ความสัมพันธ์เชิงฟังก์ชันระหว่างค่าสัมประสิทธิ์การตัดสินใจประเภทต่าง ๆ สำหรับการวิเคราะห์ การถดถอยโลจิสติกที่ได้จาก SPSS

The Functional Relationship between the Coefficients of Determination for Logistic Regression analysis produced by SPSS

แสงหล้า ชัยมงคล

ภาควิชาคณิตศาสตร์และสถิติ คณะวิทยาศาสตร์และเทคโนโลยี

มหาวิทยาลัยธรรมศาสตร์ ศูนย์รังสิต ปทุมธานี 12121

1.บทนำ

ในการวิเคราะห์การถดถอยเชิงเส้นค่าสัมประสิทธิ์การตัดสินใจ (coefficient of determination or R^2) ถูกใช้เป็นเครื่องมือในการวัดระดับความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระ หรืออาจกล่าวได้ว่า R^2 เป็นค่าที่ใช้ระบุถึงสัดส่วนของความผันแปรของตัวแปรตามที่สามารถอธิบายได้ (explained variation) ด้วยตัวแปรอิสระ และเช่นเดียวกับการวิเคราะห์การถดถอยโลจิสติก (logistic regression) ที่ค่า R^2 สามารถใช้วัดระดับความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระได้เช่นกัน เพียงแต่การใช้ค่า R^2 สำหรับการวิเคราะห์การถดถอยโลจิสติกนั้นมีความยุ่งยากกว่าค่า R^2 ของการวิเคราะห์การถดถอยเชิงเส้น เนื่องจาก R^2 ของการวิเคราะห์การถดถอยโลจิสติกคำนวณได้หลายวิธีและแต่ละวิธีนั้นไม่จำเป็นที่จะต้องมีความสัมพันธ์กันทางรูปแบบสมการคณิตศาสตร์ หรือแนวคิด ทำให้จึงยังไม่มีข้อสรุปที่ชัดเจนว่าควรจะใช้ R^2 รูปแบบใดในการวัดขนาดของความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระ ดังนั้นในปัจจุบันโปรแกรมสำเร็จรูปทางสถิติที่สำคัญ เช่น SPSS, SAS, และ STATA เป็นต้น จึงได้มีการรายงานค่า R^2 หลายๆค่าสำหรับการถดถอยโลจิสติก อนึ่งการใช้ค่า R^2 หลายๆค่าในการตรวจสอบตัวแบบเดียวนั้น นอกจากจะทำให้มีการคำนวณที่เพิ่มขึ้นแล้ว ยังก่อให้เกิดความสับสนในการอธิบายและการนำไปใช้ประโยชน์ นอกจากนี้ค่า R^2 บางตัวยังไม่สามารถอธิบายความหมายให้เข้าใจได้ง่าย จึงสมควรหาความสัมพันธ์เชิงฟังก์ชันของค่า R^2 ต่างๆเหล่านี้กับค่า R^2 ที่สามารถอธิบายความหมายได้ดีและเป็นที่เข้าใจได้ง่าย

ปัจจุบันโปรแกรมสำเร็จ SPSS (ซึ่งเป็นโปรแกรมที่นิยมใช้ทางสถิติที่สำคัญโปรแกรมหนึ่ง) ได้รายงานค่า R^2 ไว้ 2 ค่า คือ Cox and Snell R^2 (R^2_{CS}) และ Negelkerke R^2 (R^2_N) [1] ที่มีชื่อเรียกโดยทั่วไปว่า Psuedo R^2 ซึ่งค่า R^2 ทั้งสองนี้มีข้อเสียตรงที่ไม่สามารถอธิบายความหมายในแง่ที่เข้าใจได้ ดังนั้นจึงจำเป็นที่จะต้องหาความสัมพันธ์เชิงฟังก์ชันของค่า R^2 ทั้งสองนี้กับค่า อัตราส่วนความควรจะเป็น R^2 (likelihood ratio $R^2 : R^2_L$) ค่า R^2_L ซึ่งนอกจากจะมีประโยชน์ในแง่ที่สามารถใช้อธิบายได้เสมือนกับค่า R^2 ของตัวแบบการถดถอยเชิงเส้นแล้ว ยังมีรูปแบบสมการที่สัมพันธ์กับค่า R^2 ของตัวแบบการถดถอยเชิงเส้นอีกด้วย ดังนั้นในบทความนี้ ได้แสดงความสัมพันธ์เชิงฟังก์ชันของค่า R^2_{CS} และ R^2_N ของโปรแกรม SPSS กับค่า R^2_L โดยหัวข้อ 2 จะได้อธิบายค่า R^2 ทั้ง 3 ค่านี้ หัวข้อ 3 จะแสดงความสัมพันธ์เชิงฟังก์ชันระหว่าง R^2_{CS} และ R^2_N กับค่า R^2_L หัวข้อ 4 แสดงให้เห็น

ความสัมพันธ์ของค่า R^2 ต่างๆโดยใช้ตัวอย่างจริง และจะสรุปในหัวข้อ 5

2.นิยามและรายละเอียดของค่า R^2 ประเภทต่างๆ

ปัจจุบันได้มีผู้เสนอค่า R^2 สำหรับตัวแบบการถดถอยโลจิสติกไว้หลายค่า [2, 3] ซึ่งสามารถจัดกลุ่มย่อยๆได้ เช่น (1) R^2 ที่วัดด้วยหลักการของสัดส่วนการลดลงของการกระจาย (proportional reduction in dispersion), (2) R^2 ที่วัดด้วยอาศัยฟังก์ชันควรจะเป็น (likelihood function) และ (3) R^2 ที่วัดด้วยข้อมูลที่จัดอันดับ (rank information) [2] เป็นต้น ในหัวข้อนี้จะได้อธิบายเฉพาะค่า R^2_{CS}, R^2_N และ R^2_L โดยทั้ง 3 ตัวนี้ จัดว่าเป็น R^2 ในกลุ่ม (2) นอกจากนี้เฉพาะ R^2_L ที่สามารถจัดในกลุ่ม (1) ได้อีกด้วย

2.1 R^2_L : อัตราส่วนความควรจะเป็น R^2

ให้ L_0 แทน ฟังก์ชันควรจะเป็นสำหรับตัวแบบที่ประกอบด้วยค่าคงที่เท่านั้น

L_M แทน ฟังก์ชันควรจะเป็นสำหรับตัวแบบ ที่ประกอบด้วยตัวแปรอิสระที่สนใจ และ

L_S แทนฟังก์ชันควรจะเป็นสำหรับตัวแบบ เต็ม (saturated model) หรือตัวแบบที่มีจำนวนพารามิเตอร์เท่ากับจำนวนค่าสังเกต

และกำหนดให้ สถิติดีเวียนซ์ (deviance) ของตัวแบบที่มีตัวแปรอิสระ (D_M) และตัวแบบที่มีเฉพาะค่าคงที่ (D_0) เป็นดังต่อไปนี้

$$D_M \equiv -2(\log L_M - \log L_S) \equiv -2\log(L_M / L_S)$$
$$D_0 \equiv -2(\log L_0 - \log L_S) \equiv -2\log(L_0 / L_S)$$

ดังนั้นรูปแบบทั่วไปของตัววัดที่คำนวณด้วยฟังก์ชันควรจะเป็นคือ

$$G = \frac{\log L_M - \log L_0}{\log L_S - \log L_0} = \frac{D_0 - D_M}{D_0}$$

ซึ่งเป็นสัดส่วนการลดลงของดีเวียนซ์อันเนื่องจากตัวแบบที่สนใจ

สำหรับตัวแบบการถดถอยโลจิสติกที่ตัวแปรตามมีค่าที่เป็นไปได้เพียง 2 ค่า คือ 1 ถ้าเหตุการณ์ที่สนใจเกิดขึ้น และ 0 กรณีอื่นๆ ดังนั้นสำหรับตัวแบบเต็ม (saturated model) ค่าประมาณความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจเมื่อกำหนดเวกเตอร์ของตัวแปรอิสระจะมีค่าเท่ากับค่าสังเกตนั้นๆหรือเขียนเป็นสัญลักษณ์ได้ $\hat{P}(y_i = 1 \mid x_i) \equiv \hat{\pi}_i = y_i$ และฟังก์ชันควรจะเป็นจะมีค่าเป็น 1 นั่นคือ

$$L_S = \prod_{i=1}^{n} y_i^{y_i} \times (1 - y_i)^{(1-y_i)} = 1$$

นอกจากนั้น Hosmer และ Lemeshow (p.13) [4] ให้ข้อสังเกตว่า ค่าสถิติดีเวียนซ์สำหรับตัวแบบการถดถอยโลจิสติกจะมีค่าเท่ากับผลรวมความคลาดเคลื่อนกำลังสอง (sum of square error: SSE) ที่คำนวณโดยวิธีกำลังสองน้อยสุดแบบสามัญ (ordinary least square: OLS) นั่นคือ จะได้ว่าค่า $D_M = -2\log L_M$ จะมีค่าเท่ากับ SSE ของตัวแบบที่มีตัวแปรอิสระ และ $D_0 = -2\log L_0$ จะมีค่าเท่ากับ SSE ของตัวแบบที่มีเฉพาะค่าคงที่ซึ่งเปรียบเสมือนได้กับค่าผลรวมกำลังสองทั้งหมด (total sum of square: SST) ของ OLS ดังนั้น อัตราส่วนความควรจะเป็น R^2 หรือ $R_L{}^2$ สำหรับตัวแบบการถดถอยโลจิสติก สามารถเขียนได้ดังนี้

$$R^2_L = 1 - \frac{\log(L_M)}{\log(L_0)} = \frac{\log L_0 - \log L_M}{\log L_0} \qquad (2.1)$$

ซึ่งจะเห็นว่า R^2_L สามารถอธิบายถึงสัดส่วนการลดลงของความแปรผันที่เกิดขึ้นในตัวแปรตาม เนื่องมาจากตัวแปรอิสระที่อยู่ในตัวแบบ

2.2 R^2_{CS} และ R^2_N

Cox and Snell R^2 (R^2_{CS}) นิยามได้ดังนี้ [5]

$$R^2_{CS} = 1 - \exp\{-\frac{2}{n}[\log L_M - \log L_0]\} = 1 - \left(\frac{L_0}{L_M}\right)^{2/n} \qquad (2.2)$$

เมื่อ L_0 และ L_M นิยามได้เช่นเดียวกับข้างต้น เนื่องจากค่า L_M เป็นฟังก์ชันผลคูณของความน่าจะเป็น ดังนั้นจึงทำให้ค่า R_{CS}^2 มีค่าต่ำกว่า 1 และมีค่าสูงสุดเท่ากับ

$$\max(R_{CS}^2) = 1-(L_0)^{2/n}$$

Nagelkerke [6] พบว่าสำหรับตัวแบบถดถอยโลจิสติกที่มี ค่า $y=1$ และ $y=0$ ในสัดส่วนที่เท่าๆกัน แล้ว ค่าสูงสุดของค่า R_{CS}^2 จะมีค่าเพียง 0.75 เท่านั้น ดังนั้น Nagelkerke [6] จึงได้เสนอให้มีการปรับค่า R_{CS}^2 โดยการนำค่าสูงสุดไปหารค่า R_{CS}^2 เดิมและเรียกค่าที่ได้นี้ว่า R_N^2 ซึ่งสามารถคำนวณได้จากสูตร

$$R_N^2 = \frac{1-\left(\dfrac{L_0}{L_M}\right)^{2/n}}{1-L_0^{2/n}} \qquad (2.3)$$

3.ความสัมพันธ์เชิงฟังก์ชันระหว่างค่า R_{CS}^2 และ R_N^2 กับ R_L^2

จากนิยามของค่า R_{CS}^2

$$\begin{aligned} R_{CS}^2 &= 1-\left(L_0/L_M\right)^{2/n} \\ &= 1-\exp\{-\frac{2}{n}[\log L_M - \log L_0]\} \\ &= 1-\exp\{-\frac{2\log L_0}{n}\left[\frac{\log L_M - \log L_0}{\log L_0}\right]\} \end{aligned}$$

ถ้ากำหนดให้

$$\delta = -2\log L_0 / n \qquad (3.1)$$

แล้วจะได้ว่า

$$R_{CS}^2 = 1-\exp\{-\delta R_L^2\} \qquad (3.2)$$

และในกรณีที่ ค่า R_L^2 มีค่าน้อยอย่างพอเพียง (sufficiently small value) สมการ (3.2) สามารถเขียนความสัมพันธ์ในรูปเชิงเส้นได้ดังนี้

$$R_{CS}^2 \approx \delta R_L^2 \qquad (3.3)$$

Sharma (p. 95) [7] ได้ศึกษาอิทธิพลของสัมประสิทธิ์การตัดสินใจของตัวแบบแฝง (latent model) หรือ ρ^2 และโอกาสของการเกิดเหตุการณ์ที่สนใจหรือ $P(y=1)\equiv\bar{\pi}$ ที่มีต่อความสัมพันธ์เชิงฟังก์ชันระหว่าง R_{CS}^2 และ R_L^2 พบว่า สำหรับค่า ρ^2 ที่มีค่าน้อยอย่างพอเพียง แล้ว จะได้ว่า

$$R_L^2 < R_{CS}^2 \quad \text{เมื่อ } 0.2<\bar{\pi}<0.8 \qquad (3.4)$$

และ $\quad R_L^2 > R_{CS}^2 \quad$ ในกรณีอื่นๆ

ทั้งนี้เนื่องจากทราบว่าค่าสูงสุดของค่าลอการิทึมของความควรจะเป็น (log-likelihood) ของตัวแบบที่มีเฉพาะค่าคงที่

$$\log L_0 = n[\bar{\pi}\log\bar{\pi} + (1-\bar{\pi})\log(1-\bar{\pi})]$$

ดังนั้น δ สามารถเขียนเป็นฟังก์ชันของ $\bar{\pi}$ ได้ดังนี้

$$\delta = -2[\bar{\pi}\log\bar{\pi} + (1-\bar{\pi})\log(1-\bar{\pi})]$$ จะสังเกตว่า δ เป็นฟังก์ชันพาราโบลาของ $\bar{\pi}$ ที่มีค่าสูงสุดเมื่อ $\bar{\pi}=0.05$ นอกจากนั้นยังพบว่า $\delta>1$ ถ้า $0.2<\bar{\pi}<0.8$ และ $\delta<1$ ในกรณีอื่นๆ

ในทำนองเดียวกัน ความสัมพันธ์ระหว่าง R_N^2 และ R_L^2 สามารถกำหนดได้ดังต่อไปนี้

$$R_N^2 = \frac{1-\exp(-\delta R_L^2)}{1-L_0^{2/n}} \qquad (3.5)$$

และจากสมการ (3.1) จะได้ว่า

$$L_0 = \exp(-\frac{n}{2}\delta)$$

เมื่อแทน L_0 ในสมการ (3.5) จะได้ความสัมพันธ์ระหว่าง R_N^2 และ R_L^2 ดังต่อไปนี้

$$\begin{aligned} R_N^2 &= \frac{1-\exp(-\delta R_L^2)}{1-\exp(-\delta)} \\ &\approx \frac{\delta}{1-\exp(-\delta)}R_L^2 \qquad (3.6) \\ &= \delta^* R_L^2 \end{aligned}$$

เมื่อ $\delta^* = \delta/(1-\exp(-\delta))$

นอกจากนั้น Sharma (p.96) [7] ยังได้แสดงให้เห็นว่า $R_N^2 > R_L^2$ และ $R_N^2 > R_{CS}^2$ ทุกค่าของ ρ^2 และ $\bar{\pi}$

4.ตัวอย่างเชิงตัวเลขแสดงความสัมพันธ์ระหว่าง R^2 ทั้ง 3 ค่า

ตัวอย่างที่นำมาแสดงเป็นตัวอย่างเกี่ยวกับโรคหลอดเลือดหัวใจตีบ (CHD) [8] โดย $y=1$ ถ้าเกิดโรค และ $y=0$ ถ้าไม่เกิดโรค ของคนจำนวน 100 คน โดยมี

ตัวแปรอิสระคือ อายุ ผลลัพธ์จากโปรแกรม SPSS ที่สอดคล้องกับค่าต่างๆ ที่ต้องการแสดงความสัมพันธ์มีดังนี้

ตาราง 1. Dependent variable Frequencies

		Freq	Percent	Valid Percent	Cum Percent
Valid	0	57	57.0	57.0	57.0
	1	43	43.0	43.0	100.0
	Total	100	100.0	100.0	

ตาราง 1 แสดงค่าความถี่ของค่า Y จากการวิเคราะห์ความถี่ ได้ค่าความน่าจะเป็นที่คนจะเกิดโรคหลอดเลือดหัวใจตีบหรือ $P(y=1) \equiv \bar{\pi}$ เท่ากับ 0.43

ตาราง 2. Block 0: Beginning Block

Iteration History (a,b,c)

Iteration		-2 Log likelihood	Coefficients Constant
Step 0	1	136.663	-.280
	2	136.663	-.282
	3	136.663	-.282

a Constant is included in the model.

b Initial -2 Log Likelihood: 136.663

c Estimation terminated at iteration number 3 because parameter estimates changed by less than .001.

ตาราง 2 แสดงค่า $-2\log L_0$ และสัมประสิทธิ์ของตัวแบบการถดถอยโลจิสติกที่มีเฉพาะค่าคงที่ โดยทำการประมาณค่าด้วยวิธีความควรจะเป็นสูงสุดจำนวน 3 รอบ พบว่าค่าสัมประสิทธิ์การถดถอยในรอบที่ 2 และ 3 มีค่าไม่แตกต่างกัน และค่า $-2\log L_0$ มีค่าไม่เปลี่ยนแปลง ดังนั้นจึงได้ว่า ค่าประมาณของ $-2\log L_0$ มีค่าเท่ากับ 136.663

ตาราง 3. Block 1: Method = Enter

Iteration History(a,b,c,d)

Iteration		-2 Log likelihood	Coefficients Constant	AGE
Step 1	1	108.493	-4.152	.087
	2	107.366	-5.181	.108
	3	107.353	-5.308	.111
	4	107.353	-5.309	.111
	5	107.353	-5.309	.111

a Method: Enter

b Constant is included in the model.

c Initial -2 Log Likelihood: 136.663

d Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

ตาราง 3 แสดงค่า $-2\log L_M$ และสัมประสิทธิ์ของตัวแบบการถดถอยโลจิสติกที่นำตัวแปรอิสระอายุเพิ่มเข้ามาในสมการ การประมาณค่าพารามิเตอร์ต่างๆของตัวแบบด้วยวิธีความควรจะเป็นสูงสุดจะทำ 5 รอบ และได้ตัวแบบการถดถอยโลจิสติกที่แสดงความสัมพันธ์ระหว่างอายุและโอกาสเสี่ยงต่อการเกิดโรคหลอดเลือดตีบตันดังนี้

$$\frac{\log \hat{P}(y=1)}{1-\log \hat{P}(y=1)} = -5.309 + 0.111x$$

และค่า $-2\log L_M$ = 107.353

ตาราง 4. Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	107.353	.254	.341

ตาราง 4 แสดงค่า $-2\log L_M = 107.353$ สำหรับตัวแบบที่มีตัวแปรอิสระอายุ $R_{CS}^2 = 0.254$ และ $R_N^2 = 0.341$ จากค่า R_{CS}^2 สามารถอธิบายเหมือนกับค่า R^2 ของตัวแบบการถดถอยเชิงเส้นในเทอมของการปรับปรุงค่าเฉลี่ยเรขาคณิตต่อจำนวนค่าสังเกตหรือการลดลงของความคลาดเคลื่อนที่เกิดจากตัวแบบที่มีตัวแปร

อิสระเมื่อเทียบกับตัวแบบที่มีเพียงค่าคงที่ [8] ในกรณีนี้สามารถอธิบายได้ว่า อายุสามารถลดความคลาดเคลื่อนของโอกาสเสี่ยงต่อการเกิดโรคหลอดเลือดตีบตันร้อยละ 25.4 ต่อจำนวนค่าสังเกตเมื่อเทียบกับการไม่มีตัวแปรอายุ ในขณะที่ R_N^2 นั้นไม่สามารถอธิบายความหมายในแง่นี้ได้ [8] ดังนั้นเพื่อให้สามารถอธิบายความหมายของ R^2 สำหรับการถดถอยโลจิสติก เช่นเดียวกับค่า R^2 ของการถดถอยเชิงเส้นทั่วไป จำเป็นต้องคำนวณค่า R_L^2 และหาความสัมพันธ์ระหว่าง R_{CS}^2 และ R_N^2 กับ R_L^2

โปรแกรม SPSS ไม่ได้รายงานค่า R_L^2 โดยตรง แต่สามารถคำนวณได้จากสมการ (2.1) และเนื่องจาก SPSS ให้ค่า -2log-likelihood ดังนั้น ค่า R_L^2 คำนวณได้จาก

$$R_L^2 = 1 - \frac{\log L_M}{\log L_0}$$

$$= 1 - \frac{(-2\log L_M)}{(-2\log L_0)} = 1 - \frac{107.353}{136.663} = 0.214$$

หมายความว่า อายุสามารถอธิบายความคลาดเคลื่อนของโอกาสเสี่ยงต่อการเกิดโรคหลอดเลือดตีบตัน ร้อยละ 21.4 และ

$$\delta = -2\log L_0 / n$$

$$= 136.663/100 = 1.367$$

ดังนั้นค่า R_{CS}^2 สามารถคำนวณจาก R_L^2 โดยใช้ความสัมพันธ์เชิงฟังก์ชันในสมการ (3.2) จะได้ว่า

$$R_{CS}^2 = 1 - \exp\{-\delta R_L^2\}$$

$$= 1 - \exp\{(-1.367)(0.214)\}$$

$$= 0.254$$

ซึ่งมีค่าเท่ากับที่ SPSS รายงาน ในกรณีที่พิจารณาว่าค่า R_L^2 นั้นมีค่าน้อยอย่างพอเพียง ค่า R_{CS}^2 สามารถประมาณได้โดยใช้สมการ (3.3) จะได้

$$R_{CS}^2 \approx \delta R_L^2 = (1.367)(0.214) = 0.293$$

ตัวอย่างนี้ ค่า π มีค่าเท่ากับ 0.43 ซึ่งอยู่ในช่วง (0.2, 0.8) ดังนั้นจากผลการศึกษาของ Sharma [7] จะได้ว่า

$$R_L^2 < R_{CS}^2 = 0.214 < 0.254$$

ในทำนองเดียวกัน R_N^2 สามารถหาจาก R_L^2 โดยใช้ความสัมพันธ์เชิงฟังก์ชันในสมการ (3.6) จะได้

$$R_N^2 = \frac{1 - \exp(-\delta R_L^2)}{1 - \exp(-\delta)}$$

$$= \frac{1 - \exp\{-(1.367)(0.214)\}}{1 - \exp(-1.367)} = 0.341$$

หรือประมาณได้จาก

$$R_N^2 \approx \frac{\delta}{1 - \exp(-\delta)} R_L^2$$

$$= \frac{(1.367)(0.214)}{1 - \exp(-1.367)} = (1.835)(0.214) = 0.393$$

จาก R^2 ทั้ง 3 ค่าของตัวอย่างนี้ สรุปได้ว่า $R_L^2 < R_{CS}^2 < R_N^2$ ซึ่งสอดคล้องกับผลการศึกษาของ Sharma [7]

ในกรณีนี้การอธิบายความหมายของ R_{CS}^2 และ R_N^2 สามารถอธิบายในแง่จำนวนเท่าของค่า R_L^2 นั่นคือค่า R_{CS}^2 จะให้ค่าสูงกว่า ค่า R_L^2 โดยประมาณ 1.367 เท่า ทำนองเดียวกันกับค่า R_N^2 จะให้ค่าสูงกว่าค่า R_L^2 โดยประมาณ 1.835 เท่า

5. สรุป

บทความนี้ผู้เขียนได้แสดงความสัมพันธ์เชิงฟังก์ชันระหว่างค่า R^2 ที่ใช้ในโปรแกรมสำเร็จรูปทางสถิติ SPSS โดยกำหนดความสัมพันธ์ระหว่างค่า R^2 เหล่านี้ กับค่า R_L^2 ในรูปฟังก์ชันของลอการิทึมควรจะเป็นของตัวแบบที่มีเฉพาะค่าคงที่ (log likelihood of null model: $\log L_0$) ที่เป็นฟังก์ชันของความน่าจะเป็นที่จะเกิดเหตุการณ์ที่สนใจ การทราบความสัมพันธ์เชิงฟังก์ชันระหว่างค่า R^2 ต่างๆ นั้นจะทำให้ช่วยลดจำนวนค่า R^2 ที่ต้องคำนวณและยังช่วยลดความสับสนในการอธิบายและการนำค่าไปใช้ประโยชน์ นอกจากนี้ค่า R^2 บางตัวยังไม่สามารถอธิบายความหมายให้เข้าใจได้ง่าย

6. เอกสารอ้างอิง

[1] SPSS for Windows Release 11.5, Chicago: SPSS Inc., 2001.

[2] Menard, S., Coefficient of Determination for Multiple Logistic Regression Analysis, The American Statistician, Vol. 54; pp. 17 - 24, 2000.

[3] Mittlbock, M. and Schemper, M., Explained Variation for Logistic Regression, Statistics in Medicine, Vol. 15; pp. 113-121, 1996.

[4] Hosmer, D. W., and Lemeshow, S., Applied Regression Analysis, 2nd Edition., Wiley, New York, 2000.

[5] Maddala, G.S., Limitted - Dependent and Qualitative Variables in Economics, Cambridge University Press, Cambridge, 1983.

[6] Nagelkerke, N.J.D., A Note on a General Definition of Coefficient of Determination, Biometrika, Vol. 78; pp. 691-692, 1991.

[7] Sharma, D., Logistic Regression, Measure of Explained Variation, and the Base Rate Problem, Ph.D Thesis, Florida State University, Tallahassee U.S.A., 2006.

[8] Cox, D.R., and Wermuth, N., A Comment on the Coefficient of Determination for Binary Response, The American Statistician, Vol. 46; pp. 1-4, 1992.