(t|X_i, Z_i) = Z_i h_0(t)e^{X_i\beta}, \tag{22.2}$$

trong đó Z_i là biến frailty, thường được giả định tuân theo phân phối Gamma với kỳ vọng bằng 1 và phương sai θ.

22.3 Ước lượng tham số

Ước lượng tham số trong mô hình frailty thường dựa trên phương pháp hợp lý tối đa. Hàm hợp lý có dạng:

$$L(\beta, \theta) = \int \prod_{i=1}^{n} h_i(t_i|X_i, Z_i) S_i(t_i|X_i, Z_i) f(Z_i; \theta) dZ_i, \tag{22.3}$$

trong đó:

- $S_i(t)$ là hàm sống sót của cá nhân i,
- $f(Z_i; \theta)$ là hàm mật độ xác suất của biến frailty.

Việc tích phân theo Z_i thường không có công thức đóng, do đó các phương pháp xấp xỉ như Laplace hoặc Monte Carlo được sử dụng để ước lượng tham số.

22.4 Kết luận

Mô hình hồi quy hỗn hợp cung cấp một cách tiếp cận linh hoạt để mô hình hóa dữ liệu sinh tồn có sự phụ thuộc do các yếu tố không quan sát được. Việc lựa chọn phân phối frailty phù hợp và sử dụng các phương pháp tính toán hiệu quả là chìa khóa để có được ước lượng chính xác.

Phần V

Ước lượng Bayesian

Chương 23

Tổng quan về Ước lượng Bayesian

23.1 Giới thiệu về Ước lượng Bayesian

Ước lượng Bayesian dựa trên định lý Bayes để cập nhật niềm tin về tham số của một mô hình thống kê sau khi có dữ liệu quan sát. Trái ngược với phương pháp ước lượng tần suất (frequentist), trong đó tham số là cố định nhưng không xác định, phương pháp Bayesian coi tham số là một biến ngẫu nhiên với một phân phối xác suất tiên nghiệm (prior distribution).

23.1.1 Định lý Bayes

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} \tag{23.1}$$

Trong đó:

- θ là tham số cần ước lượng.
- D là tập dữ liệu quan sát được.
- $P(\theta)$ là phân phối tiên nghiệm (**prior distribution**).
- $P(D|\theta)$ là hàm khả năng (**likelihood function**).
- $P(D)$ là xác suất bằng tổng (evidence):

$$P(D) = \int P(D|\theta)P(\theta)d\theta \tag{23.2}$$

- $P(\theta|D)$ là phân phối hậu nghiệm (**posterior distribution**).

23.2 Phân phối trước, phân phối hậu nghiệm và quy trình tính toán

23.2.1 Phân phối tiên nghiệm (Prior Distribution)

Phân phối tiên nghiệm $P(\theta)$ phản ánh kiến thức hoặc niềm tin có sẵn về θ trước khi quan sát dữ liệu. Các loại phổ biến:

- **Phân phối không thông tin**: $P(\theta) \propto 1$.
- **Phân phối chuẩn**: $\theta \sim \mathcal{N}(\mu_0, \sigma_0^2)$.
- **Phân phối Beta**: $\theta \sim \text{Beta}(\alpha, \beta)$.

23.2.2 Phân phối hậu nghiệm (Posterior Distribution)

$$P(\theta|D) \propto P(D|\theta)P(\theta) \tag{23.3}$$

Ví dụ: Với mẫu $x_1, x_2, ..., x_n$ từ phân phối Gaussian:

$$P(D|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \theta)^2}{2\sigma^2}\right) \tag{23.4}$$

Nếu $P(\theta) = \mathcal{N}(\mu_0, \sigma_0^2)$, thì hậu nghiệm cũng là Gaussian:

$$P(\theta|D) = \mathcal{N}(\mu_n, \sigma_n^2) \tag{23.5}$$

Với:

$$\mu_n = \frac{\sigma_0^2 \sum x_i + \sigma^2 \mu_0}{n\sigma_0^2 + \sigma^2} \tag{23.6}$$

$$\sigma_n^2 = \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)^{-1} \tag{23.7}$$

Hoặc :

$$\mu_n = \frac{\frac{1}{\sigma_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{x}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}} \tag{23.8}$$

$$\sigma_n^2 = \frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}} \tag{23.9}$$

Quy trình tính toán trong Bayesian

1. Chọn phân phối tiên nghiệm phù hợp.
2. Định nghĩa hàm khả năng dựa trên mô hình dữ liệu.
3. Áp dụng Định lý Bayes để cập nhật phân phối hậu nghiệm.
4. Nếu cần ước lượng điểm, sử dụng kỳ vọng hậu nghiệm:

$$\hat{\theta} = \mathbb{E}[\theta|D] \tag{23.10}$$

23.3 Diễn giải ý tưởng của phương pháp ước lượng Bayesian trong kinh tế lượng

23.3.1 Định nghĩa

- **Phân phối tiên nghiệm (Prior)**: Là *niềm tin ban đầu* về một tham số trước khi quan sát dữ liệu thực tế.
- **Phân phối hậu nghiệm (Posterior)**: Là *niềm tin đã được cập nhật* sau khi quan sát dữ liệu thực tế.

Công thức Bayes thể hiện điều này:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} \tag{23.11}$$

Trong đó:

- $P(\theta)$: Tiên nghiệm (prior), phản ánh kiến thức có trước về tham số θ.
- $P(D|\theta)$: Khả năng (likelihood), thể hiện xác suất dữ liệu quan sát được nếu θ có giá trị nào đó.
- $P(D)$: Bằng chứng (evidence), đại diện cho xác suất của dữ liệu quan sát được.
- $P(\theta|D)$: Hậu nghiệm (posterior), tức là phân phối xác suất của θ sau khi cập nhật thông tin từ dữ liệu.

23.3.2 Ví dụ thực tế: Dự báo lạm phát

*** Bối cảnh**

Giả sử cần dự báo lạm phát của Việt Nam vào năm 2025 dựa trên dữ liệu lịch sử.

*** Bước 1: Xác định tiên nghiệm (Prior)**

Trước khi có dữ liệu mới, có thể giả định:

- Lạm phát thường dao động trong khoảng 2% - 5%.
- Trung bình 10 năm qua: 3%, độ lệch chuẩn 1%.
- Chọn phân phối tiên nghiệm:

$$\pi(\theta) \sim \mathcal{N}(3, 1^2) \tag{23.12}$$

*** Bước 2: Quan sát dữ liệu thực tế (Likelihood)**

Dữ liệu từ 2015-2024 cho thấy:

- Lạm phát năm 2023: 4.2%.
- Dữ liệu vĩ mô cho thấy xu hướng tăng.
- Giả sử độ lệch chuẩn: 0.5%.

*** Bước 3: Cập nhật hậu nghiệm (Posterior)**

Sử dụng công thức Bayes:

$$P(\theta|D) \sim \mathcal{N}(3.5, 0.7^2) \tag{23.13}$$

Ý nghĩa:

- Dự báo lạm phát dịch chuyển từ 3% lên 3.5%.
- Độ không chắc chắn giảm nhờ có dữ liệu mới.

*** Bước 4: Ứng dụng thực tế**

- **Chính phủ**: Điều chỉnh chính sách tiền tệ để kiểm soát lạm phát.
- **Ngân hàng**: Điều chỉnh lãi suất cho vay.
- **Doanh nghiệp**: Lập kế hoạch giá cả.

=> Tóm lại:

- Tiên nghiệm giúp tận dụng kiến thức có sẵn khi dữ liệu ít.
- Hậu nghiệm kết hợp thông tin từ dữ liệu mới để đưa ra dự báo tốt hơn.
- Bayesian giúp cải thiện dự báo kinh tế lượng ngay cả khi dữ liệu ít.

Mô tả chi tiết quá trình tính toán:

**** Thu thập dữ liệu lạm phát thực tế (2015-2024)**

Dưới đây là tỷ lệ lạm phát hàng năm của Việt Nam từ năm 2015 đến 2024, dựa trên dữ liệu từ Ngân hàng Thế giới và các nguồn khác:

Năm	Tỷ lệ lạm phát (%)
2015	0.6
2016	2.7
2017	3.5
2018	3.5
2019	2.8
2020	3.2
2021	1.8
2022	3.2
2023	3.3
2024	4.1

Bảng 23.1: Tỷ lệ lạm phát của Việt Nam (2015-2024)

** Xác định phân phối tiên nghiệm (Prior)

Giả sử rằng trước khi quan sát dữ liệu, chúng ta có niềm tin rằng tỷ lệ lạm phát trung bình của Việt Nam dao động quanh mức 3% với độ lệch chuẩn là 1%. Do đó, ta chọn phân phối tiên nghiệm cho tỷ lệ lạm phát θ là:

$$\theta \sim N(3, 1^2) \tag{23.14}$$

** Tính toán thống kê từ dữ liệu

Từ dữ liệu lạm phát thực tế, ta tính:

Trung bình mẫu ($\bar{x}$):

$$\bar{x} = \frac{0.6 + 2.7 + 3.5 + 3.5 + 2.8 + 3.2 + 1.8 + 3.2 + 3.3 + 4.1}{10} = 3.07 \tag{23.15}$$

Độ lệch chuẩn mẫu (s):

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \approx 0.87 \tag{23.16}$$

** Xác định phân phối hậu nghiệm (Posterior)

Sử dụng công thức cập nhật Bayesian, phân phối hậu nghiệm của θ là:

$$\theta|D \sim N(\mu_n, \sigma_n^2) \tag{23.17}$$

Trong đó:

$$\mu_n = \frac{\frac{1}{\sigma_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{x}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}} \tag{23.18}$$

$$\sigma_n^2 = \frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}} \tag{23.19}$$

Với các giá trị:

- $\mu_0 = 3$, $\sigma_0^2 = 1$
- $n = 10$, $\bar{x} = 3.07$
- Giả sử $\sigma^2 = s^2 = 0.87^2$

Thay vào công thức:

$$\mu_n = \frac{\frac{1}{1} \times 3 + \frac{10}{0.87^2} \times 3.07}{\frac{1}{1} + \frac{10}{0.87^2}} \approx 3.06 \tag{23.20}$$

$$\sigma_n^2 = \frac{1}{\frac{1}{1} + \frac{10}{0.87^2}} \approx 0.07 \tag{23.21}$$

Vậy, phân phối hậu nghiệm của tỷ lệ lạm phát năm 2025 là:

$$\theta|D \sim N(3.06, 0.07) \tag{23.22}$$

**** Dự báo và Kết luận**

Dựa trên phân phối hậu nghiệm, dự báo tỷ lệ lạm phát cho năm 2025 là khoảng 3.06% với độ không chắc chắn được đo bằng độ lệch chuẩn khoảng $\sqrt{0.07} \approx 0.26$. Điều này cho thấy dự báo khá tin cậy.

Lưu ý: Phương pháp trên dựa trên giả định rằng dữ liệu lạm phát hàng năm tuân theo phân phối chuẩn. Trong thực tế, cần kiểm tra các giả định này và có thể sử dụng các phương pháp phức tạp hơn để dự báo chính xác hơn.

23.4 Kết luận

Ước lượng Bayesian cung cấp một khung lý thuyết mạnh mẽ cho phân tích dữ liệu. Với MCMC, ta có thể giải quyết các mô hình phức tạp mà phương pháp tần suất gặp khó khăn.

Chương 24

Hồi quy Bayesian (Bayesian Regression)

24.1 Giới thiệu về Hồi quy Bayesian

Hồi quy Bayesian là một phương pháp thống kê dựa trên lý thuyết xác suất Bayes để ước lượng tham số của mô hình hồi quy. Thay vì coi tham số là cố định như trong phương pháp hồi quy thường quy (frequentist), phương pháp Bayesian coi tham số là biến ngẫu nhiên có phân phối xác suất.

24.2 Công thức Bayes

Công thức Bayes cơ bản được biểu diễn như sau:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} \tag{24.1}$$

Trong đó:

- θ là tham số của mô hình cần ước lượng,
- D là dữ liệu quan sát,
- $P(\theta|D)$ là phân phối hậu nghiệm (posterior) của θ sau khi quan sát dữ liệu,
- $P(D|\theta)$ là hàm hợp lý (likelihood),
- $P(\theta)$ là phân phối tiên nghiệm (prior),
- $P(D)$ là hằng số chuẩn hóa (marginal likelihood), cũng được gọi là bằng chứng (evidence).

24.3 Mô hình Hồi quy tuyến tính Bayesian

Giả sử ta có một tập dữ liệu gồm n quan sát (X, y) với $X \in \mathbb{R}^{n \times p}$ là ma trận đặc trưng và $y \in \mathbb{R}^n$ là vector đầu ra. Mô hình hồi quy tuyến tính có dạng:

$$y = X\beta + \epsilon, \tag{24.2}$$

trong đó:

- $\beta \in \mathbb{R}^p$ là vector tham số cần ước lượng,
- $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ là nhiễu Gaussian.

24.3.1 Xác định phân phối tiên nghiệm

Giả sử rằng β có phân phối tiên nghiệm Gaussian:

$$\beta \sim \mathcal{N}(\mu_0, \Sigma_0), \tag{24.3}$$

trong đó μ_0 và Σ_0 là trung bình và hiệp phương sai tiên nghiệm.

24.3.2 Xác định hàm hợp lý

Vì nhiễu ϵ tuân theo phân phối Gaussian, nên xác suất có điều kiện của y theo β là:

$$P(y|X, \beta) = \mathcal{N}(X\beta, \sigma^2 I_n). \tag{24.4}$$

24.3.3 Tính toán phân phối hậu nghiệm

Theo định lý Bayes, phân phối hậu nghiệm của β được xác định bởi:

$$P(\beta|X, y) \propto P(y|X, \beta)P(\beta). \tag{24.5}$$

Do cả $P(y|X, \beta)$ và $P(\beta)$ đều là phân phối Gaussian, nên phân phối hậu nghiệm của β cũng là Gaussian:

$$\beta|X, y \sim \mathcal{N}(\mu_n, \Sigma_n), \tag{24.6}$$

trong đó:

$$\Sigma_n = (X^T X/\sigma^2 + \Sigma_0^{-1})^{-1}, \tag{24.7}$$

$$\mu_n = \Sigma_n(X^T y/\sigma^2 + \Sigma_0^{-1}\mu_0). \tag{24.8}$$

24.4 Dự báo Bayesian

Dự báo cho một điểm dữ liệu mới x_* được tính bằng cách lấy kỳ vọng theo phân phối hậu nghiệm:

$$y_*|x_*, X, y \sim \mathcal{N}(x_*^T \mu_n, x_*^T \Sigma_n x_* + \sigma^2). \tag{24.9}$$

24.5 Ưu điểm của Hồi quy Bayesian

- Hỗ trợ mô hình hóa sự không chắc chắn bằng phân phối hậu nghiệm.
- Giúp tránh overfitting khi dữ liệu huấn luyện ít bằng cách sử dụng thông tin tiên nghiệm.

- Linh hoạt hơn so với hồi quy tuyến tính truyền thống.

Hồi quy Bayesian cung cấp một cách tiếp cận mạnh mẽ và linh hoạt cho các bài toán hồi quy, đặc biệt hữu ích khi dữ liệu ít hoặc có nhiễu. Việc sử dụng phân phối tiên nghiệm giúp bổ sung thông tin vào mô hình và cải thiện kết quả dự báo.

24.6 Các mô hình hồi quy Bayesian

24.6.1 Hồi quy tuyến tính Bayesian (Bayesian Linear Regression)

a. Giới thiệu

Hồi quy tuyến tính Bayesian (Bayesian Linear Regression) là một phương pháp thống kê cho phép kết hợp thông tin tiên nghiệm với dữ liệu quan sát để ước lượng tham số của mô hình hồi quy tuyến tính.

b. Mô hình hồi quy tuyến tính

Giả sử ta có mô hình hồi quy tuyến tính tiêu chuẩn:

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I) \tag{24.10}$$

trong đó:

- $y \in \mathbb{R}^n$ là vector giá trị đầu ra,
- $X \in \mathbb{R}^{n \times d}$ là ma trận dữ liệu,
- $\beta \in \mathbb{R}^d$ là vector tham số cần ước lượng,
- $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ là nhiễu Gaussian có phương sai σ^2.

c. Tiên nghiệm (Prior)

Trong Bayesian Regression, ta đặt giả thuyết tiên nghiệm cho β dưới dạng phân phối Gaussian:

$$p(\beta) = \mathcal{N}(\beta | \mu_0, \Sigma_0) \tag{24.11}$$

Trong đó:

- μ_0 là trung bình tiên nghiệm,
- Σ_0 là ma trận hiệp phương sai tiên nghiệm.

d. Phân phối hậu nghiệm (Posterior Distribution)

Sử dụng định lý Bayes:

$$p(\beta|X,y) \propto p(y|X,\beta)p(\beta) \tag{24.12}$$

Trong đó:

- $p(y|X,\beta)$ là phân phối khả năng (likelihood),
- $p(\beta)$ là phân phối tiên nghiệm.

Từ đó, ta có phân phối hậu nghiệm của β là một phân phối Gaussian:

$$p(\beta|X,y) = \mathcal{N}(\beta|\mu_N, \Sigma_N) \tag{24.13}$$

Trong đó:

$$\Sigma_N = (X^T X/\sigma^2 + \Sigma_0^{-1})^{-1} \tag{24.14}$$

$$\mu_N = \Sigma_N(X^T y/\sigma^2 + \Sigma_0^{-1}\mu_0) \tag{24.15}$$

e. Dự báo (Predictive Distribution)

Cho một điểm dữ liệu mới x_*, đầu ra dự đoán được tính bằng:

$$p(y_*|x_*, X, y) = \mathcal{N}(y_*|x_*^T\mu_N, x_*^T\Sigma_N x_* + \sigma^2) \tag{24.16}$$

Hồi quy tuyến tính Bayesian giúp biểu diễn độ không chắc chắn trong ước lượng tham số và dự đoán. Phương pháp này đặc biệt hữu ích khi dữ liệu ít hoặc có nhiễu.

24.6.2 Hồi quy Logistic Bayesian (Bayesian Logistic Regression)

a. Giới thiệu

Hồi quy logistic Bayesian mở rộng mô hình hồi quy logistic bằng cách đưa vào phân bố xác suất trên các tham số, giúp giảm hiện tượng quá khớp (overfitting) và cung cấp phân bố hậu nghiệm cho việc suy luận.

b. Mô hình toán học

Giả sử ta có tập dữ liệu gồm n quan sát $\{(x_i, y_i)\}_{i=1}^n$, trong đó:

- $x_i \in \mathbb{R}^d$ là vector đặc trưng của quan sát thứ i,
- $y_i \in \{0, 1\}$ là nhãn của quan sát thứ i,
- $\beta \in \mathbb{R}^d$ là vector tham số cần ước lượng.

Xác suất của nhãn y_i được mô hình hóa bởi:

$$P(y_i = 1|x_i, \beta) = \sigma(x_i^T \beta), \tag{24.17}$$

trong đó $\sigma(z)$ là hàm sigmoid:

$$\sigma(z) = \frac{1}{1 + e^{-z}}. \tag{24.18}$$

c. Hàm khả năng (Likelihood)

Với toàn bộ tập dữ liệu, hàm khả năng được viết dưới dạng:

$$P(Y|X, \beta) = \prod_{i=1}^{n} \sigma(x_i^T \beta)^{y_i} (1 - \sigma(x_i^T \beta))^{1-y_i}. \tag{24.19}$$

d. Phân bố tiên nghiệm (Prior Distribution)

Thông thường, ta chọn phân bố Gaussian làm tiên nghiệm cho β:

$$P(\beta) = \mathcal{N}(\beta|0, \lambda I), \tag{24.20}$$

trong đó λI là ma trận hiệp phương sai, kiểm soát mức độ phẳng của phân bố tiên nghiệm.

e. Phân bố hậu nghiệm (Posterior Distribution)

Sử dụng Định lý Bayes, phân bố hậu nghiệm của β là:

$$P(\beta|X, Y) \propto P(Y|X, \beta)P(\beta). \tag{24.21}$$

Do không có dạng giải tích đơn giản, ta phải dùng phương pháp xấp xỉ như:

- Lấy mẫu Monte Carlo Markov Chain (MCMC),
- Xấp xỉ Laplace,
- Biến phân (Variational Inference).

Hồi quy logistic Bayesian giúp tích hợp thông tin tiên nghiệm vào quá trình học, cải thiện khả năng suy luận và tránh quá khớp. Việc ước lượng phân bố hậu nghiệm yêu cầu các phương pháp tính toán xấp xỉ.

24.6.3 Hồi quy Bayesian với dữ liệu bảng (Bayesian Panel Data Models)

a. Giới thiệu

Dữ liệu bảng (panel data) kết hợp dữ liệu theo không gian và thời gian, thường được biểu diễn dưới dạng y_{it} với chỉ số i cho từng cá nhân và t cho từng thời điểm. Mô hình hồi quy Bayesian cung cấp một cách tiếp cận xác suất để ước lượng tham số và giải quyết vấn đề bất định.

b. Mô hình hồi quy dữ liệu bảng

Giả sử mô hình hồi quy tuyến tính dạng:

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + \varepsilon_{it}, \quad \varepsilon_{it} \sim \mathcal{N}(0, \sigma^2) \tag{24.22}$$

trong đó:

- y_{it} là biến phụ thuộc của cá nhân i tại thời điểm t.
- $\mathbf{x}_{it}$ là vector các biến độc lập.
- $\boldsymbol{\beta}$ là vector hệ số hồi quy chung.
- α_i là hiệu ứng cá nhân (random effects hoặc fixed effects).
- ε_{it} là nhiễu ngẫu nhiên.

c. Phương pháp Bayesian

Trong Bayesian, các tham số được coi là biến ngẫu nhiên với phân phối tiên nghiệm $p(\boldsymbol{\theta})$. Phân phối hậu nghiệm được tính bằng định lý Bayes:

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{24.23}$$

trong đó $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \alpha_i, \sigma^2\}$ là tập hợp các tham số cần ước lượng.

d. Phân phối tiên nghiệm

- Với $\boldsymbol{\beta}$: sử dụng tiên nghiệm không thông tin $\mathcal{N}(0, \lambda I)$.
- Với α_i: sử dụng $\mathcal{N}(0, \tau^2)$.
- Với σ^2: sử dụng phân phối nghịch gamma Inv-Gamma(a, b).

e. Phân phối hậu nghiệm và suy luận

Phân phối hậu nghiệm không có dạng đóng nên cần sử dụng phương pháp lấy mẫu Monte Carlo Markov Chain (MCMC), chẳng hạn như thuật toán Gibbs Sampling hoặc Hamiltonian Monte Carlo.

f. Mô hình Hiệu ứng Ngẫu nhiên (Random Effects)

Mô hình hiệu ứng ngẫu nhiên giả sử $\alpha_i \sim \mathcal{N}(0, \tau^2)$. Khi đó, phân phối hậu nghiệm của α_i có dạng:

$$p(\alpha_i|\mathbf{y}, \boldsymbol{\beta}, \sigma^2) \sim \mathcal{N}\left(\frac{\tau^2 \sum_t (y_{it} - \mathbf{x}_{it}'\boldsymbol{\beta})}{\sigma^2 + T\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + T\tau^2}\right) \tag{24.24}$$

Hồi quy Bayesian với dữ liệu bảng cung cấp một cách tiếp cận mềm dẻo để phân tích dữ liệu theo thời gian và không gian. Phương pháp MCMC cho phép ước lượng phân phối hậu nghiệm khi không có nghiệm giải tích.

Chương 25

Mô hình chuỗi thời gian Bayesian

25.1 Giới thiệu

Chuỗi thời gian Bayesian (Bayesian Time Series) là một phương pháp mô hình hóa dữ liệu chuỗi thời gian sử dụng lý thuyết xác suất Bayes để suy luận về các tham số không quan sát được. Phương pháp này giúp cập nhật niềm tin của chúng ta về một mô hình khi có thêm dữ liệu mới, cho phép linh hoạt hơn so với các phương pháp suy diễn tần suất (frequentist).

25.2 Mô hình tổng quát

Một mô hình chuỗi thời gian Bayesian điển hình có dạng:

$$y_t = f(y_{t-1}, y_{t-2}, ..., \theta) + \varepsilon_t, \tag{25.1}$$

với θ là vector tham số cần ước lượng, và ε_t là nhiễu trắng với $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$.

25.3 Ước lượng tham số bằng Bayes

Theo nguyên lý Bayes, ta có công thức:

$$p(\theta|Y) \propto p(Y|\theta)p(\theta), \tag{25.2}$$

trong đó:

- $p(\theta|Y)$ là phân phối hậu nghiệm (posterior distribution),
- $p(Y|\theta)$ là hàm khả năng (likelihood function),
- $p(\theta)$ là phân phối tiên nghiệm (prior distribution).

25.4 Mô hình tự hồi quy Bayesian (Bayesian AR model)

Một mô hình AR Bayesian bậc p có dạng:

$$y_t = \sum_{i=1}^{p} \phi_i y_{t-i} + \varepsilon_t, \tag{25.3}$$

trong đó ϕ_i là các hệ số hồi quy cần ước lượng.

Lựa chọn phân phối tiên nghiệm:

$$\phi_i \sim \mathcal{N}(0, \tau^2), \tag{25.4}$$

với τ^2 được xác định dựa trên dữ liệu.

25.5 Mô hình động Bayesian (Bayesian Dynamic Models)

Mô hình động có dạng:

$$y_t = X_t\beta_t + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2), \tag{25.5}$$

$$\beta_t = \beta_{t-1} + \eta_t, \quad \eta_t \sim \mathcal{N}(0, Q). \tag{25.6}$$

Quá trình β_t cho phép hệ số thay đổi theo thời gian.

25.6 Phương pháp suy luận trong Mô hình chuỗi thời gian Bayesia

Trong mô hình chuỗi thời gian Bayesian, mục tiêu chính là ước lượng **phân phối hậu nghiệm** $p(\theta|\mathcal{D})$, tức là phân phối của tham số θ sau khi quan sát dữ liệu $\mathcal{D}$. Tuy nhiên, trong hầu hết các trường hợp, phân phối hậu nghiệm **không có dạng đóng**, do đó ta phải sử dụng các phương pháp **lấy mẫu xấp xỉ** hoặc **tối ưu hóa xấp xỉ** để ước lượng nó.

Ba phương pháp chính thường được sử dụng là:

25.6.1 Phương pháp Markov Chain Monte Carlo (MCMC)

MCMC là một phương pháp tạo mẫu từ phân phối hậu nghiệm bằng cách xây dựng một chuỗi Markov có phân phối dừng chính là phân phối hậu nghiệm cần ước lượng.

*** Nguyên lý**

- Bắt đầu từ một giá trị khởi tạo θ_0.
- Tạo một chuỗi các mẫu $\theta_1, \theta_2, \ldots$ sao cho phân phối của các mẫu này tiệm cận phân phối hậu nghiệm $p(\theta|\mathcal{D})$.
- Sử dụng các thuật toán lấy mẫu như **Metropolis-Hastings** hoặc **Gibbs Sampling**.

*** Ưu điểm**

- Chính xác khi số lượng mẫu đủ lớn.

*** Nhược điểm**

- Tính toán chậm, khó hội tụ nếu không chọn hàm đề xuất tốt.

25.6.2 Phương pháp Hamiltonian Monte Carlo (HMC)

HMC là một cải tiến của MCMC, sử dụng **cơ học Hamilton** để di chuyển trong không gian tham số hiệu quả hơn.

*** Nguyên lý**

- Xem tham số θ như một vật thể chuyển động trong không gian dưới ảnh hưởng của một trường lực sinh ra từ hàm xác suất hậu nghiệm.
- Dùng phương trình Hamilton để tính toán chuyển động của tham số, giúp lấy mẫu hiệu quả hơn so với MCMC truyền thống.

*** Ưu điểm**

- Hội tụ nhanh hơn so với MCMC thông thường.
- Ít bị mắc kẹt trong các vùng mật độ thấp.

*** Nhược điểm**

- Cần tính đạo hàm của hàm hậu nghiệm (đắt về mặt tính toán).
- Khó điều chỉnh các siêu tham số.

25.6.3 Phương pháp Variational Bayes (VB)

Variational Bayes là một phương pháp tối ưu hóa thay vì lấy mẫu. Nó tìm một phân phối gần đúng $q(\theta)$ để xấp xỉ hậu nghiệm $p(\theta|\mathcal{D})$ bằng cách **tối ưu hóa một hàm mất mát** gọi là *ELBO* (Evidence Lower Bound).

*** Nguyên lý**

- Xác định một họ phân phối đơn giản $q(\theta)$ để xấp xỉ hậu nghiệm $p(\theta|\mathcal{D})$.
- Tìm $q(\theta)$ tốt nhất bằng cách tối đa hóa ELBO.

*** Ưu điểm**

- Tốc độ nhanh hơn MCMC/HMC.
- Tính toán hiệu quả hơn cho mô hình lớn.

*** Nhược điểm**

- Độ chính xác kém hơn so với MCMC/HMC.
- Phụ thuộc vào lựa chọn của họ phân phối xấp xỉ.

*** Tóm tắt lựa chọn phương pháp**

Phương pháp	Loại tiếp cận	Độ chính xác	Hiệu suất tính toán
MCMC	Lấy mẫu	Cao	Chậm
HMC	Lấy mẫu	Rất cao	Trung bình
VB	Tối ưu hóa	Trung bình	Nhanh

Nếu cần tính toán chính xác cao và có đủ tài nguyên, ta chọn **HMC hoặc MCMC**. Nếu cần tính toán nhanh hơn, có thể chọn **Variational Bayes**.

25.7 Các mô hình chuỗi thời gian Bayesian phổ biến trong kinh tế lượng

25.7.1 Bayesian Vector Autoregression (BVAR)

Mô hình Vector Autoregression (VAR) mở rộng theo hướng Bayesian nhằm giải quyết vấn đề **overfitting** khi số lượng biến kinh tế lớn.

a. Định nghĩa mô hình VAR

Một mô hình VAR bậc p có dạng:

$$Y_t = A_1 Y_{t-1} + A_2 Y_{t-2} + \cdots + A_p Y_{t-p} + \epsilon_t \tag{25.7}$$

với:

- $Y_t \in \mathbb{R}^N$ là vector gồm N biến kinh tế tại thời điểm t.
- A_i là ma trận hệ số $N \times N$.
- $\epsilon_t \sim \mathcal{N}(0, \Sigma)$ là nhiễu trắng với hiệp phương sai Σ.

Viết lại dạng vector hóa:

$$Y = X\beta + \epsilon \tag{25.8}$$

với:

- $Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_T \end{bmatrix}$,

- $X = \begin{bmatrix} Y_0 & Y_{-1} & \dots & Y_{-p+1} \\ Y_1 & Y_0 & \dots & Y_{-p+2} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{T-1} & Y_{T-2} & \dots & Y_{T-p} \end{bmatrix}$,
- β là vector hóa của tất cả A_i,
- $\epsilon \sim \mathcal{N}(0, I_T \otimes \Sigma)$.

b. Hồi quy Bayesian với tiên nghiệm Minnesota

Để tránh overfitting, ta áp đặt tiên nghiệm lên β:

$$\beta \sim \mathcal{N}(\mu_\beta, \Omega_\beta) \tag{25.9}$$

- **Minnesota prior** giả định các hệ số của biến trễ cao hơn có phương sai nhỏ hơn:
$$\text{Var}(A_{i,jk}) = \frac{\lambda^2}{i^2}, \quad i = 1, 2, \dots, p \tag{25.10}$$
với λ là tham số điều chỉnh độ chặt của tiên nghiệm.
- Tiên nghiệm về ma trận hiệp phương sai nhiễu:
$$\Sigma \sim \text{Inverse-Wishart}(\Psi, \nu) \tag{25.11}$$
- Phân phối hậu nghiệm của β và Σ:
$$p(\beta, \Sigma|Y) \propto p(Y|\beta, \Sigma)p(\beta)p(\Sigma) \tag{25.12}$$

Ước lượng bằng **Gibbs Sampling** để lấy mẫu từ phân phối hậu nghiệm.

25.7.2 Bayesian State-Space Models

Mô hình **Bayesian State-Space** mô tả động thái của các trạng thái ẩn trong nền kinh tế.

a. Định nghĩa mô hình

Mô hình gồm hai phương trình:

- **Phương trình trạng thái (State Equation)**:
$$S_t = FS_{t-1} + \eta_t, \quad \eta_t \sim \mathcal{N}(0, Q) \tag{25.13}$$
- **Phương trình quan sát (Observation Equation)**:
$$Y_t = HS_t + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, R) \tag{25.14}$$

trong đó:

- S_t là trạng thái ẩn của nền kinh tế.
- Y_t là dữ liệu quan sát được.
- F, H là ma trận hệ số.
- η_t, ϵ_t là nhiễu Gaussian.

b. Bayesian Estimation

Tiên nghiệm:

$$S_0 \sim \mathcal{N}(\mu_0, \Sigma_0) \tag{25.15}$$

$$Q \sim \text{Inverse-Wishart}(\Psi_Q, \nu_Q) \tag{25.16}$$

$$R \sim \text{Inverse-Wishart}(\Psi_R, \nu_R) \tag{25.17}$$

Ước lượng bằng **Particle Filtering** hoặc **Markov Chain Monte Carlo (MCMC)**.

Mô hình chuỗi thời gian Bayesian cung cấp một cách tiếp cận linh hoạt và mạnh mẽ để xử lý dữ liệu chuỗi thời gian, đặc biệt khi có sự không chắc chắn về tham số hoặc khi các tham số có thể thay đổi theo thời gian.

Chương 26

Ước lượng Bayesian với dữ liệu nhỏ hoặc không đầy đủ

26.1 Giới thiệu

Ước lượng Bayesian cung cấp một phương pháp mạnh mẽ để xử lý dữ liệu nhỏ hoặc không đầy đủ bằng cách kết hợp thông tin tiên nghiệm (prior information) với dữ liệu quan sát được để suy luận về tham số. Khi dữ liệu ít hoặc bị thiếu, mô hình Bayesian có thể sử dụng tiên nghiệm thích hợp để cải thiện độ chính xác của ước lượng.

Gọi θ là tham số cần ước lượng, và D là tập dữ liệu quan sát được (có thể nhỏ hoặc không đầy đủ). Theo quy tắc Bayes:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} \tag{26.1}$$

trong đó:
- $p(\theta)$ là tiên nghiệm về θ,
- $p(D|\theta)$ là xác suất có điều kiện của dữ liệu (likelihood),
- $p(D)$ là hệ số chuẩn hóa, còn gọi là **bằng chứng** (evidence),
- $p(\theta|D)$ là phân phối hậu nghiệm của tham số.

26.2 Chọn tiên nghiệm phù hợp với dữ liệu nhỏ

Khi dữ liệu nhỏ, lựa chọn tiên nghiệm đóng vai trò quan trọng vì nó có thể ảnh hưởng mạnh đến kết quả. Một số lựa chọn phổ biến:

26.2.1 Tiên nghiệm không thông tin (Non-informative prior)

- Khi không có nhiều thông tin trước, ta chọn tiên nghiệm phẳng (uniform) hoặc Jeffreys prior:

$$p(\theta) \propto 1 \tag{26.2}$$

hoặc

$$p(\theta) \propto \frac{1}{\sqrt{I(\theta)}} \tag{26.3}$$

với $I(\theta)$ là thông tin Fisher.

26.2.2 Tiên nghiệm thông tin (Informative prior)

- Nếu có thông tin trước về tham số, ta dùng phân phối Gaussian, Gamma, Beta, v.v.
- Ví dụ: Nếu θ là trung bình của phân phối chuẩn, ta có thể chọn tiên nghiệm Gaussian:

$$\theta \sim \mathcal{N}(\mu_0, \sigma_0^2) \tag{26.4}$$

26.2.3 Tiên nghiệm co rút (Shrinkage prior)

- Khi số lượng tham số lớn hơn số quan sát, ta sử dụng tiên nghiệm Laplace (Lasso Bayesian) hoặc tiên nghiệm Gaussian chuẩn hóa (Ridge Bayesian):

$$p(\theta) \propto e^{-\lambda|\theta|} \tag{26.5}$$

hoặc

$$p(\theta) \propto e^{-\lambda\theta^2} \tag{26.6}$$

26.3 Ước lượng Bayesian với dữ liệu không đầy đủ

Khi dữ liệu bị thiếu, ta cần tích hợp phân phối hậu nghiệm theo những giá trị không quan sát được.

26.3.1 Phương pháp tích phân biên (Marginalization)

Giả sử dữ liệu bị thiếu là X_m, dữ liệu quan sát là X_o. Khi đó, ta tính hậu nghiệm biên:

$$p(\theta|X_o) = \int p(\theta|X_o, X_m)p(X_m|X_o)dX_m \tag{26.7}$$

26.3.2 Phương pháp Gibbs Sampling (MCMC)

1. Lấy mẫu $X_m^{(t+1)} \sim p(X_m|\theta^{(t)}, X_o)$
2. Lấy mẫu $\theta^{(t+1)} \sim p(\theta|X_o, X_m^{(t+1)})$

26.3.3 Phương pháp EM Bayesian (Expectation-Maximization Bayesian)

1. **Bước E**: Tính kỳ vọng hậu nghiệm có điều kiện $Q(\theta) = E[\log p(X|\theta)|X_o]$.
2. **Bước M**: Cập nhật tham số bằng cách tối đa hóa:

$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta)p(\theta) \tag{26.8}$$

26.4 Ví dụ: Ước lượng trung bình với dữ liệu nhỏ hoặc bị thiếu

Giả sử ta có một tập dữ liệu nhỏ $X = \{x_1, x_2, ..., x_n\}$ từ phân phối chuẩn $X_i \sim \mathcal{N}(\mu, \sigma^2)$, với σ^2 đã biết.

1. **Chọn tiên nghiệm**:

$$\mu \sim \mathcal{N}(\mu_0, \tau_0^2) \tag{26.9}$$

2. **Tính phân phối hậu nghiệm**:

$$p(X|\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \tag{26.10}$$

$$\mu|X \sim \mathcal{N}(\mu_n, \tau_n^2) \tag{26.11}$$

với:

$$\mu_n = \frac{\frac{\mu_0}{\tau_0^2} + \frac{n\bar{x}}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \tag{26.12}$$

$$\tau_n^2 = \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \tag{26.13}$$

26.5 Ví dụ minh họa: Ước Lượng Bayesian trong kinh tế lượng

Giả sử chúng ta có một bộ dữ liệu nhỏ về mối quan hệ giữa **tăng trưởng GDP** (Y) và **chi tiêu chính phủ** (X) của một số quốc gia trong một khoảng thời gian ngắn. Vì số quan sát ít, ước lượng theo phương pháp truyền thống (OLS) có thể không đáng tin cậy. Ta sẽ sử dụng **hồi quy Bayesian** để cải thiện kết quả.

26.5.1 Mô hình Hồi Quy Bayesian

Mô hình hồi quy tuyến tính có dạng:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{26.14}$$

với:

- Y_i là tăng trưởng GDP của quốc gia i.
- X_i là chi tiêu chính phủ của quốc gia i.
- β_0, β_1 là các tham số cần ước lượng.
- ϵ_i là sai số ngẫu nhiên tuân theo phân phối chuẩn với phương sai σ^2.

26.5.2 Thiết Lập Phân Phối Tiên Nghiệm (Prior)

Vì dữ liệu nhỏ, ta sử dụng phân phối tiên nghiệm để đưa thêm thông tin từ các nghiên cứu trước. Giả sử:

$$\begin{aligned}\beta_0 &\sim \mathcal{N}(0, 10)\\ \beta_1 &\sim \mathcal{N}(0.5, 0.2)\\ \sigma^2 &\sim \text{Inverse-Gamma}(2, 2)\end{aligned}$$

26.5.3 Tính Phân Phối Hậu Nghiệm (Posterior)

Theo định lý Bayes, ta có:

$$P(\beta_0, \beta_1, \sigma^2|Y, X) \propto P(Y|X, \beta_0, \beta_1, \sigma^2)P(\beta_0, \beta_1, \sigma^2) \tag{26.15}$$

với $P(Y|X, \beta_0, \beta_1, \sigma^2)$ là hàm hợp lý (likelihood) của mô hình hồi quy chuẩn.

Vì phân phối hậu nghiệm không có dạng đóng (closed-form), ta phải sử dụng **phương pháp MCMC (Markov Chain Monte Carlo)** để lấy mẫu từ hậu nghiệm.

26.5.4 Ước Lượng Bằng Gibbs Sampling

Quy trình Gibbs Sampling gồm các bước:

1. **Khởi tạo giá trị ban đầu** cho $\beta_0, \beta_1, \sigma^2$.
2. **Cập nhật β_0 và β_1 từ phân phối có điều kiện**:

$$P(\beta_0, \beta_1|Y, X, \sigma^2) = \mathcal{N}(\mu, \Sigma) \tag{26.16}$$

   với μ, Σ là các tham số ước lượng từ dữ liệu.

3. **Cập nhật σ^2 từ phân phối có điều kiện**:

$$P(\sigma^2|Y, X, \beta_0, \beta_1) = \text{Inverse-Gamma}(\alpha', \beta') \tag{26.17}$$

4. **Lặp lại** bước 2 và 3 cho đến khi đạt hội tụ.

26.5.5 Kết Quả và Ứng Dụng

Sau khi chạy MCMC, ta có thể lấy giá trị trung bình của các mẫu β_0, β_1 để có ước lượng tốt hơn về tác động của chi tiêu chính phủ đến tăng trưởng GDP. Với Bayesian, ta còn có **độ tin cậy của ước lượng** thông qua phân phối hậu nghiệm, giúp tránh overfitting khi dữ liệu nhỏ.

26.5.6 Ứng dụng thực tế

- Phân tích tác động của chính sách tài khóa khi dữ liệu chưa đủ lớn.
- Dự báo tăng trưởng GDP với ít quan sát.
- Hỗ trợ ra quyết định khi dữ liệu kinh tế bị thiếu hoặc không hoàn chỉnh.

26.6 Kết luận

- Bayesian inference rất phù hợp khi dữ liệu nhỏ hoặc bị thiếu.
- Tiên nghiệm đóng vai trò quan trọng trong việc điều chỉnh kết quả ước lượng.
- Các phương pháp như Gibbs Sampling và EM Bayesian giúp xử lý dữ liệu không đầy đủ.

Chương 27

Phương pháp MCMC trong Kinh tế lượng Bayesian

27.1 Tổng quan về MCMC trong kinh tế lượng Bayesian

Trong bối cảnh kinh tế lượng Bayesian, chúng ta quan tâm đến việc ước lượng **phân phối hậu nghiệm** của tham số θ dựa trên dữ liệu quan sát $\mathcal{D}$. Theo định lý Bayes:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} \tag{27.1}$$

trong đó:

- $p(\theta|\mathcal{D})$: Phân phối hậu nghiệm của tham số θ.
- $p(\mathcal{D}|\theta)$: Hàm hợp lý (likelihood).
- $p(\theta)$: Phân phối tiên nghiệm (prior).
- $p(\mathcal{D}) = \int p(\mathcal{D}|\theta)p(\theta)d\theta$: Hệ số chuẩn hóa.

Vấn đề chính là tích phân trên mẫu không khả thi trong hầu hết các trường hợp, do đó chúng ta cần sử dụng **MCMC** để lấy mẫu từ phân phối hậu nghiệm.

27.2 Phương pháp Markov Chain Monte Carlo (MCMC)

MCMC là một tập hợp các thuật toán dùng để lấy mẫu từ phân phối hậu nghiệm bằng cách xây dựng một **chuỗi Markov** hội tụ đến phân phối đích $p(\theta|\mathcal{D})$. Hai thuật toán phổ biến trong kinh tế lượng Bayesian là:

- **Thuật toán Metropolis-Hastings**
- **Thuật toán Gibbs Sampling**

27.3 Thuật toán Metropolis-Hastings (MH)

Thuật toán Metropolis-Hastings hoạt động như sau:

1. Khởi tạo giá trị ban đầu $\theta^{(0)}$.
2. **Lặp lại** cho $t = 1, 2, \ldots, T$:
   (a) Sinh **đề xuất** $\theta^* \sim q(\theta^*|\theta^{(t-1)})$.
   (b) Tính toán **tỷ số chấp nhận**:
   $$r = \frac{p(\mathcal{D}|\theta^*)p(\theta^*)}{p(\mathcal{D}|\theta^{(t-1)})p(\theta^{(t-1)})} \times \frac{q(\theta^{(t-1)}|\theta^*)}{q(\theta^*|\theta^{(t-1)})} \tag{27.2}$$
   (c) Chấp nhận θ^* với xác suất $\alpha = \min(1, r)$:
   $$\theta^{(t)} = \begin{cases} \theta^* & \text{với xác suất } \alpha \\ \theta^{(t-1)} & \text{với xác suất } 1-\alpha \end{cases} \tag{27.3}$$

27.4 Thuật toán Gibbs Sampling

Gibbs Sampling cập nhật từng tham số θ_j theo quy tắc:

1. Khởi tạo giá trị $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_d^{(0)})$.
2. **Lặp lại** cho $t = 1, 2, \ldots, T$:
   (a) Cập nhật từng thành phần $j = 1, 2, \ldots, d$ theo phân phối điều kiện:
   $$\theta_j^{(t)} \sim p(\theta_j|\theta_1^{(t)}, \ldots, \theta_{j-1}^{(t)}, \theta_{j+1}^{(t-1)}, \ldots, \theta_d^{(t-1)}, \mathcal{D}) \tag{27.4}$$

27.5 Ứng dụng MCMC vào kinh tế lượng Bayesian

27.5.1 Ví dụ: Mô hình hồi quy Bayesian

Xét mô hình hồi quy tuyến tính:

$$y_i = X_i\beta + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{27.5}$$

Với:

- y: Biến phụ thuộc.
- X: Ma trận biến độc lập.
- β: Vector hệ số hồi quy.

Phân phối tiên nghiệm: Chọn phân phối tiên nghiệm Gaussian cho β:

$$\beta \sim \mathcal{N}(\beta_0, \Sigma_0) \tag{27.6}$$

Phân phối hậu nghiệm: Không có dạng đóng, cần MCMC để lấy mẫu từ:

$$p(\beta|X, y) \propto p(y|X, \beta)p(\beta) \tag{27.7}$$

27.5.2 MCMC với Gibbs Sampling

Khi giả định σ^2 có tiên nghiệm inverse-Gamma, ta có:

$$p(\beta|\sigma^2, X, y) \sim \mathcal{N}(\hat{\beta}, V_\beta) \tag{27.8}$$

$$p(\sigma^2|\beta, X, y) \sim \text{Inverse-Gamma}(\alpha, \beta) \tag{27.9}$$

Gibbs Sampling thực hiện:

1. Lấy mẫu β từ phân phối Gaussian hậu nghiệm.
2. Lấy mẫu σ^2 từ phân phối inverse-Gamma.
3. Lặp lại cho đến khi hội tụ.

27.6 Tổng kết

MCMC là phương pháp phổ biến trong kinh tế lượng Bayesian, giúp lấy mẫu từ phân phối hậu nghiệm trong các mô hình phức tạp.

Phần VI

Phân tích dữ liệu chuỗi thời gian (Time Series Data)

Chương 28

Tổng quan về chuỗi thời gian

28.1 Các khái niệm cơ bản: tính dừng, tự tương quan, mùa vụ

28.1.1 Tính Dừng (Stationarity)

Một chuỗi thời gian $\{Y_t\}$ được gọi là dừng (stationary) nếu các đặc điểm thống kê của nó không thay đổi theo thời gian. Điều này có nghĩa là:

- Giá trị kỳ vọng $E(Y_t)$ không thay đổi theo thời gian:

$$E(Y_t) = \mu, \quad \forall t. \tag{28.1}$$

- Phương sai $\text{Var}(Y_t)$ là hằng số:

$$\text{Var}(Y_t) = \sigma^2, \quad \forall t. \tag{28.2}$$

- Hàm tự hiệp phương sai $\gamma(k) = \text{Cov}(Y_t, Y_{t+k})$ chỉ phụ thuộc vào độ trễ k mà không phụ thuộc vào thời điểm t:

$$\gamma(k) = E[(Y_t - \mu)(Y_{t+k} - \mu)]. \tag{28.3}$$

Chuỗi thời gian không dừng có thể được chuyển thành chuỗi dừng bằng phương pháp sai phân (differencing):

$$\Delta Y_t = Y_t - Y_{t-1}. \tag{28.4}$$

28.1.2 Tự Tương Quan (Autocorrelation)

Tự tương quan đo lường mối quan hệ giữa các quan sát của chuỗi thời gian tại các thời điểm khác nhau. Hệ số tự tương quan bậc k (Autocorrelation Function - ACF) được định nghĩa như sau:

$$\rho_k = \frac{\gamma(k)}{\gamma(0)} = \frac{E[(Y_t - \mu)(Y_{t+k} - \mu)]}{\sigma^2}, \quad k = 0, 1, 2, \ldots \tag{28.5}$$

Một chuỗi thời gian có thể có tự tương quan dương ($\rho_k > 0$) hoặc tự tương quan âm ($\rho_k < 0$).

Bài kiểm định Durbin-Watson thường được sử dụng để kiểm tra tự tương quan trong mô hình hồi quy:

$$DW = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}, \tag{28.6}$$

trong đó e_t là phần dư của mô hình hồi quy.

28.1.3 Tính Mùa Vụ (Seasonality)

Mùa vụ xảy ra khi một chuỗi thời gian có các mẫu lặp lại theo chu kỳ cố định. Một mô hình phổ biến để mô tả tính mùa vụ là:

$$Y_t = T_t + S_t + I_t, \tag{28.7}$$

trong đó:

- T_t: thành phần xu hướng (trend component),
- S_t: thành phần mùa vụ (seasonal component),
- I_t: thành phần ngẫu nhiên (irregular component).

Mô hình hồi quy với biến giả mùa vụ có dạng:

$$Y_t = \beta_0 + \sum_{i=1}^{s-1} \beta_i D_{it} + \epsilon_t, \tag{28.8}$$

trong đó D_{it} là các biến giả đại diện cho các giai đoạn trong chu kỳ mùa vụ.

28.2 Biểu diễn và phân tích dữ liệu chuỗi thời gian

28.2.1 Giới thiệu

Dữ liệu chuỗi thời gian là tập hợp các quan sát được thu thập theo thứ tự thời gian. Phân tích chuỗi thời gian nhằm mục đích mô tả, mô hình hóa và dự báo dữ liệu trong tương lai.

28.2.2 Biểu diễn dữ liệu chuỗi thời gian

Dữ liệu chuỗi thời gian được ký hiệu là $\{Y_t\}_{t=1}^{T}$, trong đó Y_t là giá trị quan sát tại thời điểm t. Một số phương pháp biểu diễn dữ liệu chuỗi thời gian:

- **Biểu đồ chuỗi thời gian (Time Series Plot)**: Trực quan hóa sự thay đổi của Y_t theo thời gian t.
- **Biểu đồ phân phối (Histogram)**: Phân bố xác suất của chuỗi dữ liệu.
- **Biểu đồ tự tương quan (ACF - Autocorrelation Function)**: Biểu diễn mức độ tương quan của một quan sát với chính nó ở các độ trễ khác nhau.

- **Biểu đồ PACF (Partial Autocorrelation Function)**: Tính toán tương quan giữa các quan sát sau khi loại bỏ ảnh hưởng của các quan sát trung gian.

28.2.3 Phân tích dữ liệu chuỗi thời gian

Kiểm tra tính dừng

Tính dừng của chuỗi thời gian rất quan trọng trong việc xây dựng mô hình dự báo. Một chuỗi thời gian được gọi là dừng nếu:

- Kỳ vọng $E(Y_t) = \mu$ không thay đổi theo thời gian.
- Phương sai $Var(Y_t) = \sigma^2$ là hằng số.
- Hàm tự tương quan $\gamma(k) = Cov(Y_t, Y_{t-k})$ chỉ phụ thuộc vào độ trễ k.

Kiểm định tính dừng phổ biến:

- Kiểm định Dickey-Fuller (ADF Test): H_0: Chuỗi có gốc đơn vị (không dừng).
- Kiểm định KPSS: H_0: Chuỗi là dừng.

Phân tích tự tương quan

Tự tương quan đo lường mức độ tương quan giữa các quan sát trong chuỗi thời gian:

$$\rho_k = \frac{Cov(Y_t, Y_{t-k})}{Var(Y_t)} \tag{28.9}$$

Hàm tự tương quan ACF giúp phát hiện tính dừng và nhận dạng mô hình ARIMA.

Phân tích mùa vụ

Mùa vụ thể hiện sự lặp lại có chu kỳ trong dữ liệu chuỗi thời gian. Mô hình phổ biến:

- Mô hình hồi quy với biến giả mùa vụ:

$$Y_t = \beta_0 + \sum_{j=1}^{s-1} \beta_j D_{jt} + \epsilon_t \tag{28.10}$$

  trong đó D_{jt} là biến giả mùa vụ.

- Mô hình SARIMA: Mở rộng ARIMA với thành phần mùa vụ $(p, d, q) \times (P, D, Q)_s$.

28.2.4 Kết luận

Biểu diễn và phân tích chuỗi thời gian là bước quan trọng giúp nhận diện tính chất của dữ liệu, từ đó lựa chọn mô hình phù hợp để dự báo.

Chương 29

Mô hình ARIMA và các biến thể

29.1 Mô hình AR, MA, ARMA, ARIMA

29.1.1 Mô hình AR (AutoRegressive)

Mô hình tự hồi quy bậc p (AR(p)) được biểu diễn như sau:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \epsilon_t, \tag{29.1}$$

trong đó:

- Y_t là giá trị của chuỗi thời gian tại thời điểm t.
- ϕ_i là hệ số của mô hình.
- ϵ_t là nhiễu trắng với $\mathbb{E}[\epsilon_t] = 0$ và $\text{Var}(\epsilon_t) = \sigma^2$.

29.1.2 Mô hình MA (Moving Average)

Mô hình trung bình trượt bậc q (MA(q)) được biểu diễn như sau:

$$Y_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}. \tag{29.2}$$

29.1.3 Mô hình ARMA

Mô hình kết hợp giữa AR và MA bậc (p, q) (ARMA(p, q)) được biểu diễn như:

$$Y_t = \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t. \tag{29.3}$$

29.1.4 Mô hình ARIMA

Mô hình ARIMA(p, d, q) mở rộng ARMA bằng cách thêm phép sai phân:

$$(1 - B)^d Y_t = \sum_{i=1}^{p} \phi_i (1 - B)^d Y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t, \tag{29.4}$$

trong đó B là toán tử trễ ($BY_t = Y_{t-1}$).

29.2 Phương pháp chọn bậc mô hình tối ưu

29.2.1 Tiêu chí AIC (Akaike Information Criterion)

$$\text{AIC} = -2\ln(L) + 2k, \tag{29.5}$$

trong đó:

- L là giá trị cực đại của hàm hợp lý.
- k là số tham số trong mô hình.

29.2.2 Tiêu chí BIC (Bayesian Information Criterion)

$$\text{BIC} = -2\ln(L) + k\ln(n), \tag{29.6}$$

trong đó n là số quan sát.

29.3 Dự báo bằng ARIMA

Dự báo bằng mô hình ARIMA sử dụng phương pháp ngoại suy:

$$\hat{Y}_{t+h} = E[Y_{t+h}|Y_t, Y_{t-1}, \dots] \tag{29.7}$$

Phương pháp phổ biến là Box-Jenkins, gồm các bước:

1. Kiểm tra tính dừng.
2. Xác định bậc p, d, q.
3. Ước lượng tham số.
4. Kiểm tra mô hình.
5. Dự báo.

Chương 30

Mô hình ARCH/GARCH

30.1 Biến động tài chính và mô hình ARCH/GARCH

Trong lĩnh vực tài chính, các chuỗi thời gian thường có đặc điểm *phương sai không cố định* (heteroskedasticity), nghĩa là mức độ biến động của dữ liệu thay đổi theo thời gian. Mô hình ARCH (Autoregressive Conditional Heteroskedasticity) và GARCH (Generalized Autoregressive Conditional Heteroskedasticity) được sử dụng để mô tả hiện tượng này.

30.1.1 Mô hình ARCH

Giả sử một chuỗi thời gian $\{r_t\}$ được mô tả bởi phương trình:

$$r_t = \mu + \epsilon_t, \tag{30.1}$$

với ϵ_t là sai số ngẫu nhiên có kỳ vọng bằng 0, nhưng phương sai có thể thay đổi theo thời gian. Mô hình ARCH(p) giả định rằng phương sai có dạng:

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i \epsilon_{t-i}^2, \tag{30.2}$$

với $\alpha_0 > 0$, $\alpha_i \geq 0$ để đảm bảo phương sai luôn dương.

30.1.2 Mô hình GARCH

Mô hình GARCH(p, q) mở rộng ARCH bằng cách thêm vào các bậc trễ của chính phương sai có điều kiện:

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2, \tag{30.3}$$

với $\alpha_i \geq 0$, $\beta_j \geq 0$, và $\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j < 1$ để đảm bảo tính ổn định.

30.2 Ước lượng tham số và dự báo biến động

30.2.1 Ước lượng tham số

Thông thường, các tham số của mô hình ARCH/GARCH được ước lượng bằng phương pháp *hợp lý tối đa* (Maximum Likelihood Estimation - MLE). Hàm hợp lý có dạng:

$$L(\theta) = -\frac{1}{2}\sum_{t=1}^{T}\left(\log(2\pi) + \log(\sigma_t^2) + \frac{\epsilon_t^2}{\sigma_t^2}\right), \tag{30.4}$$

trong đó θ là vector chứa các tham số cần ước lượng.

30.2.2 Dự báo biến động

Dự báo phương sai có điều kiện được thực hiện bằng cách sử dụng phương trình hồi quy của GARCH. Ví dụ, với mô hình GARCH(1,1), phương sai kỳ vọng tại thời điểm $t+1$ được tính bằng:

$$\mathbb{E}[\sigma_{t+1}^2|\mathcal{F}_t] = \alpha_0 + \alpha_1\epsilon_t^2 + \beta_1\sigma_t^2. \tag{30.5}$$

Dự báo nhiều bước có thể được tính đệ quy bằng cách lặp lại công thức trên.

Phần VII

Kinh tế lượng không gian (Spatial Econometrics)

Chương 31

Tổng quan về kinh tế lượng không gian

31.1 Khái niệm và tầm quan trọng của kinh tế lượng không gian

31.1.1 Khái niệm

Kinh tế lượng không gian là nhánh của kinh tế lượng tập trung vào các mô hình thống kê có tính đến mối quan hệ không gian giữa các đơn vị quan sát. Trong bối cảnh này, dữ liệu không gian có thể là:

- **Dữ liệu vùng (areal data)**: Thuộc tính của các vùng địa lý (GDP theo tỉnh, tỉ lệ thất nghiệp theo quận).
- **Dữ liệu điểm (point data)**: Dữ liệu có tọa độ cụ thể (vị trí doanh nghiệp, điểm đo ô nhiễm).
- **Dữ liệu dòng chảy (flow data)**: Biểu thị sự di chuyển giữa các khu vực (dòng di cư, dòng vốn đầu tư).

Khi dữ liệu có tính không gian, giả định tính độc lập của phần dư trong hồi quy tuyến tính OLS bị vi phạm do tồn tại tự tương quan không gian.

31.1.2 Tầm quan trọng

Kinh tế lượng không gian quan trọng vì:

- Hầu hết hiện tượng kinh tế - xã hội đều có sự liên kết không gian.
- Mô hình hồi quy truyền thống bỏ qua yếu tố không gian có thể dẫn đến ước lượng chệch.
- Cung cấp công cụ phân tích chính xác hơn cho các nhà hoạch định chính sách.

31.2 Ứng dụng thực tế trong kinh tế, xã hội và môi trường

31.2.1 Kinh tế

- Phát triển vùng: Nghiên cứu tác động lan tỏa của đầu tư công.
- Bất động sản: Định giá nhà dựa trên đặc điểm không gian.

31.2.2 Xã hội

- Dịch tễ học: Phân tích sự lây lan của dịch bệnh.
- Bất bình đẳng thu nhập: Tương quan không gian của mức sống.

31.2.3 Môi trường

- Ô nhiễm không khí: Dòng chảy không khí mang theo chất ô nhiễm.
- Biến đổi khí hậu: Ảnh hưởng lan tỏa của thiên tai.

31.3 Sự khác biệt giữa kinh tế lượng truyền thống và kinh tế lượng không gian

31.3.1 Mô hình hồi quy tuyến tính truyền thống

Mô hình OLS chuẩn có dạng:

$$Y = X\beta + \varepsilon \tag{31.1}$$

với giả định không có tự tương quan không gian.

31.3.2 Mô hình kinh tế lượng không gian

Mô hình Spatial Lag (SLM):

$$Y = \rho WY + X\beta + \varepsilon \tag{31.2}$$

Mô hình Spatial Error (SEM):

$$Y = X\beta + u, \quad u = \lambda Wu + \varepsilon \tag{31.3}$$

31.4 Các thách thức khi phân tích dữ liệu không gian

31.4.1 Xác định mối quan hệ không gian

Lựa chọn ma trận trọng số không gian W rất quan trọng, các phương pháp phổ biến:

- Láng giềng k gần nhất (k-nearest neighbors)
- Khoảng cách nghịch đảo (inverse distance weighting)
- Dựa trên biên giới hành chính (contiguity-based weights)

31.4.2 Kiểm định tự tương quan không gian

Moran's I:

$$I = \frac{N}{\sum_i \sum_j W_{ij}} \times \frac{\sum_i \sum_j W_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{\sum_i (Y_i - \bar{Y})^2} \tag{31.4}$$

Geary's C:

$$C = \frac{(N-1)\sum_i \sum_j W_{ij}(Y_i - Y_j)^2}{2\sum_i (Y_i - \bar{Y})^2} \tag{31.5}$$

31.4.3 Lựa chọn mô hình thích hợp

Cần chọn mô hình phù hợp dựa trên kiểm định:

- Kiểm định Lagrange Multiplier (LM) để quyết định giữa SLM và SEM.
- So sánh Akaike Information Criterion (AIC) giữa các mô hình.

31.5 Tóm tắt chương

- Kinh tế lượng không gian quan trọng vì nhiều hiện tượng kinh tế có sự lan tỏa theo không gian.
- Khác với mô hình OLS truyền thống, mô hình không gian đưa vào ma trận trọng số W.
- Có hai mô hình chính: Spatial Lag Model (SLM) và Spatial Error Model (SEM).
- Phân tích dữ liệu không gian gặp nhiều thách thức.

Chương 32

Các Khái Niệm Cơ Bản trong Kinh tế lượng Không gian

32.1 Sự Phụ Thuộc Không Gian (Spatial Dependence)

Định nghĩa: Sự phụ thuộc không gian là hiện tượng giá trị của một biến tại một vị trí bị ảnh hưởng bởi giá trị của biến đó tại các vị trí lân cận.

Biểu diễn toán học:

$$y_i = f(y_j), \quad \forall j \in N(i) \tag{32.1}$$

Trong đó:

- y_i là giá trị của biến quan sát tại vị trí i.
- $N(i)$ là tập hợp các vị trí lân cận của i.

Mô hình cơ bản: Sự phụ thuộc không gian có thể được mô hình hóa bằng một **ma trận trọng số không gian** W, trong đó:

$$Y = WY + \varepsilon \tag{32.2}$$

với W là ma trận trọng số không gian, ε là nhiễu.

32.2 Tự Tương Quan Không Gian (Spatial Autocorrelation)

Định nghĩa: Hiện tượng các quan sát không gian có tương quan với nhau theo một mô hình xác định.

Chỉ số Moran's I:

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \times \frac{\sum_i \sum_j w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_i (y_i - \bar{y})^2} \tag{32.3}$$

Trong đó:

- w_{ij} là phần tử trong ma trận trọng số không gian W.

- y_i, y_j là giá trị của biến quan sát tại vị trí i, j.
- $\bar{y}$ là giá trị trung bình của biến y.

Ý nghĩa: Nếu $I > 0$ thì có tự tương quan dương (các giá trị gần nhau có xu hướng giống nhau). Nếu $I < 0$ thì có tự tương quan âm (các giá trị gần nhau có xu hướng khác nhau).

32.3 Ma Trận Trọng Số Không Gian (Spatial Weight Matrix)

32.3.1 Các phương pháp xây dựng ma trận trọng số không gian

Ma trận trọng số không gian W được xây dựng dựa trên mối quan hệ giữa các khu vực theo các phương pháp khác nhau.

32.3.2 Ma trận k-nearest neighbors

Xây dựng dựa trên k láng giềng gần nhất của mỗi điểm. Mỗi phần tử w_{ij} của ma trận W được xác định như sau:

$$w_{ij} = \begin{cases} 1, & \text{nếu } j \text{ là một trong } k \text{ láng giềng gần nhất của } i \\ 0, & \text{ngược lại} \end{cases} \tag{32.4}$$

32.3.3 Ma trận khoảng cách nghịch đảo

Trọng số giữa hai điểm i và j được xác định theo khoảng cách:

$$w_{ij} = \frac{1}{d_{ij}^{\alpha}} \tag{32.5}$$

Với:

- d_{ij} là khoảng cách giữa điểm i và j.
- α là một số mũ (thường lấy $\alpha = 2$).

32.3.4 Ma trận Queen và Rook

- **Ma trận Queen:** Hai vùng được coi là láng giềng nếu chúng có **chung cạnh hoặc chung đỉnh**.
- **Ma trận Rook:** Hai vùng được coi là láng giềng nếu chúng **chỉ có chung cạnh**.

Chương 33

Các Mô Hình Kinh Tế Lượng Không Gian

33.1 Mô hình hồi quy không gian tuyến tính (SLM)

33.1.1 Công thức và ý nghĩa

Mô hình hồi quy không gian tuyến tính (Spatial Lag Model - SLM) được biểu diễn bởi phương trình:

$$y = \rho W y + X\beta + \varepsilon, \tag{33.1}$$

trong đó:

- y là vector kết quả (biến phụ thuộc).
- W là ma trận trọng số không gian.
- ρ là hệ số tự tương quan không gian.
- X là ma trận biến giải thích.
- β là vector hệ số hồi quy.
- ε là nhiễu trắng có phân phối chuẩn $N(0, \sigma^2 I)$.

33.1.2 Ứng dụng của SLM

SLM được ứng dụng rộng rãi trong các lĩnh vực như:

- Định giá bất động sản dựa trên ảnh hưởng từ các khu vực lân cận.
- Phân tích tác động lan truyền của chính sách kinh tế.
- Nghiên cứu dịch tễ học để xác định mô hình lây lan bệnh dịch.

33.2 Mô hình sai số không gian (SEM)

33.2.1 Công thức và ý nghĩa

Mô hình sai số không gian (Spatial Error Model - SEM) được biểu diễn bởi phương trình:

$$y = X\beta + u, \tag{33.2}$$

trong đó:

$$u = \lambda W u + \varepsilon. \tag{33.3}$$

Thành phần sai số u có cấu trúc không gian thông qua ma trận trọng số W.

33.2.2 Khi nào nên sử dụng SEM?

SEM phù hợp khi có sự phụ thuộc không gian trong sai số nhưng không nhất thiết trong biến phụ thuộc.

33.3 Mô hình Durbin không gian (SDM)

33.3.1 Công thức tổng quát

Mô hình Durbin không gian (Spatial Durbin Model - SDM) mở rộng từ SLM:

$$y = \rho W y + X\beta + WX\theta + \varepsilon. \tag{33.4}$$

33.3.2 Sự khác biệt giữa SDM và SLM

SDM bao gồm các biến giải thích có tác động lan truyền không gian thông qua WX.

33.4 Các mô hình mở rộng khác

- Mô hình SAR: $y = \rho W y + X\beta + \varepsilon$.
- Mô hình SLX: $y = X\beta + WX\theta + \varepsilon$.
- Mô hình GWR: Hồi quy địa phương với tham số thay đổi theo vị trí.
- Mô hình Bayesian Spatial: Sử dụng Bayesian inference để ước lượng mô hình không gian.

Chương 34

Kiểm định và Phương pháp Ước lượng

34.1 Kiểm định Moran's I

Moran's I là một kiểm định thống kê dùng để phát hiện tự tương quan không gian trong dữ liệu. Công thức của Moran's I được biểu diễn như sau:

$$I = \frac{N}{W} \frac{\sum_i \sum_j w_{ij}(X_i - \bar{X})(X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2} \tag{34.1}$$

trong đó:

- N là số quan sát.
- X_i và X_j là giá trị của biến tại vị trí i và j.
- $\bar{X}$ là giá trị trung bình của biến.
- w_{ij} là phần tử của ma trận trọng số không gian.
- $W = \sum_i \sum_j w_{ij}$ là tổng các phần tử của ma trận trọng số.

34.2 Kiểm định Lagrange Multiplier (LM)

LM là một kiểm định để xác định mô hình phù hợp giữa Spatial Lag Model (SLM) và Spatial Error Model (SEM). Công thức kiểm định LM cho SLM:

$$LM_{lag} = \frac{(e'WX\hat{\beta})^2}{\sigma^2 \sum_i \sum_j w_{ij}^2} \tag{34.2}$$

Công thức kiểm định LM cho SEM:

$$LM_{error} = \frac{(e'We)^2}{\sigma^2 tr(W'W)} \tag{34.3}$$

34.3 Ước lượng mô hình không gian

34.3.1 Phương pháp OLS

Phương pháp bình phương nhỏ nhất thông thường (OLS) sử dụng công thức:

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{34.4}$$

34.3.2 Phương pháp MLE

Ước lượng hợp lý cực đại (MLE) trong mô hình không gian có dạng:

$$L(\beta, \rho, \sigma^2) = -\frac{N}{2}\log(2\pi\sigma^2) + \log|I - \rho W| - \frac{1}{2\sigma^2}(Y - \rho WY - X\beta)'(Y - \rho WY - X\beta) \tag{34.5}$$

34.3.3 Phương pháp GMM

Ước lượng Moment tổng quát (GMM) dựa trên điều kiện moment:

$$E[g(X, \beta)] = 0 \tag{34.6}$$

Với $g(X, \beta)$ là các moment conditions thu được từ dữ liệu.

34.3.4 So sánh các phương pháp ước lượng

- OLS đơn giản nhưng không hiệu quả khi có tự tương quan không gian.
- MLE cho kết quả tốt hơn nhưng yêu cầu giả định về phân phối.
- GMM linh hoạt nhưng phụ thuộc vào cách chọn moment conditions.

Chương 35

Ứng dụng Kinh tế lượng Không gian

35.1 Phân tích giá bất động sản theo vị trí địa lý

Giá bất động sản phụ thuộc nhiều vào vị trí địa lý. Giả sử ta có dữ liệu về giá nhà theo từng quận tại TP. Hồ Chí Minh. Mô hình hồi quy không gian có thể được biểu diễn như sau:

$$P_i = \rho W P_i + X_i \beta + \epsilon_i \tag{35.1}$$

với:

- P_i là giá nhà tại khu vực i.
- W là ma trận trọng số không gian phản ánh mối quan hệ giữa các khu vực.
- X_i bao gồm các yếu tố như diện tích, số phòng, tiện ích.
- ρ là hệ số phản ánh tác động không gian.
- ϵ_i là sai số ngẫu nhiên.

Mô hình này giúp xác định mức độ ảnh hưởng của vị trí đến giá bất động sản.

35.2 Đánh giá tác động chính sách kinh tế vùng

Chính sách đầu tư cơ sở hạ tầng có thể ảnh hưởng đến tăng trưởng kinh tế vùng. Mô hình không gian phù hợp là mô hình Durbin không gian (SDM):

$$Y = \rho W Y + X\beta + WX\theta + \epsilon \tag{35.2}$$

Ứng dụng thực tế: phân tích tác động của dự án cao tốc Bắc Nam lên GDP các tỉnh miền Trung.

35.3 Ứng dụng trong nghiên cứu môi trường

Sử dụng dữ liệu ô nhiễm không khí (PM2.5) từ các trạm quan trắc, ta có thể mô hình hóa sự lan truyền của khí thải bằng mô hình sai số không gian (SEM):

$$Y = X\beta + \mu, \quad \mu = \lambda W\mu + \epsilon \tag{35.3}$$

Ứng dụng: xác định khu vực có mức độ ô nhiễm cao nhất và tìm giải pháp kiểm soát.

35.4 Dự báo mô hình kinh tế vùng với dữ liệu không gian

Dự đoán GDP theo tỉnh thành dựa trên mô hình SLM:

$$GDP_i = \rho WGDP_i + X_i\beta + \epsilon_i \tag{35.4}$$

Ứng dụng: lập kế hoạch phát triển kinh tế vùng.

Chương 36

Công cụ và Phần mềm Phân tích Không gian với Python

36.1 Giới thiệu

Python là một trong những ngôn ngữ lập trình được sử dụng rộng rãi trong phân tích dữ liệu không gian nhờ vào hệ sinh thái phong phú và dễ sử dụng. Chương này sẽ giới thiệu những thư viện quan trọng, các kỹ thuật phân tích và ứng dụng thực tế trong phân tích không gian.

36.2 Các thư viện Python cho phân tích không gian

36.2.1 GeoPandas

- Hỗ trợ làm việc với dữ liệu không gian trong môi trường pandas.
- Xử lý và thao tác dữ liệu hình học (points, lines, polygons).

36.2.2 Shapely

- Cung cấp các phép toán hình học như giao cắt, hợp nhất, chéo nhau.
- Hữu ích trong kiểm tra quan hệ giữa các đối tượng không gian.

36.2.3 PySAL (Python Spatial Analysis Library)

- Chuyên dành cho kinh tế lượng không gian.
- Hỗ trợ phân cố không gian, hồi quy không gian, v.v.

36.2.4 Rasterio

- Dùng để xử lý dữ liệu raster (GIS, ảnh vệ tinh).
- Hỗ trợ thao tác và phân tích raster hiệu quả.

36.3 Phân tích không gian với Python

36.3.1 Kiểm định Moran's I

- Kiểm tra tính tự tương không gian của một biến số.
- Sử dụng PySAL để tính Moran's I và vẽ biểu đồ.

36.3.2 Hồi quy không gian (Spatial Regression)

- Mô hình SAR (Spatial Autoregressive Model) và SEM (Spatial Error Model).
- Thực hành hồi quy không gian bằng PySAL.

36.3.3 Tạo bản đồ nóng (Hotspot Analysis)

- Xác định vùng tập trung cao/thấp của một biến kinh tế.
- Sử dụng Getis-Ord Gi* để phân tích.

36.4 Trực quan hóa dữ liệu không gian

36.4.1 Matplotlib và GeoPandas

- Vẽ bản đồ không gian với GeoPandas.
- Tuỳ chỉnh màu sắc và kích thước.

36.4.2 Folium

- Tạo bản đồ tương tác.
- Chồng lớp dữ liệu GIS lên bản đồ Leaflet.

36.5 Ứng dụng thực tế

- Phân tích thị trường bất động sản.
- Dự báo chính sách kinh tế theo khu vực.
- Mô hình dự đoán ô nhiễm môi trường.

36.6 Kết luận

Chương này đã giới thiệu những công cụ Python quan trọng trong phân tích dữ liệu không gian, từ các kỹ thuật phân tích cho đến trực quan hóa. Python giúp chúng ta khai thác hiệu quả thông tin không gian, từ đó nâng cao chất lượng quyết định trong kinh tế và quy hoạch.

Phần VIII

Machine Learning trong kinh tế lượng

Chương 37

Các phương pháp Machine Learning trong Kinh tế lượng

37.1 Giới thiệu về Machine Learning trong Kinh tế lượng

Machine Learning (ML) ngày càng được ứng dụng rộng rãi trong kinh tế lượng nhằm nâng cao độ chính xác của mô hình dự báo, kiểm định mô hình và xử lý dữ liệu lớn. Sự khác biệt chính giữa ML và các phương pháp kinh tế lượng truyền thống là cách tiếp cận dữ liệu: ML tập trung vào tính linh hoạt và tối ưu hóa dự báo, trong khi kinh tế lượng truyền thống thường dựa trên lý thuyết kinh tế và kiểm định giả thuyết.

Machine Learning (ML) trong kinh tế lượng là một lĩnh vực kết hợp giữa các phương pháp học máy và phân tích kinh tế lượng để khám phá mô hình và dự báo dữ liệu kinh tế. ML cung cấp các công cụ mạnh mẽ để xử lý dữ liệu lớn và phi tuyến tính, mở rộng khả năng phân tích so với các mô hình hồi quy truyền thống.

37.1.1 Các Khái Niệm Cơ Bản

37.1.2 Mô hình hồi quy và Machine Learning

Trong kinh tế lượng, mô hình hồi quy tuyến tính thường được sử dụng để ước lượng quan hệ giữa biến phụ thuộc Y và các biến độc lập $X_1, X_2, \ldots, X_p$ như sau:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \tag{37.1}$$

với ε là nhiễu trắng.

Machine Learning mở rộng các phương pháp này bằng cách sử dụng các mô hình phi tuyến tính và thuật toán tối ưu hóa.

Mô hình học có giám sát và không giám sát

- Học có giám sát: Bao gồm hồi quy (Regression) và phân loại (Classification).
- Học không giám sát: Bao gồm phân cụm (Clustering) và giảm chiều dữ liệu (Dimensionality Reduction).

37.1.3 Các Phương Pháp Machine Learning Phổ Biến trong Kinh tế lượng

Hồi quy Ridge và Lasso

Khi có đa cộng tuyến, ta sử dụng hồi quy Ridge và Lasso để giảm phương sai:

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{37.2}$$

$$\hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{37.3}$$

Cây quyết định và Rừng ngẫu nhiên

Cây quyết định chia không gian dữ liệu thành các vùng nhỏ bằng cách tối ưu một tiêu chí nhất định (như giảm phương sai trong hồi quy hoặc giảm entropy trong phân loại). Rừng ngẫu nhiên sử dụng nhiều cây quyết định để cải thiện tính tổng quát.

Mạng Nơ-ron nhân tạo (Artificial Neural Networks - ANN)

ANN mô phỏng cấu trúc của não người để tìm kiếm mô hình trong dữ liệu:

$$y = f(W_2 \cdot \sigma(W_1 X + b_1) + b_2) \tag{37.4}$$

trong đó: - W_1, W_2 là trọng số, - b_1, b_2 là bias, - σ là hàm kích hoạt (như ReLU, sigmoid, tanh).

Học tăng cường (Reinforcement Learning)

Học tăng cường tìm cách tối ưu hành động dựa trên phần thưởng:

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s',a') - Q(s,a)] \tag{37.5}$$

37.1.4 Ứng dụng Machine Learning trong Kinh tế lượng

- Dự báo kinh tế vĩ mô (GDP, lạm phát, thất nghiệp)
- Phân tích tài chính (giá cổ phiếu, biến động thị trường)
- Phân loại rủi ro tín dụng
- Phân tích hành vi tiêu dùng

37.1.5 Kết luận

Machine Learning cung cấp nhiều công cụ mạnh mẽ cho kinh tế lượng, giúp cải thiện độ chính xác dự báo và phát hiện các mô hình ẩn trong dữ liệu. Việc kết hợp ML với các mô hình kinh tế truyền thống có thể mở rộng khả năng phân tích và giải thích dữ liệu kinh tế.

37.2 Hồi quy tuyến tính mở rộng: Ridge, Lasso, Elastic Net

37.2.1 Giới thiệu

Hồi quy tuyến tính mở rộng bao gồm các phương pháp Ridge Regression, Lasso Regression và Elastic Net Regression, được thiết kế để xử lý vấn đề đa cộng tuyến và chọn biến trong mô hình hồi quy.

37.2.2 Hồi quy Ridge

Hồi quy Ridge mở rộng mô hình hồi quy tuyến tính bằng cách thêm một thành phần phạt L_2 vào hàm mất mát:

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2, \tag{37.6}$$

trong đó $\lambda > 0$ là tham số điều chỉnh mức độ phạt.

37.2.3 Hồi quy Lasso

Hồi quy Lasso sử dụng chuẩn L_1 để tạo ra sự co rút của các hệ số hồi quy về 0:

$$\hat{\beta}^{lasso} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{37.7}$$

37.2.4 Elastic Net Regression

Elastic Net kết hợp cả hai phương pháp trên bằng cách sử dụng cả chuẩn L_1 và L_2:

$$\hat{\beta}^{elastic} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2. \tag{37.8}$$

Elastic Net có thể giải quyết tốt vấn đề chọn biến khi số lượng biến rất lớn.

37.2.5 Kết luận

Mô hình Ridge, Lasso và Elastic Net là các phương pháp mở rộng của hồi quy tuyến tính nhằm kiểm soát đa cộng tuyến và cải thiện khả năng tổng quát hóa của mô hình.

37.3 Mô hình cây quyết định và boosting (Random Forest, XGBoost)

37.3.1 Mô hình Cây Quyết Định

Cây quyết định là một phương pháp học có giám sát được sử dụng cho các bài toán phân loại và hồi quy. Cấu trúc của cây bao gồm:
- Nút gốc (root node)
- Nút trung gian (internal nodes)
- Lá (leaf nodes)

Quá trình xây dựng cây quyết định dựa trên việc chia nhỏ dữ liệu theo các tiêu chí như Entropy hoặc Gini Index.

Entropy và Thông tin Thu được

Entropy của một tập dữ liệu S được định nghĩa là:

$$H(S) = -\sum_{i=1}^{c} p_i \log_2 p_i \tag{37.9}$$

trong đó p_i là xác suất xuất hiện của lớp thứ i.

Thông tin thu được khi phân tách theo thuộc tính A:

$$IG(S, A) = H(S) - \sum_{v \in \text{values}(A)} \frac{|S_v|}{|S|} H(S_v) \tag{37.10}$$

Chỉ số Gini

Chỉ số Gini được tính theo công thức:

$$Gini(S) = 1 - \sum_{i=1}^{c} p_i^2 \tag{37.11}$$

37.3.2 Random Forest

Random Forest là một tập hợp của nhiều cây quyết định, mỗi cây được huấn luyện trên một tập con dữ liệu khác nhau bằng phương pháp Bootstrap Aggregating (Bagging).

Dự đoán của Random Forest được tính bằng trung bình hoặc số phiếu từ các cây:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} h_t(x) \tag{37.12}$$

trong đó $h_t(x)$ là dự đoán của cây thứ t và T là tổng số cây trong rừng.

37.4 XGBoost

XGBoost (Extreme Gradient Boosting) là một phương pháp boosting sử dụng đạo hàm bậc hai để tối ưu hóa.

Hàm mất mát của XGBoost:

$$L = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{t=1}^{T} \Omega(f_t) \tag{37.13}$$

trong đó $\Omega(f_t)$ là hàm phạt độ phức tạp của cây.

Cập nhật trọng số bằng đạo hàm bậc hai:

$$g_i = \frac{\partial l(y_i, \hat{y}_i)}{\partial \hat{y}_i}, \quad h_i = \frac{\partial^2 l(y_i, \hat{y}_i)}{\partial \hat{y}_i^2} \tag{37.14}$$

Mô hình được cập nhật bằng cách tối ưu hóa hàm mục tiêu dựa trên đạo hàm này.

37.5 Machine Learning nhân quả: Double ML, Causal Inference

Machine Learning nhân quả (Causal Machine Learning) là lĩnh vực kết hợp giữa học máy và suy luận nhân quả để xác định mối quan hệ nguyên nhân - kết quả từ dữ liệu quan sát. Trong đó, Double Machine Learning (Double ML) là một phương pháp quan trọng giúp ước lượng tác động nhân quả trong mô hình có nhiều nhiễu.

37.5.1 Suy luận nhân quả (Causal Inference)

Suy luận nhân quả dựa trên mô hình **khung phản thực tế** (Potential Outcome Framework) của Rubin:

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0), \tag{37.15}$$

trong đó:

- $Y_i(1)$, $Y_i(0)$ là kết quả tiềm năng khi cá thể i nhận hoặc không nhận can thiệp D_i.
- $D_i \in \{0, 1\}$ là biến chỉ định (treatment indicator).

Tác động nhân quả trung bình (ATE - Average Treatment Effect):

$$\tau = \mathbb{E}[Y(1) - Y(0)]. \tag{37.16}$$

Do dữ liệu chỉ quan sát một trong hai trạng thái của Y_i, cần sử dụng các phương pháp ước lượng nhân quả như:**Double Machine Learning**.

37.5.2 Double Machine Learning (Double ML)

Double ML được sử dụng để khử nhiễu từ các biến gây nhiễu (confounders) trong mô hình nhân quả. Giả sử mô hình hồi quy tổng quát:

$$Y = D\theta_0 + g(X) + \epsilon, \tag{37.17}$$

$$D = m(X) + v. \tag{37.18}$$

Bước 1: Dự đoán D bằng mô hình machine learning:

$$\hat{D} = \hat{m}(X). \tag{37.19}$$

Bước 2: Dự đoán Y bằng mô hình machine learning:

$$\hat{Y} = \hat{g}(X). \tag{37.20}$$

Bước 3: Xây dựng phần dư và ước lượng bằng phương pháp hai bước (orthogonalization):

$$\tilde{Y} = Y - \hat{g}(X), \quad \tilde{D} = D - \hat{m}(X). \tag{37.21}$$

Bước 4: Hồi quy phần dư để ước lượng tác động nhân quả:

$$\hat{\theta}_0 = \mathbb{E}[\tilde{D}^T\tilde{D}]^{-1}\mathbb{E}[\tilde{D}^T\tilde{Y}]. \tag{37.22}$$

Double ML giúp giảm thiên lệch do mô hình hóa sai số trong X và đảm bảo tính vững của ước lượng.

37.5.3 Kết luận

Double ML là một kỹ thuật mạnh mẽ trong suy luận nhân quả khi kết hợp Machine Learning với phương pháp ước lượng cổ điển. Việc áp dụng Double ML trong các bài toán kinh tế lượng giúp cải thiện độ chính xác và tính tin cậy của các phân tích nhân quả.

37.6 Deep Learning trong phân tích kinh tế

Deep Learning (Học sâu) là một nhánh của Machine Learning sử dụng mạng neuron nhân tạo nhiều lớp để học các biểu diễn dữ liệu phức tạp. Trong phân tích kinh tế, Deep Learning được áp dụng để dự báo, phân loại, và tối ưu hóa ra quyết định.

37.6.1 Mạng Neuron Nhân Tạo (Artificial Neural Network - ANN)

Một mạng neuron nhân tạo cơ bản bao gồm các lớp sau:

- Lớp đầu vào (Input Layer)

- Lớp ẩn (Hidden Layers)
- Lớp đầu ra (Output Layer)

Mỗi neuron trong một lớp nhận đầu vào từ các neuron của lớp trước và tính toán một hàm kích hoạt:

$$z_i = \sum_{j=1}^{n} w_j x_j + b, \tag{37.23}$$

trong đó w_j là trọng số, x_j là đầu vào, và b là hệ số tự do (bias).

Hàm kích hoạt phổ biến:

- Hàm sigmoid: $\sigma(z) = \frac{1}{1+e^{-z}}$
- Hàm ReLU: $f(z) = \max(0, z)$
- Hàm tanh: $\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$

37.6.2 Lan Truyền Ngược (Backpropagation)

Quá trình huấn luyện mạng neuron sử dụng thuật toán lan truyền ngược để tối ưu hóa trọng số dựa trên đạo hàm của hàm mất mát. Hàm mất mát phổ biến trong hồi quy:

$$L(\hat{y}, y) = \frac{1}{2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{37.24}$$

Đạo hàm của mất mát theo trọng số được tính bằng quy tắc dây chuyền:

$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial z}{\partial w}. \tag{37.25}$$

37.6.3 Ứng Dụng trong Kinh Tế

- Dự báo kinh tế: Mạng LSTM để phân tích chuỗi thời gian.
- Phân loại rủi ro tài chính: Sử dụng mô hình DNN để đánh giá tín dụng.
- Ra quyết định tối ưu: Deep Reinforcement Learning trong tối ưu hóa danh mục đầu tư.

37.6.4 Kết Luận

Deep Learning cung cấp các công cụ mạnh mẽ để phân tích dữ liệu kinh tế phức tạp. Việc áp dụng ANN, LSTM, và Reinforcement Learning có thể cải thiện đáng kể hiệu quả dự báo và tối ưu hóa trong kinh tế học.

Chương 38

Xử lý dữ liệu lớn trong Kinh tế lượng

38.1 Tiền xử lý dữ liệu kinh tế (missing data, outliers, scaling)

Trong phân tích kinh tế lượng, dữ liệu thường có kích thước lớn và chứa nhiều vấn đề cần xử lý trước khi đưa vào mô hình. Một số bước quan trọng trong tiền xử lý dữ liệu bao gồm:

38.1.1 Xử lý dữ liệu bị thiếu (Missing Data)

Giả sử ta có một tập dữ liệu $X \in \mathbb{R}^{n \times p}$, trong đó một số giá trị bị thiếu. Một số phương pháp xử lý dữ liệu bị thiếu phổ biến:

- **Loại bỏ các hàng hoặc cột có dữ liệu bị thiếu:** Nếu số lượng dữ liệu bị thiếu nhỏ, ta có thể loại bỏ dòng hoặc cột chứa giá trị đó.
- **Điền giá trị trung bình:** Nếu biến X_j có giá trị bị thiếu, ta có thể điền bằng trung bình:

$$X_{ij} = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ij}, \quad \text{với } X_{ij} \neq \text{NA} \tag{38.1}$$

- **Điền bằng phương pháp hồi quy:** Ước lượng giá trị bị thiếu dựa trên hồi quy với các biến khác:

$$X_{ij} = \beta_0 + \sum_{k \neq j} \beta_k X_{ik} + \varepsilon_i \tag{38.2}$$

38.1.2 Xử lý ngoại lai (Outliers)

Dữ liệu ngoại lai có thể làm sai lệch kết quả phân tích. Một số phương pháp phát hiện và xử lý:

- **Sử dụng z-score:** Nếu X_j có phân phối chuẩn, điểm ngoại lai có thể xác định bằng:

$$z_i = \frac{X_{ij} - \mu_j}{\sigma_j}, \quad |z_i| > \tau \Rightarrow X_{ij} \text{ là ngoại lai} \tag{38.3}$$

- **Sử dụng phương pháp IQR (Interquartile Range):**

$$IQR = Q_3 - Q_1, \quad X_{ij} \text{ là ngoại lai nếu } X_{ij} < Q_1 - 1.5IQR \text{ hoặc } X_{ij} > Q_3 + 1.5IQR \tag{38.4}$$

38.1.3 Chuẩn hóa và Tỷ lệ hóa (Scaling)

Khi các biến có đơn vị đo khác nhau, cần chuẩn hóa dữ liệu để đảm bảo tính ổn định của mô hình:

- **Chuẩn hóa Min-Max:** Biến đổi dữ liệu về khoảng $[0, 1]$:

$$X'_{ij} = \frac{X_{ij} - \min(X_j)}{\max(X_j) - \min(X_j)} \tag{38.5}$$

- **Chuẩn hóa Z-score:** Đưa dữ liệu về phân phối chuẩn với trung bình 0 và độ lệch chuẩn 1:

$$X'_{ij} = \frac{X_{ij} - \mu_j}{\sigma_j} \tag{38.6}$$

38.2 Chọn biến và giảm chiều dữ liệu (PCA, Feature Selection)

Trong phân tích dữ liệu lớn, việc chọn biến phù hợp và giảm chiều dữ liệu giúp cải thiện hiệu suất mô hình, giảm nhiễu và tăng độ chính xác trong dự đoán. Hai phương pháp chính là **Phân tích thành phần chính (PCA)** và **Chọn biến (Feature Selection)**.

38.2.1 Phân tích thành phần chính (PCA)

PCA là một kỹ thuật giảm chiều dữ liệu dựa trên biến đổi tuyến tính để tìm ra các hướng chính của dữ liệu.

Mô hình toán học của PCA

Cho một tập dữ liệu $X \in \mathbb{R}^{n \times p}$ với n quan sát và p biến số. PCA thực hiện phép biến đổi như sau:

1. Chuẩn hóa dữ liệu: Trung bình bằng 0 và phương sai bằng 1. 2. Tính ma trận hiệp phương sai:

$$\Sigma = \frac{1}{n} X^T X \tag{38.7}$$

3. Tính các trị riêng và vector riêng của Σ:

$$\Sigma v_i = \lambda_i v_i \tag{38.8}$$

với λ_i là trị riêng, v_i là vector riêng. 4. Chọn k thành phần chính đầu tiên tương ứng với k trị riêng lớn nhất để tạo không gian mới:

$$Z = XV_k \tag{38.9}$$

với V_k là ma trận chứa k vector riêng tương ứng.

38.2.2 Chọn biến (Feature Selection)

Feature Selection giúp chọn ra tập biến quan trọng nhất để cải thiện hiệu suất mô hình mà không làm mất quá nhiều thông tin.

Các phương pháp chọn biến

1. **Phương pháp lọc (Filter Methods)**: - Sử dụng các tiêu chí thống kê như hệ số tương quan, ANOVA, hoặc test χ^2 để đánh giá tầm quan trọng của từng biến độc lập. - Chỉ số Information Gain (IG):

$$IG(Y, X) = H(Y) - H(Y|X) \tag{38.10}$$

với $H(Y)$ là entropy của biến đích Y.

2. **Phương pháp bọc (Wrapper Methods)**: - Sử dụng thuật toán học máy như Recursive Feature Elimination (RFE) để chọn biến tốt nhất.

3. **Phương pháp nhúng (Embedded Methods)**: - Sử dụng các mô hình như Lasso (L1 regularization):

$$\min_{\beta} \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{38.11}$$

giúp loại bỏ biến không quan trọng.

38.2.3 Kết luận

Cả PCA và Feature Selection đều đóng vai trò quan trọng trong xử lý dữ liệu lớn. PCA giảm chiều dựa trên biến đổi tuyến tính, trong khi Feature Selection giúp loại bỏ các biến không quan trọng để cải thiện hiệu suất mô hình.

38.3 Xử lý dữ liệu bảng (panel data) bằng ML

38.3.1 Giới thiệu về Dữ Liệu Bảng

Dữ liệu bảng (panel data) là dạng dữ liệu kết hợp giữa dữ liệu chuỗi thời gian (time series) và dữ liệu chéo (cross-sectional). Dữ liệu bảng có dạng:

$$Y_{it} = X_{it}\beta + \epsilon_{it}, \quad i = 1, \ldots, N; \quad t = 1, \ldots, T, \tag{38.12}$$

trong đó:

- Y_{it} là biến phụ thuộc của đối tượng i tại thời điểm t.
- X_{it} là vector các biến độc lập.
- β là vector hệ số hồi quy.
- ϵ_{it} là sai số.

38.3.2 Mô hình Dữ Liệu Bảng trong Machine Learning

Các phương pháp Machine Learning có thể được áp dụng để phân tích dữ liệu bảng, bao gồm:

Hồi Quy Tuyến Tính Mở Rộng

- **Ridge Regression**: Giảm phương sai bằng cách thêm penalty $\lambda \sum_{j=1}^{p} \beta_j^2$ vào hàm mất mát.
- **Lasso Regression**: Chọn biến tự động bằng penalty $\lambda \sum_{j=1}^{p} |\beta_j|$.
- **Elastic Net**: Kết hợp Ridge và Lasso với trọng số α.

Random Forest và Gradient Boosting

- Random Forest: Dùng cây quyết định để nắm bắt tính phi tuyến.
- Gradient Boosting (XGBoost, LightGBM): Tối ưu hóa dự đoán thông qua boosting.

38.3.3 Mô hình Học Sâu với Dữ Liệu Bảng

Mô hình mạng nơ-ron nhân tạo (Neural Networks) có thể được áp dụng với dữ liệu bảng. Hàm mất mát được tối ưu hóa thường là:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - \hat{Y}_{it})^2 + \lambda ||\theta||^2. \tag{38.13}$$

38.3.4 Tổng kết

Dữ liệu bảng có thể được xử lý bằng nhiều phương pháp Machine Learning khác nhau. Hồi quy mở rộng, mô hình cây, boosting, và deep learning đều có ứng dụng quan trọng trong phân tích kinh tế.

38.4 Dữ liệu thời gian thực và vấn đề xử lý dữ liệu lớn

38.4.1 Giới thiệu

Dữ liệu thời gian thực (Real-time Data) đóng vai trò quan trọng trong các hệ thống hiện đại như giao dịch tài chính, Internet of Things (IoT), và phân tích kinh tế lượng. Xử lý dữ liệu lớn (Big Data Processing) đòi hỏi các phương pháp tối ưu để đảm bảo tính hiệu quả và chính xác trong thời gian thực.

38.4.2 Mô hình toán học trong xử lý dữ liệu thời gian thực

Dòng dữ liệu (Data Stream Processing)

Dữ liệu thời gian thực có thể được mô hình hóa như một luồng dữ liệu liên tục X_t, với:

$$X_t = \{x_1, x_2, ..., x_t, ...\}, \quad x_t \in \mathbb{R}^d \tag{38.14}$$

Mỗi x_t là một vector dữ liệu đến tại thời điểm t.

Cửa sổ trượt (Sliding Window Model)

Một cách tiếp cận phổ biến là sử dụng cửa sổ trượt kích thước w, chỉ xem xét các điểm dữ liệu gần nhất:

$$W_t = \{x_{t-w+1}, ..., x_t\} \tag{38.15}$$

Điều này giúp giảm tải tính toán và lưu trữ.

Xử lý dữ liệu song song và phân tán

Để xử lý dữ liệu lớn trong thời gian thực, các mô hình phân tán như Hadoop, Spark Streaming, Flink được sử dụng. Mô hình MapReduce có thể được biểu diễn dưới dạng toán học như sau:

$$\text{Map} : f_m : \mathbb{R}^d \to \mathbb{R}^{d'} \tag{38.16}$$

$$\text{Reduce} : f_r : \mathbb{R}^{d'} \times \mathbb{R}^{d'} \to \mathbb{R}^{d'} \tag{38.17}$$

Mô hình dự báo dữ liệu thời gian thực

Dữ liệu thời gian thực thường được dự báo bằng các mô hình như ARIMA, LSTM:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma^2) \tag{38.18}$$

Mô hình LSTM sử dụng cổng nhớ để cập nhật trạng thái ẩn h_t:

$$h_t = f(W_h h_{t-1} + W_x X_t + b_h) \tag{38.19}$$

38.4.3 Kết luận

Việc xử lý dữ liệu thời gian thực và dữ liệu lớn đòi hỏi các phương pháp tính toán hiệu quả, bao gồm mô hình trượt, xử lý song song và dự báo thời gian thực.

Chương 39

Hồi quy và Dự báo với Machine Learning

39.1 So sánh hồi quy truyền thống với ML

Hồi quy truyền thống (như hồi quy tuyến tính) và Machine Learning (ML) có sự khác biệt rõ rệt về mô hình hóa dữ liệu, cách tiếp cận và mục tiêu tối ưu hóa. Dưới đây là một phân tích toán học chi tiết:

Hồi quy truyền thống

Hồi quy tuyến tính cổ điển có dạng:

$$Y = X\beta + \varepsilon, \tag{39.1}$$

trong đó:

- Y là biến phụ thuộc (đầu ra),
- X là ma trận các biến độc lập (đặc trưng),
- β là vector hệ số hồi quy,
- ε là nhiễu có phân phối chuẩn $\mathcal{N}(0, \sigma^2)$.

Hệ số β được ước lượng bằng phương pháp bình phương tối thiểu thông thường (OLS - Ordinary Least Squares):

$$\hat{\beta} = (X^T X)^{-1} X^T Y. \tag{39.2}$$

Machine Learning trong Hồi quy

Machine Learning sử dụng các mô hình phức tạp hơn như hồi quy Ridge, Lasso, và phương pháp phi tuyến như Random Forest, Neural Networks. Cách tiếp cận này có đặc điểm:

Ridge Regression: Ổn định hóa bằng thêm một thành phần phạt $\lambda\|\beta\|^2$:

$$\hat{\beta}_{ridge} = (X^T X + \lambda I)^{-1} X^T Y. \tag{39.3}$$

Lasso Regression: Sử dụng phạt ℓ_1 để chọn biến:

$$\hat{\beta}_{lasso} = \arg\min_{\beta} \left(\|Y - X\beta\|^2 + \lambda \|\beta\|_1 \right). \tag{39.4}$$

Random Forest và Gradient Boosting: Các mô hình dựa trên cây quyết định, không có dạng công thức đơn giản mà sử dụng tập hợp cây quyết định để dự báo.

Neural Networks: Mô hình phi tuyến, tối ưu hóa dựa trên lan truyền ngược và hàm mất mát như:

$$L = \sum_{i=1}^{n} (Y_i - f(X_i; \theta))^2. \tag{39.5}$$

39.1.1 So sánh Hiệu suất

Các mô hình ML thường có hiệu suất tốt hơn khi dữ liệu phi tuyến hoặc có nhiều biến tương quan, nhưng yêu cầu nhiều dữ liệu hơn và tính toán phức tạp hơn.

39.2 Mô hình XGBoost, Random Forest và hồi quy phi tuyến

39.2.1 Giới thiệu

Mô hình XGBoost, Random Forest và hồi quy phi tuyến là các phương pháp mạnh mẽ trong Machine Learning để xử lý các quan hệ phi tuyến giữa biến đầu vào và đầu ra. Chúng giúp cải thiện độ chính xác dự báo so với các mô hình hồi quy tuyến tính truyền thống.

39.2.2 Random Forest

Random Forest (RF) là một tập hợp của nhiều cây quyết định, trong đó:

- Mỗi cây được xây dựng từ một tập con ngẫu nhiên của dữ liệu.
- Kết quả cuối cùng được lấy trung bình (đối với bài toán hồi quy) hoặc theo số phiếu bầu (đối với phân loại).

Công thức dự báo của Random Forest được biểu diễn như sau:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} f_t(x), \tag{39.6}$$

trong đó T là số lượng cây trong rừng, $f_t(x)$ là hàm dự đoán của cây thứ t.

39.2.3 XGBoost

XGBoost (Extreme Gradient Boosting) là một phương pháp tăng cường (boosting) dựa trên việc xây dựng cây quyết định tuần tự nhằm giảm thiểu hàm mất mát. Hàm mất mát tổng thể trong XGBoost có dạng:

$$L = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{t=1}^{T} \Omega(f_t), \tag{39.7}$$

trong đó:

- $l(y_i, \hat{y}_i)$ là hàm mất mát giữa giá trị thực và giá trị dự báo,
- $\Omega(f_t)$ là hàm điều chuẩn để kiểm soát độ phức tạp của cây quyết định.

Mô hình mới $F(x)$ được cập nhật theo công thức:

$$F^{(t)}(x) = F^{(t-1)}(x) + \eta f_t(x), \tag{39.8}$$

trong đó η là tốc độ học.

39.2.4 Hồi quy Phi tuyến

Hồi quy phi tuyến mô tả mối quan hệ giữa biến đầu vào và đầu ra bằng một hàm phi tuyến như:

$$y = f(x, \theta) + \varepsilon, \tag{39.9}$$

trong đó $f(x, \theta)$ có thể là một hàm mũ, logarit, sigmoid hoặc một mạng nơ-ron. Một số mô hình phổ biến:

- Hồi quy đa thức: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p + \varepsilon$.
- Hồi quy logarit: $y = \beta_0 + \beta_1 \log(x) + \varepsilon$.
- Hồi quy hàm mũ: $y = \beta_0 e^{\beta_1 x} + \varepsilon$.

39.2.5 Kết luận

Random Forest, XGBoost và hồi quy phi tuyến là các công cụ mạnh mẽ để dự báo và phân tích dữ liệu có tính phi tuyến, giúp cải thiện đáng kể độ chính xác so với mô hình tuyến tính truyền thống.

39.3 Đánh giá mô hình dự báo (MAPE, RMSE, R-squared)

39.3.1 Giới thiệu

Trong kinh tế lượng và machine learning, đánh giá mô hình dự báo là một bước quan trọng để kiểm tra mức độ chính xác của mô hình. Ba chỉ số phổ biến để đánh giá mô hình dự báo bao gồm:

- Sai số phần trăm tuyệt đối trung bình (Mean Absolute Percentage Error - MAPE)
- Căn bậc hai của trung bình bình phương sai số (Root Mean Square Error - RMSE)
- Hệ số xác định (R^2 - R-squared)

39.3.2 Sai số phần trăm tuyệt đối trung bình (MAPE)

MAPE đo lường sai số tương đối trung bình của các dự báo so với giá trị thực tế:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \tag{39.10}$$

Trong đó:

- y_i là giá trị thực tế tại quan sát thứ i
- $\hat{y}_i$ là giá trị dự báo tại quan sát thứ i
- n là số quan sát

Giá trị MAPE càng nhỏ thì mô hình dự báo càng chính xác.

39.3.3 Căn bậc hai của trung bình bình phương sai số (RMSE)

RMSE đo lường sai số dự báo trung bình theo đơn vị gốc của biến phụ thuộc:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{39.11}$$

RMSE càng nhỏ thì mô hình càng chính xác, vì sai số bình phương được giảm thiểu.

39.3.4 Hệ số xác định (R^2 - R-squared)

Hệ số xác định đo lường mức độ giải thích của biến độc lập đối với biến phụ thuộc:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{39.12}$$

Trong đó $\bar{y}$ là giá trị trung bình của biến phụ thuộc. Giá trị R^2 nằm trong khoảng $[0, 1]$, với R^2 càng gần 1 thì mô hình càng phù hợp.

39.3.5 Kết luận

Các chỉ số MAPE, RMSE, và R^2 cung cấp các cách tiếp cận khác nhau để đánh giá mô hình dự báo. Tùy vào mục tiêu phân tích mà ta có thể sử dụng từng chỉ số phù hợp.

39.4 ML và các phương pháp Bayes trong dự báo kinh tế lượng

39.4.1 Giới thiệu

Machine Learning (ML) và phương pháp Bayes ngày càng được sử dụng rộng rãi trong dự báo kinh tế lượng do khả năng xử lý dữ liệu lớn và cập nhật thông tin theo cách xác suất.

39.4.2 Phương pháp Bayes trong Dự Báo

Phương pháp Bayes dựa trên định lý Bayes:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} \tag{39.13}$$

trong đó:

- $P(\theta|D)$ là xác suất hậu nghiệm của tham số θ sau khi quan sát dữ liệu D.
- $P(D|\theta)$ là hàm khả năng (likelihood) của dữ liệu.
- $P(\theta)$ là xác suất tiên nghiệm của tham số θ.
- $P(D)$ là xác suất biên của dữ liệu.

Mô hình hồi quy Bayes

Mô hình hồi quy Bayes mở rộng hồi quy tuyến tính thông qua cách tiếp cận xác suất:

$$y = X\beta + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I) \tag{39.14}$$

Trong mô hình Bayes, β được xem là một biến ngẫu nhiên với phân phối tiên nghiệm $P(\beta)$, thường được chọn là phân phối Gauss:

$$\beta \sim \mathcal{N}(\mu_0, \Sigma_0) \tag{39.15}$$

Khi có dữ liệu quan sát, phân phối hậu nghiệm của β cũng là một phân phối Gauss:

$$P(\beta|D) \propto P(D|\beta)P(\beta) \tag{39.16}$$

Ước lượng bằng phương pháp Bayes

Giá trị kỳ vọng của β trong mô hình hồi quy Bayes được tính như sau:

$$\mathbb{E}[\beta|D] = (X^T X + \Sigma_0^{-1})^{-1}(X^T y + \Sigma_0^{-1}\mu_0) \tag{39.17}$$

$$\text{Var}(\beta|D) = (X^T X + \Sigma_0^{-1})^{-1}\sigma^2 \tag{39.18}$$

39.4.3 Kết hợp Machine Learning và Bayes trong Dự Báo

Bayesian Neural Networks (BNN)

Mạng nơ-ron Bayes sử dụng phân phối tiên nghiệm trên các trọng số W:

$$P(W|D) = \frac{P(D|W)P(W)}{P(D)} \tag{39.19}$$

BNN thường được huấn luyện bằng các phương pháp xấp xỉ như Variational Inference (VI) hoặc Markov Chain Monte Carlo (MCMC).

Gaussian Processes (GP) trong dự báo kinh tế

Mô hình Gaussian Process (GP) có thể được sử dụng trong dự báo thời gian với phân phối tiên nghiệm Gauss:

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')) \tag{39.20}$$

Trong đó:

- $m(x)$ là hàm trung bình.
- $k(x, x')$ là hàm hiệp phương sai (kernel function).

39.4.4 Kết luận

Phương pháp Bayes mang lại lợi thế trong dự báo kinh tế lượng nhờ khả năng cập nhật thông tin và biểu diễn sự không chắc chắn trong mô hình. Kết hợp Machine Learning với Bayesian Inference giúp tăng độ chính xác và tính linh hoạt của các mô hình dự báo.

Chương 40

Machine Learning trong phân tích nhân quả và chính sách

40.1 Machine Learning nhân quả (Causal Forest, Double ML)

Machine Learning nhân quả kết hợp các phương pháp học máy với mô hình kinh tế lượng để ước lượng tác động nhân quả trong các chính sách kinh tế và xã hội.

40.1.1 Causal Forest

Causal Forest (Rừng Nhân Quả) là một phương pháp mở rộng từ Random Forest để ước lượng hiệu ứng nhân quả không đồng nhất giữa các nhóm. Mô hình dựa trên phương pháp cây quyết định với trọng số thích nghi.

Giả sử mô hình nhân quả:

$$Y = \tau(X)D + f(X) + \varepsilon \tag{40.1}$$

trong đó:

- Y là biến kết quả,
- D là biến can thiệp (treatment),
- $\tau(X)$ là hiệu ứng nhân quả không đồng nhất cần ước lượng,
- $f(X)$ là hàm kiểm soát,
- ε là sai số.

Causal Forest xây dựng cây quyết định để chia dữ liệu dựa trên biến X, sau đó tính hiệu ứng nhân quả trong từng nhóm.

40.1.2 Double Machine Learning (Double ML)

Double ML là một phương pháp dựa trên việc kết hợp các mô hình hồi quy học máy với ước lượng hiệu ứng nhân quả. Phương pháp này bao gồm hai bước chính:

1. Hồi quy Y trên X và hồi quy D trên X bằng Machine Learning để loại bỏ yếu tố gây nhiễu.

2. Ước lượng hiệu ứng nhân quả bằng phương pháp học máy tuyến tính hoặc phi tuyến tính.

Mô hình Double ML có dạng:

$$Y = g(X) + \tau D + U \tag{40.2}$$

$$D = m(X) + V \tag{40.3}$$

Trong đó U, V là phần dư.

Ước lượng hiệu ứng nhân quả bằng cách sử dụng sai số có điều chỉnh:

$$\hat{\tau} = \frac{\sum_{i=1}^{n}(\tilde{Y}_i - \bar{\tilde{Y}})(\tilde{D}_i - \bar{\tilde{D}})}{\sum_{i=1}^{n}(\tilde{D}_i - \bar{\tilde{D}})^2} \tag{40.4}$$

trong đó $\tilde{Y}_i$ và $\tilde{D}_i$ là phần dư sau khi loại bỏ ảnh hưởng của X.

40.2 Xác định tác động chính sách bằng ML

40.2.1 Giới thiệu

Trong kinh tế lượng, việc đánh giá tác động của các chính sách là một vấn đề quan trọng. Truyền thống, các phương pháp như hồi quy tuyến tính, mô hình sai biệt-kép (Difference-in-Differences, DID) và phương pháp sử dụng biến công cụ (Instrumental Variables, IV) được áp dụng. Tuy nhiên, Machine Learning (ML) cung cấp các công cụ mạnh mẽ hơn để xác định tác động chính sách, đặc biệt trong các mô hình phi tuyến tính và dữ liệu lớn.

40.2.2 Mô hình tổng quát

Giả sử ta có một tập dữ liệu $\{(X_i, D_i, Y_i)\}_{i=1}^{N}$, trong đó:

- X_i là vector đặc trưng (covariates) của cá nhân/thực thể i.
- $D_i \in \{0, 1\}$ là biến chỉ thị chính sách (được nhận hay không).
- Y_i là kết quả đầu ra (outcome) mà ta muốn đo lường.

Mô hình tác động chính sách được biểu diễn như sau:

$$Y_i = \tau(X_i)D_i + f(X_i) + \epsilon_i, \tag{40.5}$$

trong đó $\tau(X_i)$ là tác động chính sách điều kiện theo đặc trưng X_i, $f(X_i)$ là thành phần kiểm soát và ϵ_i là nhiễu.

40.2.3 Phương pháp Machine Learning trong đánh giá chính sách

Causal Forest

Causal Forest (Athey và Imbens, 2016) là một mô hình dựa trên Random Forest để ước lượng tác động chính sách dị biệt theo từng cá nhân:

$$\hat{\tau}(X) = \mathbb{E}[Y|X, D = 1] - \mathbb{E}[Y|X, D = 0]. \tag{40.6}$$

Mô hình sử dụng cây quyết định để chia nhóm đồng nhất về đặc trưng X, sau đó ước lượng tác động trung bình có điều kiện.

Double Machine Learning (Double ML)

Phương pháp Double ML (Chernozhukov et al., 2018) sử dụng hồi quy Lasso hoặc Random Forest để điều chỉnh nhiễu trước khi ước lượng tác động chính sách:

$$\tilde{Y}_i = Y_i - \hat{f}(X_i), \tag{40.7}$$

$$\tilde{D}_i = D_i - \hat{g}(X_i), \tag{40.8}$$

$$\tau = \frac{\sum_{i=1}^{N} \tilde{Y}_i \tilde{D}_i}{\sum_{i=1}^{N} \tilde{D}_i^2}. \tag{40.9}$$

Phương pháp này giúp giảm sai lệch do nhiễu và lựa chọn biến tự động.

40.2.4 Kết luận

Machine Learning mang lại những công cụ mạnh mẽ để xác định tác động chính sách, đặc biệt khi dữ liệu có nhiều biến số và quan hệ phi tuyến tính. Causal Forest và Double ML là hai phương pháp phổ biến giúp nâng cao độ chính xác của ước lượng.

40.3 Kiểm định giả thuyết và ML trong phân tích chính sách

40.3.1 Giới thiệu

Kiểm định giả thuyết là một công cụ quan trọng trong phân tích chính sách. Khi kết hợp với Machine Learning (ML), phương pháp này có thể giúp phát hiện mối quan hệ nhân quả, đánh giá tác động của chính sách, và cải thiện độ chính xác của mô hình dự báo.

40.3.2 Kiểm định giả thuyết thống kê

Giả sử chúng ta có hai giả thuyết:

- H_0: Không có tác động của chính sách.
- H_1: Chính sách có tác động đáng kể.

Với dữ liệu $X = \{x_1, x_2, ..., x_n\}$ và biến phụ thuộc $Y = \{y_1, y_2, ..., y_n\}$, kiểm định giả thuyết truyền thống có thể được thực hiện bằng kiểm định t hoặc kiểm định F trong mô hình hồi quy:

$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2). \tag{40.10}$$

Thống kê kiểm định t được tính bằng:

$$t = \frac{\hat{\beta}_1}{\text{SE}(\hat{\beta}_1)}, \tag{40.11}$$

trong đó $\text{SE}(\hat{\beta}_1)$ là sai số chuẩn của $\hat{\beta}_1$.

40.3.3 Machine Learning và kiểm định giả thuyết

Machine Learning có thể được sử dụng để kiểm định giả thuyết theo nhiều cách khác nhau, bao gồm:

- Phương pháp permutation test: Xáo trộn dữ liệu và so sánh phân phối của các hệ số.
- Phương pháp bootstrap: Lấy mẫu lại nhiều lần để ước lượng phân phối tham số.
- Double Machine Learning (DML): Sử dụng ML để kiểm soát biến gây nhiễu trong mô hình nhân quả.

Permutation Test

Xáo trộn ngẫu nhiên biến độc lập X và ước lượng mô hình nhiều lần để xây dựng phân phối giả định của hệ số hồi quy:

$$P = \frac{|\hat{\beta}_1 - \hat{\beta}_1^{perm}|}{\sigma_{perm}}. \tag{40.12}$$

Bootstrap

Với phương pháp bootstrap, ta tạo nhiều tập dữ liệu mới bằng cách lấy mẫu lại với thay thế, sau đó ước lượng lại mô hình trên từng tập dữ liệu:

$$\hat{\beta}_1^{(b)} = \frac{\sum_{i=1}^{n} w_i X_i Y_i}{\sum_{i=1}^{n} w_i X_i^2}, \tag{40.13}$$

trong đó w_i là trọng số bootstrap.

Double Machine Learning

DML tách quy trình ước lượng thành hai bước:

1. Dự báo biến độc lập bằng mô hình ML: $\hat{X} = f(Z) + \eta$.
2. Hồi quy phần dư để loại bỏ biến nhiễu: $\hat{Y} = \beta_1 \hat{X} + \varepsilon$.

40.3.4 Ứng dụng thực tế

Một ví dụ thực tế là đánh giá tác động của chính sách trợ cấp kinh tế. Ta có thể sử dụng mô hình hồi quy tuyến tính thông thường và so sánh kết quả với các phương pháp ML như Random Forest hoặc XGBoost để kiểm tra tính vững chắc của kết quả.

40.3.5 Kết luận

Kiểm định giả thuyết kết hợp với ML giúp phân tích chính sách trở nên mạnh mẽ hơn, đặc biệt trong các tình huống có nhiều biến gây nhiễu. Các phương pháp như permutation test, bootstrap và DML là những công cụ hữu ích trong lĩnh vực này.

Chương 41

Machine Learning trong phân tích dữ liệu bảng (Panel Data)

41.1 Xử lý dữ liệu bảng lớn với ML

41.1.1 Giới thiệu

Dữ liệu bảng (panel data) là loại dữ liệu kết hợp giữa chuỗi thời gian và dữ liệu chéo. Xử lý dữ liệu bảng lớn với Machine Learning (ML) đòi hỏi các phương pháp tối ưu hóa, mô hình phi tuyến và các thuật toán có khả năng khai thác mối quan hệ phức tạp trong dữ liệu.

41.1.2 Mô hình hóa dữ liệu bảng

Một mô hình dữ liệu bảng tổng quát có dạng:

$$y_{it} = \alpha + X_{it}\beta + u_{it}, \tag{41.1}$$

với:

- y_{it} là biến phụ thuộc của cá thể i tại thời điểm t.
- X_{it} là vector các biến độc lập.
- β là vector các hệ số hồi quy.
- u_{it} là sai số.

Trong Machine Learning, ta có thể sử dụng các mô hình phi tuyến như Random Forest, Gradient Boosting hoặc Neural Networks để ước lượng hàm hồi quy:

$$y_{it} = f(X_{it}) + \epsilon_{it}, \tag{41.2}$$

nơi $f(X_{it})$ là một hàm phi tuyến được học bởi mô hình ML.

41.1.3 Các phương pháp Machine Learning trong dữ liệu bảng

Random Forest và Gradient Boosting

Random Forest và Gradient Boosting Trees (GBT) là các thuật toán ML phổ biến để xử lý dữ liệu bảng lớn. Chúng có thể bắt được các quan hệ phi tuyến và tương tác giữa các biến.

Mô hình mạng nơ-ron nhân tạo (Neural Networks)

Mạng nơ-ron có thể mở rộng để xử lý dữ liệu bảng bằng cách sử dụng kiến trúc như Long Short-Term Memory (LSTM) hoặc Transformer để nắm bắt quan hệ theo thời gian giữa các quan sát.

41.1.4 Ước lượng tham số và đánh giá mô hình

Các phương pháp ước lượng tham số phổ biến bao gồm Regularized Regression (Ridge, Lasso), Bayesian Inference và Maximum Likelihood Estimation (MLE). Các tiêu chí đánh giá mô hình gồm:

- Mean Squared Error (MSE)
- Mean Absolute Error (MAE)
- R-squared (R^2)
- Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC)

41.1.5 Kết luận

Machine Learning mở ra nhiều hướng mới trong phân tích dữ liệu bảng lớn, giúp nâng cao khả năng dự báo và phân tích chính sách kinh tế lượng.

41.2 So sánh Fixed Effects, Random Effects với ML

41.2.1 Giới thiệu

Trong kinh tế lượng dữ liệu bảng (panel data), hai phương pháp phổ biến để kiểm soát các hiệu ứng không quan sát được là mô hình hiệu ứng cố định (Fixed Effects - FE) và mô hình hiệu ứng ngẫu nhiên (Random Effects - RE). Cùng với sự phát triển của Machine Learning (ML), các phương pháp học máy cũng được áp dụng để xử lý dữ liệu bảng.

41.2.2 Mô hình Hiệu ứng Cố định (Fixed Effects Model)

Mô hình hiệu ứng cố định có dạng tổng quát:

$$y_{it} = \beta X_{it} + \alpha_i + \varepsilon_{it}, \tag{41.3}$$

trong đó:

- y_{it} là biến phụ thuộc của cá nhân i tại thời điểm t,
- X_{it} là vector các biến độc lập,
- α_i là hiệu ứng cố định của cá nhân i,
- ε_{it} là sai số.

Mô hình FE kiểm soát sự không đồng nhất giữa các cá nhân bằng cách đưa vào các hiệu ứng cố định α_i, ước lượng thường được thực hiện bằng phương pháp bình phương bé nhất trong nhóm (Within estimator).

41.2.3 Mô hình Hiệu ứng Ngẫu nhiên (Random Effects Model)

Mô hình hiệu ứng ngẫu nhiên có dạng:

$$y_{it} = \beta X_{it} + u_i + \varepsilon_{it}, \tag{41.4}$$

trong đó $u_i \sim N(0, \sigma_u^2)$ là thành phần ngẫu nhiên đại diện cho hiệu ứng cá nhân. Mô hình RE giả định rằng u_i không tương quan với X_{it}, do đó có thể ước lượng bằng bình phương bé nhất tổng quát (GLS).

41.2.4 Ứng dụng Machine Learning trong Dữ liệu Bảng

Machine Learning (ML) có thể thay thế hoặc kết hợp với các mô hình FE/RE để dự báo và phân tích dữ liệu bảng. Một số phương pháp ML phổ biến:

- Cây quyết định (Decision Trees), Rừng ngẫu nhiên (Random Forest), XGBoost: Có thể phát hiện mối quan hệ phi tuyến giữa các biến.
- Hồi quy Ridge, Lasso: Kiểm soát đa cộng tuyến tốt hơn FE/RE.
- Mạng nơ-ron nhân tạo (Neural Networks): Có thể xử lý dữ liệu bảng phức tạp.

Một cách tiếp cận kết hợp là sử dụng Fixed Effects trong bước tiền xử lý để loại bỏ các yếu tố không quan sát được, sau đó áp dụng ML để cải thiện dự báo.

41.2.5 Kết luận

Mô hình hiệu ứng cố định và hiệu ứng ngẫu nhiên có lợi thế khi dữ liệu có cấu trúc bảng truyền thống, nhưng ML mang lại khả năng phát hiện mô hình phức tạp hơn. Việc kết hợp cả hai phương pháp có thể giúp nâng cao độ chính xác trong dự báo kinh tế lượng.

41.3 Ứng dụng ML vào phân tích tác động theo thời gian

41.3.1 Giới thiệu

Phân tích tác động theo thời gian là một phương pháp quan trọng trong kinh tế lượng và khoa học dữ liệu nhằm xác định sự thay đổi của một biến phụ thuộc theo thời gian do tác động của một yếu tố nhất định. Machine Learning (ML) có thể hỗ trợ trong việc phân tích dữ liệu thời gian thực, cải thiện độ chính xác và phát hiện các mẫu phức tạp.

41.3.2 Mô hình hóa tác động theo thời gian

Giả sử chúng ta có một tập dữ liệu chuỗi thời gian $\{y_t, X_t\}_{t=1}^{T}$, trong đó y_t là biến phụ thuộc và X_t là tập hợp các biến độc lập.

Mô hình truyền thống

Một cách tiếp cận phổ biến trong kinh tế lượng là sử dụng mô hình hồi quy:

$$y_t = \beta_0 + \beta_1 X_t + \epsilon_t, \tag{41.5}$$

trong đó $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ là nhiễu trắng.

Tuy nhiên, mô hình này không thể nắm bắt được các yếu tố phi tuyến hoặc sự phụ thuộc phức tạp theo thời gian.

Ứng dụng Machine Learning

Machine Learning có thể cải thiện phân tích bằng cách sử dụng các mô hình như:

- **Random Forest**: Một tập hợp các cây quyết định có thể nắm bắt mối quan hệ phi tuyến.
- **XGBoost**: Một phương pháp boosting giúp tối ưu hóa dự báo chuỗi thời gian.
- **Recurrent Neural Networks (RNNs)**: Mạng nơ-ron hồi quy giúp phát hiện xu hướng và sự phụ thuộc theo thời gian.

Ví dụ, sử dụng Random Forest để dự báo tác động theo thời gian:

$$y_t = f(X_t) + \epsilon_t, \tag{41.6}$$

trong đó $f(\cdot)$ là một hàm phi tuyến ước lượng bởi mô hình ML.

41.3.3 Ước lượng và đánh giá mô hình

Để đánh giá mô hình, ta có thể sử dụng các chỉ số:

- **MSE (Mean Squared Error)**:

$$\text{MSE} = \frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2. \tag{41.7}$$

- **MAPE (Mean Absolute Percentage Error)**:

$$\text{MAPE} = \frac{100\%}{T}\sum_{t=1}^{T}\left|\frac{y_t - \hat{y}_t}{y_t}\right|. \tag{41.8}$$

- **R-squared**:

$$R^2 = 1 - \frac{\sum_{t=1}^{T}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{T}(y_t - \bar{y})^2}. \tag{41.9}$$

41.3.4 Kết luận

Ứng dụng Machine Learning vào phân tích tác động theo thời gian giúp phát hiện các mối quan hệ phức tạp, cải thiện độ chính xác và dự báo tốt hơn so với các phương pháp truyền thống. Tuy nhiên, việc lựa chọn mô hình phù hợp và kiểm định tính ổn định của mô hình vẫn là một thách thức quan trọng.

Chương 42

Machine Learning trong phân tích dữ liệu chuỗi thời gian

42.1 Mô hình hóa dữ liệu chuỗi thời gian bằng ML

Dữ liệu chuỗi thời gian là một tập hợp các quan sát được thu thập theo thứ tự thời gian. Cho một chuỗi thời gian $\{y_t\}_{t=1}^{T}$, mục tiêu chính là dự báo giá trị tương lai y_{T+h} dựa trên các quan sát trong quá khứ.

42.1.1 Định nghĩa chuỗi thời gian

Một chuỗi thời gian $\{y_t\}$ được coi là một hàm số theo thời gian:

$$y_t = f(t) + \epsilon_t, \tag{42.1}$$

trong đó $f(t)$ là thành phần có hệ thống và ϵ_t là nhiễu ngẫu nhiên.

42.1.2 Mô hình hồi quy tuyến tính

Một trong những phương pháp đơn giản để phân tích chuỗi thời gian là mô hình hồi quy tuyến tính:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 x_t + \epsilon_t, \tag{42.2}$$

trong đó:
- y_t là giá trị cần dự báo,
- y_{t-1} là giá trị trễ của chuỗi,
- x_t là biến giải thích,
- $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ là nhiễu.

42.1.3 Mô hình ARIMA

Mô hình tự hồi quy tích hợp trung bình trượt (ARIMA) có dạng:

$$\phi(B)(1 - B)^d y_t = \theta(B)\epsilon_t, \tag{42.3}$$

trong đó:
- B là toán tử trễ ($By_t = y_{t-1}$),
- $\phi(B)$ là đa thức tự hồi quy bậc p,
- $\theta(B)$ là đa thức trung bình trượt bậc q,
- d là số lần lấy sai phân.

42.1.4 Mô hình dựa trên Machine Learning

Các phương pháp Machine Learning có thể được sử dụng để dự báo chuỗi thời gian, bao gồm:

- Hồi quy Ridge, Lasso
- Mạng nơ-ron nhân tạo (ANN)
- Mạng nơ-ron hồi quy (RNN, LSTM, GRU)

Một mô hình Perceptron đa tầng (MLP - Multi-Layer Perceptron) có thể được biểu diễn như sau:

$$\hat{y}_t = \sigma\left(W_2 \cdot \sigma(W_1 x_t + b_1) + b_2\right), \tag{42.4}$$

trong đó:

- $x_t \in \mathbb{R}^n$ là vector đầu vào tại thời điểm t,
- $W_1 \in \mathbb{R}^{m\times n}$ và $W_2 \in \mathbb{R}^{1\times m}$ là ma trận trọng số,
- $b_1 \in \mathbb{R}^m$ và $b_2 \in \mathbb{R}$ là các hệ số điều chỉnh (bias),
- $\sigma(\cdot)$ là hàm kích hoạt phi tuyến (ví dụ: ReLU, sigmoid, tanh),
- $\hat{y}_t$ là giá trị dự báo tại thời điểm t.

Các hàm kích hoạt phổ biến trong MLP bao gồm:

- Hàm sigmoid:
$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{42.5}$$
- Hàm ReLU:
$$\sigma(x) = \max(0, x) \tag{42.6}$$
- Hàm tanh:
$$\sigma(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{42.7}$$

Mô hình MLP học tham số W_1, W_2, b_1, b_2 thông qua tối ưu hóa hàm mất mát, chẳng hạn như hàm lỗi bình phương trung bình (MSE):

$$L = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2. \tag{42.8}$$

42.2 So sánh ARIMA, VAR với ML (XGBoost, LSTM)

42.2.1 Mô hình ARIMA

ARIMA (Autoregressive Integrated Moving Average) là một mô hình thống kê truyền thống được sử dụng để phân tích chuỗi thời gian. Mô hình có dạng tổng quát:

$$\phi(B)(1-B)^d y_t = \theta(B)\epsilon_t, \tag{42.9}$$

trong đó:
- $\phi(B)$ là đa thức tự hồi quy (AR) bậc p,
- $\theta(B)$ là đa thức trung bình trượt (MA) bậc q,
- d là số lần lấy sai phân để làm dừng chuỗi.

42.2.2 Mô hình VAR (Vector Autoregression)

VAR mở rộng mô hình AR để bao gồm nhiều biến:

$$Y_t = A_1 Y_{t-1} + A_2 Y_{t-2} + \cdots + A_p Y_{t-p} + \epsilon_t, \tag{42.10}$$

trong đó:
- Y_t là vector các biến,
- A_i là ma trận hệ số,
- ϵ_t là vector nhiễu trắng.

42.2.3 XGBoost trong dự báo chuỗi thời gian

XGBoost (Extreme Gradient Boosting) là một thuật toán cây quyết định tăng cường có thể được sử dụng để dự báo chuỗi thời gian bằng cách biến chuỗi thời gian thành một bài toán hồi quy:

$$y_t = f(x_t) + \epsilon_t, \tag{42.11}$$

trong đó:
- x_t là vector đặc trưng của chuỗi thời gian tại thời điểm t,
- $f(x_t)$ là một tổ hợp tuyến tính của nhiều cây quyết định.

Hàm mất mát của XGBoost thường sử dụng dạng:

$$L = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_j \Omega(f_j), \tag{42.12}$$

trong đó $\Omega(f_j)$ là độ phức tạp của cây quyết định.

42.2.4 LSTM trong dự báo chuỗi thời gian

LSTM (Long Short-Term Memory) là một dạng mạng nơ-ron hồi quy (RNN) chuyên dùng để dự báo chuỗi thời gian. Trạng thái ẩn h_t của LSTM được tính

như sau:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f), \tag{42.13}$$
$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i), \tag{42.14}$$
$$\tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c), \tag{42.15}$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t, \tag{42.16}$$
$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o), \tag{42.17}$$
$$h_t = o_t \odot \tanh(c_t), \tag{42.18}$$

trong đó:
- f_t, i_t, o_t lần lượt là cổng quên, cổng nhập và cổng đầu ra,
- c_t là bộ nhớ,
- h_t là trạng thái ẩn,
- W, U, b là các tham số học.

42.3 Phát hiện xu hướng và cú sốc kinh tế bằng ML

Phát hiện xu hướng và cú sốc kinh tế là một vấn đề quan trọng trong kinh tế lượng tài chính. Các phương pháp Machine Learning có thể hỗ trợ phân tích các đặc điểm của chuỗi thời gian kinh tế bằng cách phát hiện các biến động bất thường và xu hướng dài hạn.

42.3.1 Phát hiện xu hướng bằng hồi quy tuyến tính

Xu hướng của một chuỗi thời gian y_t có thể được mô hình hóa bằng phương trình hồi quy tuyến tính đơn giản:

$$y_t = \alpha + \beta t + \epsilon_t, \tag{42.19}$$

trong đó:
- α là hằng số,
- β là hệ số xu hướng,
- t là thời gian,
- ϵ_t là sai số ngẫu nhiên.

Nếu $\beta > 0$, chuỗi có xu hướng tăng, nếu $\beta < 0$, chuỗi có xu hướng giảm.

42.3.2 Phát hiện cú sốc kinh tế bằng kiểm định thay đổi cấu trúc

Một cú sốc kinh tế có thể gây ra thay đổi đột ngột trong cấu trúc dữ liệu chuỗi thời gian. Kiểm định Chow có thể được sử dụng để phát hiện sự thay đổi cấu trúc tại thời điểm T_c:

$$F = \frac{(RSS_r - RSS_u)/k}{RSS_u/(n-2k)}, \tag{42.20}$$

trong đó:
- RSS_r là tổng bình phương dư của mô hình gộp (không có thay đổi cấu trúc),
- RSS_u là tổng bình phương dư của mô hình có thay đổi cấu trúc,
- k là số tham số trong mô hình.

Nếu giá trị F lớn hơn giá trị tới hạn, ta bác bỏ giả thuyết không và kết luận rằng có sự thay đổi cấu trúc.

42.3.3 Phát hiện xu hướng phi tuyến bằng Machine Learning

Các mô hình phi tuyến như LSTM hoặc XGBoost có thể học các xu hướng ẩn trong dữ liệu:

- Mạng LSTM có thể phát hiện xu hướng dài hạn và các mẫu phức tạp bằng cách sử dụng cơ chế bộ nhớ dài hạn.
- XGBoost có thể phát hiện các yếu tố quan trọng ảnh hưởng đến xu hướng và cú sốc kinh tế.

Mô hình LSTM có thể được biểu diễn như sau:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f), \tag{42.21}$$
$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i), \tag{42.22}$$
$$\tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c), \tag{42.23}$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t, \tag{42.24}$$
$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o), \tag{42.25}$$
$$h_t = o_t \odot \tanh(c_t), \tag{42.26}$$

trong đó $\odot$ là phép nhân từng phần tử, và các biến f_t, i_t, o_t, c_t lần lượt là các cổng quên, đầu vào, đầu ra và trạng thái bộ nhớ.

Phần IX

Phương pháp Monte Carlo và mô phỏng

Chương 43

Giới thiệu phương pháp Monte Carlo

43.1 Tổng quan về phương pháp Monte Carlo

Phương pháp Monte Carlo là một kỹ thuật tính toán dựa trên mô phỏng ngẫu nhiên để tìm nghiệm số của các bài toán phức tạp. Trong kinh tế lượng, Monte Carlo được sử dụng để:

- Kiểm định tính chính xác của các ước lượng thống kê.
- Phân tích tính chất phân phối của các thống kê kiểm định.
- Đánh giá hiệu suất của các mô hình kinh tế lượng.

Cho một mô hình kinh tế lượng:

$$Y = f(X, \theta) + \varepsilon \tag{43.1}$$

trong đó:

- Y là biến phụ thuộc,
- X là ma trận biến độc lập,
- θ là vector tham số cần ước lượng,
- $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ là nhiễu ngẫu nhiên.

Monte Carlo có thể được sử dụng để phân tích các tính chất của ước lượng $\hat{\theta}$.

43.2 Mô phỏng Monte Carlo trong kiểm định giả thuyết kinh tế lượng

43.2.1 Mô tả quy trình Monte Carlo

Để đánh giá một ước lượng $\hat{\theta}$, ta sử dụng quy trình mô phỏng Monte Carlo sau:

1. Chọn một giá trị thật của tham số θ_0 và xác định mô hình kinh tế lượng.
2. Sinh ngẫu nhiên tập dữ liệu giả lập (X, Y) theo mô hình:

$$Y_i = X_i\theta_0 + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{43.2}$$

3. Ước lượng tham số $\hat{\theta}$ bằng phương pháp ước lượng như OLS:

$$\hat{\theta} = (X'X)^{-1}X'Y \tag{43.3}$$

4. Lặp lại bước 2 và 3 với nhiều bộ dữ liệu giả lập (X, Y) để có các ước lượng $\hat{\theta}^{(1)}, \hat{\theta}^{(2)}, \ldots, \hat{\theta}^{(M)}$.
5. Phân tích phân phối của $\hat{\theta}$ để kiểm tra tính chệch, phương sai và hiệu suất của phương pháp ước lượng.

43.2.2 Tính chệch và phương sai của ước lượng

Từ các lần lặp Monte Carlo, ta có thể tính kỳ vọng và phương sai của ước lượng:

$$E[\hat{\theta}] \approx \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}^{(m)} \tag{43.4}$$

$$\mathrm{Var}(\hat{\theta}) \approx \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\theta}^{(m)} - E[\hat{\theta}])^2 \tag{43.5}$$

- Nếu $E[\hat{\theta}] = \theta_0$, phương pháp ước lượng là không chệch.
- Nếu phương sai nhỏ, phương pháp có độ chính xác cao.

43.2.3 Ứng dụng trong kiểm định giả thuyết

Trong kiểm định giả thuyết, Monte Carlo được dùng để kiểm tra phân phối của các thống kê kiểm định. Ví dụ, với kiểm định giả thuyết:

$$H_0 : \theta = \theta_0 \tag{43.6}$$

ta có thể sử dụng thống kê t:

$$t = \frac{\hat{\theta} - \theta_0}{s_{\hat{\theta}}} \tag{43.7}$$

- $s_{\hat{\theta}}$ là độ lệch chuẩn của $\hat{\theta}$.
- Nếu Monte Carlo cho thấy phân phối của t không tuân theo phân phối chuẩn t, thì kiểm định truyền thống có thể không đáng tin cậy.

Chương 44

Ứng dụng mô phỏng trong kinh tế

44.1 Mô phỏng dữ liệu kinh tế lượng

44.1.1 Mô hình tổng quát

Chúng ta xét một mô hình hồi quy tuyến tính bội như sau:

$$GDP_t = \beta_0 + \beta_1 Investment_t + \beta_2 Consumption_t + \varepsilon_t \tag{44.1}$$

với:

- GDP_t là tổng sản phẩm quốc nội tại thời điểm t,
- $Investment_t$ là đầu tư tại thời điểm t,
- $Consumption_t$ là tiêu dùng tại thời điểm t,
- $\varepsilon_t \sim N(0, \sigma^2)$ là nhiễu ngẫu nhiên có phân phối chuẩn với trung bình bằng 0 và phương sai σ^2,
- $\beta_0, \beta_1, \beta_2$ là các tham số cần ước lượng.

44.1.2 Mô phỏng dữ liệu bằng phương pháp Monte Carlo

Bước 1: Xác định số lượng quan sát
Chọn số lượng quan sát T, ví dụ $T = 1000$, để mô phỏng dữ liệu.

Bước 2: Xác định giá trị thực của các tham số
Chọn giá trị thực của các tham số:

$$\beta_0 = 5, \quad \beta_1 = 2, \quad \beta_2 = 3 \tag{44.2}$$

Bước 3: Sinh dữ liệu giả lập cho biến độc lập
Sinh các biến $Investment_t$ và $Consumption_t$ từ phân phối chuẩn hoặc phân phối khác phù hợp với thực tế. Ví dụ:

$$Investment_t \sim N(10, 5^2), \quad Consumption_t \sim N(20, 10^2) \tag{44.3}$$

Bước 4: Sinh nhiễu ngẫu nhiên ε_t
Sinh ε_t từ phân phối chuẩn:

$$\varepsilon_t \sim N(0, 2^2) \tag{44.4}$$

Bước 5: Tính giá trị của GDP_t theo phương trình hồi quy

$$GDP_t = 5 + 2Investment_t + 3Consumption_t + \varepsilon_t \tag{44.5}$$

Bước 6: Ước lượng tham số hồi quy
Sử dụng phương pháp bình phương nhỏ nhất (OLS - Ordinary Least Squares) để ước lượng các tham số β. Hệ số ước lượng được tính theo công thức:

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{44.6}$$

trong đó:

- X là ma trận thiết kế:

$$X = \begin{bmatrix} 1 & Investment_1 & Consumption_1 \\ 1 & Investment_2 & Consumption_2 \\ \vdots & \vdots & \vdots \\ 1 & Investment_T & Consumption_T \end{bmatrix} \tag{44.7}$$

- Y là vector kết quả:

$$Y = \begin{bmatrix} GDP_1 \\ GDP_2 \\ \vdots \\ GDP_T \end{bmatrix} \tag{44.8}$$

- $\hat{\beta}$ là vector các ước lượng:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} \tag{44.9}$$

Bước 7: Lặp lại quy trình
Lặp lại các bước trên N lần (ví dụ $N = 1000$) để tạo ra nhiều bộ dữ liệu khác nhau và phân tích sự ổn định của các ước lượng $\hat{\beta}$.

44.1.3 Đánh giá tính ổn định của ước lượng

Sau khi thực hiện mô phỏng Monte Carlo, ta có thể phân tích sự ổn định của các ước lượng bằng cách:

- Tính **trung bình của các ước lượng** $\hat{\beta}$ trên toàn bộ các lần lặp. Nếu $E[\hat{\beta}] \approx \beta$, thì ước lượng không chệch.

- Tính **phương sai của các ước lượng** để đánh giá độ tin cậy của mô hình. Nếu phương sai nhỏ, mô hình có độ tin cậy cao.
- Vẽ **biểu đồ phân phối của** $\hat{\beta}$ để kiểm tra tính chuẩn của ước lượng.

Tóm lại:

- Mô phỏng dữ liệu kinh tế lượng giúp kiểm tra tính ổn định của ước lượng bằng phương pháp Monte Carlo.
- Ta có thể sinh ngẫu nhiên các biến đầu vào, sau đó sử dụng OLS để ước lượng tham số.
- Việc lặp lại nhiều lần giúp đánh giá tính ổn định của các ước lượng β.

Dưới đây là mã Python để thực hiện mô phỏng Monte Carlo theo mô hình kinh tế lượng đã đề cập:

```
import numpy as np
import statsmodels.api as sm
import matplotlib.pyplot as plt

def monte_carlo_simulation(n_simulations=1000, n_samples=100):
    beta_0, beta_1, beta_2 = 2, 0.5, 0.3
    beta_estimates = []

    for _ in range(n_simulations):
        investment = np.random.normal(50, 10, n_samples)
        consumption = np.random.normal(30, 5, n_samples)
        epsilon = np.random.normal(0, 5, n_samples)

        gdp = beta_0 + beta_1 * investment + beta_2 * consumption + epsilon

        X = np.column_stack((np.ones(n_samples), investment, consumption))
        y = gdp

        model = sm.OLS(y, X).fit()
        beta_estimates.append(model.params)

    beta_estimates = np.array(beta_estimates)

    plt.figure(figsize=(10, 5))
```

```
25      plt.hist(beta_estimates[:, 1], bins=30, alpha=0.6,
    label="Beta 1")
26      plt.hist(beta_estimates[:, 2], bins=30, alpha=0.6,
    label="Beta 2")
27      plt.axvline(beta_1, color='r', linestyle='dashed',
    linewidth=2, label='True Beta 1')
28      plt.axvline(beta_2, color='g', linestyle='dashed',
    linewidth=2, label='True Beta 2")
29      plt.xlabel("Gia tri uoc luong")
30      plt.ylabel("Tan suat")
31      plt.title("Phan phoi cua uoc luong Beta")
32      plt.legend()
33      plt.show()
34
35      return beta_estimates
36
37  beta_estimates = monte_carlo_simulation()
38
```

Mã Python trên thực hiện mô phỏng Monte Carlo với các tham số đã nêu, sử dụng hồi quy OLS để ước lượng hệ số và trực quan hóa phân phối của các ước lượng.

44.2 Mô hình dự báo tài chính bằng mô phỏng - Minh họa với trường hợp mô phỏng giá cổ phiếu bằng ARIMA + Monte Carlo

44.2.1 Giới thiệu phương pháp

Phương pháp này kết hợp mô hình **ARIMA** để ước lượng chuỗi thời gian giá cổ phiếu và phương pháp **Monte Carlo** để mô phỏng nhiều kịch bản giá trong tương lai với các sai số ngẫu nhiên.

- **ARIMA (AutoRegressive Integrated Moving Average)**: Dự báo giá cổ phiếu dựa trên cấu trúc tự hồi quy và trung bình trượt của chuỗi thời gian.
- **Monte Carlo Simulation**: Dự báo bằng cách tạo nhiều đường giá tương lai với nhiễu ngẫu nhiên để đánh giá sự không chắc chắn.

44.2.2 Mô hình ARIMA

Mô hình ARIMA(p, d, q) được xác định bởi:

$$(1 - \sum_{i=1}^{p} \phi_i L^i)(1 - L)^d Y_t = (1 + \sum_{j=1}^{q} \theta_j L^j)\epsilon_t \tag{44.10}$$

Trong đó:

- p là số bậc tự hồi quy (AR).
- d là số lần lấy sai phân (I).
- q là số bậc trung bình trượt (MA).
- $\epsilon_t \sim N(0, \sigma^2)$ là nhiễu trắng.

Sau khi ước lượng mô hình ARIMA từ dữ liệu lịch sử, ta sẽ sử dụng nó để dự báo xu hướng giá tương lai.

44.2.3 Mô phỏng Monte Carlo dựa trên ARIMA

Sau khi có mô hình ARIMA, ta mô phỏng Monte Carlo bằng cách:

1. Lấy **giá dự báo từ ARIMA**.
2. Thêm **nhiễu ngẫu nhiên** $\epsilon_t \sim N(0, \hat{\sigma}^2)$.
3. Lặp lại nhiều lần để tạo các kịch bản khác nhau.

Công thức dự báo giá cổ phiếu:

$$S_{t+1} = \hat{S}_{t+1}^{ARIMA} + \sigma \cdot Z_t \tag{44.11}$$

với $Z_t \sim N(0, 1)$, mô phỏng nhiều kịch bản giá khác nhau.

Dưới đây là mã Python để thực hiện mô phỏng giá cổ phiếu bằng mô hình ARIMA kết hợp với Monte Carlo.

Cài đặt thư viện

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from statsmodels.tsa.arima.model import ARIMA


num_simulations = 1000
forecast_days = 30
```

Đọc dữ liệu và ước lượng mô hình ARIMA

```

np.random.seed(42)
returns = np.random.normal(0, 1, 500)
```

```
prices = 100 + np.cumsum(returns)

df = pd.DataFrame({'Price': prices})


model = ARIMA(df['Price'], order=(1, 1, 1))
model_fit = model.fit()
```

Mô phỏng Monte Carlo

```

simulated_paths = np.zeros((num_simulations, forecast_days)
    )

for i in range(num_simulations):
forecast = model_fit.forecast(steps=forecast_days)
noise = np.random.normal(0, np.std(forecast), forecast_days
    )
simulated_paths[i, :] = forecast + noise
```

Hiển thị kết quả

```

plt.figure(figsize=(10, 5))
plt.plot(simulated_paths.T, color='lightblue', alpha=0.2)
plt.plot(np.mean(simulated_paths, axis=0), color='red',
    label='Trung binh du bao')
plt.legend()
plt.title('Mo phong gia co phieu bang ARIMA + Monte Carlo')
plt.xlabel('Ngay')
plt.ylabel('Gia co phieu')
plt.show()
```

Phương pháp kết hợp ARIMA và mô phỏng Monte Carlo giúp dự báo kịch bản giá cổ phiếu với độ không chắc chắn được mô hình hóa bằng nhiễu ngẫu nhiên.

Phần X

Phụ lục

Tài liệu tham khảo

Angrist, J. D., and J.-S. Pischke (2009): "Mostly Harmless Econometrics: An Empiricist's Companion," *Princeton University Press*.

Wooldridge, J. M. (2019): *Introductory Econometrics: A Modern Approach*. Cengage Learning, Boston, MA, 7th edn.