การวิเคราะห์ข้อมูลด้วยโปรแกรม SPSS : กรณี MLR

PH·KKU 40th ANNIVERSARY Faculty of Public Health, Khon Kaen University

ACEP ASEAN Cancer Epidemiology and Prevention Research Group

รองศาสตราจารย์ ดร. พงษ์เดช สารการ

สาขาวิชาวิทยาการระบาดและชีวสถิติ, คณะสาธารณสุขศาสตร์, มหาวิทยาลัยขอนแก่น

E-mail : spongd@kku.ac.th Website : https://pongdechonline.com

แนวทางการวิเคราะห์ Multiple Linear Regression ด้วยโปรแกรม SPSS

สำหรับการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรอิสระหลายตัวกับผลลัพธ์แบบต่อเนื่อง

1 ความหมายและโครงสร้างของตัวแบบ

Multiple linear regression ใช้เมื่อผลลัพธ์เป็นตัวแปรต่อเนื่อง และต้องการศึกษาความสัมพันธ์ของตัวแปรอิสระตั้งแต่ 2 ตัวขึ้นไป

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

- Y = ตัวแปรตาม
- $X_1 \ldots X_p$ = ตัวแปรอิสระ
- β_0 = ค่าคงที่
- $\beta_1 \ldots \beta_p$ = ค่าสัมประสิทธิ์ถดถอย
- ε = ค่าความคลาดเคลื่อน

คำถามหลัก: เมื่อควบคุมตัวแปรอื่นคงที่ X แต่ละตัวสัมพันธ์กับ Y หรือไม่?

2 คำถามวิจัย วัตถุประสงค์ และสมมติฐาน

คำถามวิจัย (ตัวอย่าง):

"อายุ ดัชนีมวลกาย การสูบบุหรี่ และการออกกำลังกาย มีความสัมพันธ์กับระดับความดันโลหิตหรือไม่?"

วัตถุประสงค์:

เพื่อศึกษาความสัมพันธ์ของตัวแปรอิสระหลายตัวกับผลลัพธ์แบบต่อเนื่อง

สมมติฐานรวมของโมเดล:

H_0: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$

H_1: มีอย่างน้อย 1 ตัวที่ $\beta \neq 0$

สมมติฐานรายตัวแปร:

H_0: $\beta_i = 0$

H_1: $\beta_i \neq 0$

กำหนดระดับนัยสำคัญ $\alpha = 0.05$

3 การกำหนดตัวแปร

- **ตัวแปรตาม (Dependent variable):** ต้องเป็นข้อมูลต่อเนื่อง เช่น ความดันโลหิต ระดับน้ำตาล คะแนนคุณภาพชีวิต
- **ตัวแปรอิสระ (Independent variables):** อาจเป็นต่อเนื่องหรือจัดกลุ่ม
- **ตัวแปรจัดกลุ่มต้องแปลงเป็น dummy variables**

ตัวอย่าง

Y : systolic blood pressure
X_1 : age
X_2 : BMI
X_3 : smoking (0=ไม่สูบ, 1=สูบ)
X_4 : exercise score

หมายเหตุ: หลีกเลี่ยงการใส่ตัวแปรที่ซ้ำซ้อนกันมากเกินไป

4 ข้อตกลงเบื้องต้นที่ต้องตรวจสอบ

- ความเป็นเชิงเส้นระหว่างตัวแปรอิสระกับผลลัพธ์ (Linearity)
- ความเป็นอิสระของค่าคลาดเคลื่อน (Independence of errors)
- การแจกแจงปกติของค่าคลาดเคลื่อน (Normality of residuals)
- ความแปรปรวนคงที่ (Homoscedasticity)
- ไม่มี multicollinearity รุนแรง
- ไม่มี influential outliers มากเกินไป

เกณฑ์/แนวทางย่อ

- Tolerance > 0.1, VIF < 10
- Durbin–Watson ใกล้ 2
- ตรวจ residual plot, histogram, P–P plot, scatterplot

5 การเตรียมข้อมูลในไฟล์ SPSS

จัดรูปแบบข้อมูลเป็น Wide format
(หนึ่งแถว = หนึ่งหน่วยตัวอย่าง)

id	age	bmi	smoking	exercise	sbp
1	45	24.3	0	60	128
2	52	27.8	1	30	142
3	34	22.1	0	90	118
4	60	29.5	1	20	150
...	...	...	...	...	...

สิ่งที่ควรตรวจ

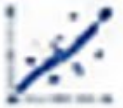
- missing values
- coding ของตัวแปร
- outliers
- type ของตัวแปร ใน Variable View
- label / value labels

คำหนดตัวแปรตามเป็น Scale และตรวจความถูกต้องก่อนวิเคราะห์

6 ขั้นตอนการวิเคราะห์ในโปรแกรม SPSS

1. นำตัวแปรตามใส่ช่อง Dependent
2. นำตัวแปรอิสระใส่ช่อง Independent(s)
3. เลือก Method เช่น Enter
   (อาจใช้ Stepwise / Forward / Backward ตามวัตถุประสงค์)
4. คลิก Statistics แล้วเลือก
   Estimates, Model fit, Confidence intervals, Collinearity diagnostics, Descriptives, Part and partial correlations
5. คลิก Plots เพื่อตรวจ ZPRED กับ ZRESID และเลือก Normal probability plot
6. คลิก Save หากต้องการ residuals หรือ predicted values
7. คลิก OK เพื่อวิเคราะห์

7 ผลลัพธ์สำคัญที่ต้องดู

1 Descriptive statistics	ค่าพื้นฐานของข้อมูล	
2 Model summary	R, R^2, Adjusted R^2, Std. Error	
3 ANOVA table	F statistic และ p-value ของโมเดลรวม	
4 Coefficients table	B, SE, Beta, t, p-value, 95% CI	
5 Collinearity statistics	Tolerance, VIF	
6 Residual diagnostics	histogram, P-P plot, residual scatterplot	

8 แนวทางการแปลผล

A โมเดลรวมมีนัยสำคัญ

หาก p-value จาก ANOVA < 0.05
แสดงว่าโมเดลโดยรวมมีความสัมพันธ์กับผลลัพธ์

B การแปลค่าสัมประสิทธิ์ B

"เมื่อควบคุมตัวแปรอื่นคงที่ หาก X เพิ่ม 1 หน่วย คาดว่า Y เปลี่ยนไป B หน่วย"

C การแปล standardized Beta

ใช้เปรียบเทียบความสำคัญสัมพัทธ์ของตัวแปรอิสระ

D การพิจารณา R^2 / Adjusted R^2

อธิบายสัดส่วนความแปรปรวนของ Y ที่โมเดลอธิบายได้
ย้ำให้พิจารณาร่วมกับ 95% CI และความหมายเชิงปฏิบัติ

9 รูปแบบการรายงานผล

ตัวอย่างการรายงาน

ผลการวิเคราะห์ multiple linear regression พบว่าโมเดลมีนัยสำคัญทางสถิติ ($F = ..., p < 0.05$) และสามารถอธิบายความแปรปรวนของผลลัพธ์ได้ร้อยละ ... (Adjusted $R^2 = ...$). เมื่อควบคุมตัวแปรอื่นคงที่ อายุ ($B = ...$, 95% CI ..., $p = ...$) และ BMI ($B = ...$, 95% CI ..., $p = ...$) มีความสัมพันธ์กับระดับความดันโลหิตอย่างมีนัยสำคัญ

สิ่งที่ควรรายงาน

- B (Unstandardized coefficient)
- SE (มาตรฐาน)
- Beta (Standardized coefficient)
- 95% CI
- p-value
- R^2 / Adjusted R^2
- F statistic ของโมเดลรวม

สรุป Workflow 6 ขั้นตอน

หัวใจของ **Multiple Linear Regression** คือ การประเมินความสัมพันธ์ของตัวแปรอิสระหลายตัว กับผลลัพธ์แบบต่อเนื่อง ภายใต้การควบคุมตัวแปรอื่นในโมเดล

Thank you

E-mail : spongd@kku.ac.th

https://pongdechonline.com