การวิเคราะห์การถดถอยลอจิสติกด้วยเอกเซล
Logistic Regression by Excel Solver

มนตรี พิริยะกุล[1*]

บทคัดย่อ

ในการวิเคราะห์การถดถอยลอจิสติกซึ่งเป็นกรณีหนึ่งของการถดถอยพหุที่ตัวแปรตามมีค่าเป็น 0 กับ 1 นั้นมีซอฟท์แวร์ทางสถิติช่วยการวิเคราะห์อยู่หลายตัว แต่ในหลายกรณีที่นักวิจัยจำเป็นต้องเรียกใช้ผลการวิเคราะห์โดยต้องการส่งผลไปกลับเช่นในกรณีส่งผลมาที่เมนูโต้ตอบในงาน DSS ที่เสนองานใน VBA การวิเคราะห์การถดถอยลอจิสติกด้วยเอกเซลจะช่วยให้ทำงานได้ง่ายขึ้นโดยเฉพาะเมื่อมีปัญหาการส่งผลไปกลับระหว่างซอฟท์แวร์

คำสำคัญ: การถดถอยลอจิสติก, excel solver

Abstract

There are several statistical software for analyzing logistic regression which is a special case of multiple regression analysis that dependent variable assumes values of 0 or 1 only. However, in case of linking between excel application in VBA environment, DSS dialogue for example, logistic regression with excel would better support the between software data transfer.

Keywords: logistic regression, excel solver

บทนำ

การถดถอยลอจิสติก คือ การถดถอยพหุที่พัฒนาขึ้นมาเพื่อใช้ในกรณีที่ตัวแปรตามมีลักษณะเป็นตัวแปรกลุ่ม (categorical variable) หรือตัวแปรนามบัญญัติ (nominal variable) ในกรณีนี้นักวิจัยจะกำหนดรหัสให้แก่ค่าตัวแปรเป็นเลข 0 กับ 1 เช่น (มนตรี พิริยะกุล, 2544)

1 = องค์การปรับปรุงแก้ไขข้อบกพร่องตามคำแนะนำของลูกค้า
0 = องค์การไม่ปรับปรุงแก้ไขข้อบกพร่องตามคำแนะนำของลูกค้า
1 = พนักงานยอมรับการหมุนเวียนเปลี่ยนงาน
0 = พนักงานไม่ยอมรับการหมุนเวียนเปลี่ยนงาน
1 = ลูกค้าตั้งใจกลับมาซื้อซ้ำ
0 = ลูกค้าตั้งใจไม่กลับมาซื้อซ้ำ
1 = คนไข้ป่วยเป็นโรคไต
0 = คนไข้ไม่ป่วยเป็นโรคไต
1 = สถานประกอบการล้มละลาย
0 = สถานประกอบการไม่ล้มละลาย

ในกรณีเหล่านี้หากวิเคราะห์ข้อมูลด้วยการถดถอยพหุจะพบว่าค่าพยากรณ์จากสมการถดถอยจะมีค่าใด ๆ ที่มักจะไม่ใช่ 0 หรือ 1 และมีความเป็นไปได้ที่จะมีค่าน้อยกว่า 0 หรือมากกว่า 1 ซึ่งมีผลให้ไม่อาจแปลผลได้

จะสังเกตได้ว่าค่าตัวแปรมีความหยาบกว่ากรณีที่เป็นตัวแปรในมาตราช่วงหรืออัตราส่วน ทั้งนี้ขึ้นอยู่กับเป้าหมายของการศึกษาว่าจะศึกษาอะไร ในกรณีระบบสนับสนุนการตัดสินใจ (decision support system--DSS) การถดถอยลอจิสติกนับเป็นส่วนหนึ่งของฐานตัวแบบที่ใช้ประกอบการ

[1] รองศาสตราจารย์ ดร. ภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยรามคำแหง, E-mail: mpiriyakul@yahoo.com

หลักเกณฑ์ทางทฤษฎี

จากสมการ $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \ldots + \beta_kX_k + u$ ... (1)

โดยที่ Y = 0, 1 และ $X_1, X_2, \ldots, X_k$เป็นตัวแปรในมาตราวัดใด ๆ และ u คือส่วนเหลือ (residual)

ให้ P_i = ความน่าจะเป็นที่ Y_iจะมีค่าเท่ากับ 1

Q_i= 1- P_i = ความน่าจะเป็นที่ Y_iจะมีค่าเท่ากับ 0

ดังนั้นจากสมการ(1) คือ $Y_i = \beta_0 + \beta_1X_{1i} + \beta_2X_{2i} + \ldots + \beta_kX_{ki} + u_i$จะมีค่าคาดหมายและความผันแปรดังนี้คือ (Hosmer, Lemeshow and Sturdivant, 2013)

$E(Y_i) = E\{\beta_0 + \beta_1X_{1i} + \beta_2X_{2i} + \ldots + \beta_kX_{ki} + u_i\}$

$= \beta_0 + \beta_1X_{1i} + \beta_2X_{2i} + \ldots + \beta_kX_{ki}$ ...(2)

แต่ $E(Y_i) = \sum_0^1\{Y_i * prob(Y_i)\} = 1*P_i + 0* Q_i = P_i$ ...(3)

ผลจากสมการ (2) และสมการ (3)แสดงว่า

$P_i = \beta_0 + \beta_1X_{1i} + \beta_2X_{2i} + \ldots + \beta_kX_{ki}$ ...(4)

และจากสมการ(1) คือ $Y_i = \beta_0 + \beta_1X_{1i} + \beta_2X_{2i} + \ldots + \beta_kX_{ki} + u$

แสดงว่า $Y_i = P_i + u_i$ ... (5)

จากสมการ (3) จะพบข้อขัดแย้งกับข้อตกลงการถดถอย 2 ประการสำคัญคือ

1. จากสมการ (1) พบว่า$u_i = Y_i–(\beta_0 + \beta_1X_{1i} + \beta_2X_{2i} + \ldots + \beta_kX_{ki})$ และอาศัยความรู้จากสมการ (5) แสดงว่า u_i มีค่าเพียง 2 ค่าคือ

$1– (\beta_0 + \beta_1X_{1i} + \beta_2X_{2i} + \ldots + \beta_kX_{ki})$และ

$0 – (\beta_0 + \beta_1X_{1i} + \beta_2X_{2i} + \ldots + \beta_kX_{ki})$

ซึ่งเป็นตัวแปรสุ่มเบอร์นูลลิ มิใช่ตัวแปรสุ่มปกติตามข้อตกลงการถดถอย

2. $V(u_i) = V(Y_i) = E\{Y_i – E(Y_i)\}^2$ และเมื่ออาศัยความรู้ตามสมการ (3) และสมการ (4) จะพบว่า

$V(Y_i) = \sum_0^1\left(Y_i - E(Y_i)\right)^2 prob(Y_i)$

$= (1-P_i)^2*P_i + (0- P_i)^2*Qi = Q_i^2P_i + P_i^2Q_i$

$= P_iQ_i(Q_i + P_i)$

$= P_iQ_i$

$= (\beta_0 + \beta_1X_{1i} + \beta_2X_{2i} + \ldots + \beta_kX_{ki})\{1-(\beta_0 + \beta_1X_{1i} + \beta_2X_{2i} + \ldots + \beta_kX_{ki})\}$

แสดงว่าความผันแปรของ u มิได้คงที่ตามข้อตกลงแต่กลับแปรไปตามค่าของ $X_1, X_2, \ldots, X_k$

เนื่องจาก $P_i = \beta_0 + \beta_1X_{1i} + \beta_2X_{2i} + \ldots + \beta_kX_{ki}$ จึงมีความเป็นไปได้ที่ค่าพยากรณ์ของ P_i อาจตกอยู่นอกช่วง [0,1] เราจึงเปลี่ยนค่า P_iเป็นค่าออร์ดิเนตใต้โค้งลอจิสติก ดังนี้คือกำหนดให้

$P_i = \frac{e^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k}}{1+e^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k}} = \frac{1}{1+e^{-(\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\cdots+\beta_kX_{ki})}}$... (6)

จากสมการ (5) คือ $Y_i = P_i + u_i$และเนื่องจาก Y_i = 0 ,1 แสดงว่าY_iแจกแจงแบบBernoulli (P) คือ $f(y) = P_i^{Y_i}(1 - P_i)^{1-Y_i}$

ดังนั้น เพื่อประมาณค่า P_i โดยวิธี maximum likelihood จะพบว่า likelihood function คือ

$L = \prod_i^n f(y_i)$

$= \prod_i^n P_i^{Y_i}(1 - P_i)^{1-Y_i}$

$\ln L = \ln\{\prod_i^n P_i^{Y_i}(1 - P_i)^{1-Y_i}\}$

$\ln L = \sum_i^n \ln\{P_i^{Y_i}(1 - P_i)^{1-Y_i}\}$

$\ln L = \sum_i^n Y_i \ln P_i + \sum_i^n(1 - Y_i)\ln(1 - P_i)$ ... (7)

สมการ (7) นี้คือ สมการที่เราจะนำไป maximize ด้วย Excel Solver

กระบวนการ MLE พัฒนาต่อไปเพื่อแสดงให้เห็นการใช้ odd และ ln(odd) ดังนี้ คือ จากสมการ (7)

$\ln L = [\sum_i^n Y_i \ln P_i - \sum_i^n Y_i \ln(1-P_i)] + \sum_i^n \ln(1-P_i)$

$= \sum_i^n Y_i \ln\left(\frac{P_i}{1-P_i}\right) + \sum_i^n \ln(1-P_i)$ ...(8)

เพื่อให้เข้าใจง่ายสำหรับการพิสูจน์ต่อไปนี้ผู้เขียนจะกำหนดให้ $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k = A$

นั่นคือ $e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k} = e^A$

จาก (6) $P_i = \frac{e^{A_i}}{1+e^{A_i}} = \frac{1}{1+e^{-A_i}}$ ดังนั้น $1- P_i = 1- \frac{e^{A_i}}{1+e^{A_i}} = 1-\frac{1}{1+e^{-A_i}} = \frac{1}{1+e^{A_i}}$

จึงพบว่า $\frac{P_i}{1-P_i} = \frac{\frac{1}{1+e^{-A}}}{\frac{1}{1+e^{A}}} = \frac{1+e^{A_i}}{1+e^{-A_i}} = e^{A_i}$

นั่นคือ $\frac{P_i}{1-P_i} = e^{\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\cdots+\beta_k X_{ki}}$

ดังนั้นสมการ (8) จึงเปลี่ยนเป็น

$\ln L = \sum_i^n Y_i(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}) - \sum_i^n \ln\left(1 + e^{\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\cdots+\beta_k X_{ki}}\right)$

ผลจากการประมาณค่าด้วยวิธี MLE คือ $\frac{\partial \ln L}{\partial \beta_i} = 0$; i = 1, 2, ..., k จะได้ค่าประมาณของ $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ คือ $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ และความน่าจะเป็นที่สิ่งที่สนใจจะเกิดขึ้นคือ

$Pr(Y_i=1) = P_i = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X_{1i}+\hat{\beta}_2 X_{2i}+\cdots+\hat{\beta}_k X_{ki}}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X_{1i}+\hat{\beta}_2 X_{2i}+\cdots+\hat{\beta}_k X_{ki}}}$; i = 1, 2, 3, ..., n ซึ่งจะมีค่าตกอยู่ในช่วง [0,1]

อนึ่ง $\frac{P_i}{1-P_i} = \frac{1+e^{(\hat{\beta}_0+\hat{\beta}_1 X_{1i}+\hat{\beta}_2 X_{2i}+\cdots+\hat{\beta}_k X_{ki})}}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 X_{1i}+\hat{\beta}_2 X_{2i}+\cdots+\hat{\beta}_k X_{ki})}} = e^{\hat{\beta}_0+\hat{\beta}_1 X_{i1}+\hat{\beta}_2 X_{2i}+\cdots+\hat{\beta}_k X_{ki}}$

เรียก $\frac{P_i}{1-P_i}$ ว่า odd ซึ่งจะพบว่ามีค่า 0 ถึง 100 หรือ 1,000 แล้วแต่การกำหนดจำนวนจุดทศนิยมและเรียก ln(odd) ว่า logit(p)

ดังนั้น $odd_i = e^{\hat{\beta}_0+\hat{\beta}_1 X_{1i}+\hat{\beta}_2 X_{2i}+\cdots+\hat{\beta}_k X_{ki}}$ ...(9)

$\ln(odd_i) = \text{logit}(p_i) = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki}$ ...(10)

สมการ (9) ใช้สำหรับเปรียบเทียบโอกาสประสบผลสำเร็จเปรียบเทียบระหว่างหน่วยสังเกตที่ i กับหน่วยสำรวจที่ j คือ

จากสมการ (9)

$odd_m = e^{\hat{\beta}_0+\hat{\beta}_1 X_{1m}+\hat{\beta}_2 X_{2m}+\cdots+\hat{\beta}_k X_{km}}$ และ

$odd_j = e^{\hat{\beta}_0+\hat{\beta}_1 X_{1j}+\hat{\beta}_2 X_{2j}+\cdots+\hat{\beta}_k X_{kj}}$

ดังนั้น odd ratio (OR) คือ

$$\frac{odd_m}{odd_j} = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X_{1m}+\hat{\beta}_2 X_{2m}+\cdots+\hat{\beta}_k X_{km}}}{e^{\hat{\beta}_0+\hat{\beta}_1 X_{1j}+\hat{\beta}_2 X_{2j}+\cdots+\hat{\beta}_k X_{kj}}}$$

$$\frac{odd_m}{odd_j} = e^{\hat{\beta}_1(X_{1m}-X_{1j})+\hat{\beta}_2(X_{2m}-X_{2j})+\cdots+\hat{\beta}_k(X_{km}-X_{kj})}$$

เมื่อแทนค่า $(X_{1m} - X_{1j}), (X_{2m} - X_{2j}), \ldots, (X_{km} - X_{kj})$ จะได้ค่าเป็นตัวเลขที่แสดงให้เห็นว่าหน่วยสำรวจที่ m ซึ่งมีคุณสมบัติต่างจากหน่วยสำรวจที่ j จะมีโอกาสประสบความสำเร็จ คือ Pr(Y=1) ต่างกันเพียงใด เกณฑ์คือ

ถ้า OR = 1 หน่วยที่ mและหน่วยที่ j มีโอกาสสำเร็จเท่ากัน

ถ้า OR > 1 หน่วยที่ m มีโอกาสสำเร็จมากกว่าหน่วยที่ j

ถ้า OR < 1 หน่วยที่ m มีโอกาสสำเร็จน้อยกว่าหน่วยที่ j

การอ่านผลและการทดสอบสมมุติฐาน

การอ่านผลและการทดสอบสมมุติฐานจะกระทำกับสมการ (10) โดยที่เครื่องหมายบวกหรือลบของ $\hat{\beta}_j$ ใช้บ่งชี้ดังนี้คือ ถ้า $\hat{\beta}_j$ มีเครื่องหมายบวกแสดงว่าถ้า X_j มีค่าเพิ่มขึ้นจะส่งผลให้โอกาสประสบผลสำเร็จสูงขึ้น แต่ถ้า $\hat{\beta}_j$ มีเครื่องหมายลบแสดงว่าถ้า X_j มีค่าเพิ่มขึ้นจะส่งผลให้โอกาสประสบผลสำเร็จลดลง ขณะที่การตรวจสอบนัยสำคัญทางสถิติมีไว้เพื่อบ่งชี้ว่าตัวแปรอิสระมีผลกระทบต่อโอกาสประสบผลสำเร็จจริงหรือไม่ (ณ ระดับนัยสำคัญที่กำหนด) แต่ในการพยากรณ์จะไม่สนใจว่ามีนัยสำคัญหรือไม่ ส่วนการระบุว่าตัวแปรอิสระใดมีอิทธิพลต่อโอกาสประสบผลสำเร็จมากกว่ากันให้พิจารณาจาก odd ratio (OR) ซึ่งเป็นการเปรียบเทียบระหว่างสถานการณ์ คือ เมื่อ X_j เพิ่มค่าขึ้น 1 หน่วยจากเดิมจะมีผลให้โอกาสประสบผลสำเร็จเปลี่ยนไปอย่างไรดังนี้ (Kleinbaum and Mitchel, 2010; Menard, 2001)

จากสมการ(10) คือ $\ln(odd_1) = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki}$

ให้ X_j เพิ่มค่าเป็น $X_j + 1$ จะได้ ln(odd2) ดังนี้

$\ln(odd_2) = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_j (X_{ji} + 1) + \cdots + \hat{\beta}_k X_{ki}$

ดังนั้น $\ln(odd_2) - \ln(odd_1) = \{\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_j (X_{ji} + 1) + \cdots + \hat{\beta}_k X_{ki}\}$

$\{\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + + \hat{\beta}_j X_{ji} + \cdots + \hat{\beta}_k X_{ki}\} = \hat{\beta}_j (X_{ji} + 1) - \hat{\beta}_j X_{ji} = \hat{\beta}_j$

นั่นคือ $\ln(\frac{odd_2}{odd_1}) = \hat{\beta}_j$

ดังนั้น $\frac{odd_2}{odd_1} = e^{\hat{\beta}_j}$; j = 1, 2, 3, ... , k

หมายความว่า X_j; j = 1, 2, 3, ... , k เปลี่ยนค่าไป 1 หน่วยจะมีผลให้โอกาสสำเร็จสูงกว่าหรือต่ำกว่าโอกาสไม่สำเร็จร้อยละเท่าไร ตัวอย่างเช่น $\frac{odd_2}{odd_1} = e^{\hat{\beta}_1}$ = 1.91 (โปรแกรม SPSS เรียกว่า Exp(β_1))หมายความว่าถ้า X_1 มีค่าสูงขึ้น 1 หน่วยจะมีผลให้ OR เพิ่มขึ้นร้อยละ 91 (คือ 100*1.91-100 = 91) ซึ่งก็คือ โอกาสที่จะประสบผลสำเร็จสูงขึ้นร้อยละ 91 นั่นเอง และสมมุติว่า $\frac{odd_2}{odd_1} = e^{\hat{\beta}_2}$ = 0.85 แสดงว่าถ้า X_2 มีค่าสูงขึ้น 1 หน่วยจะมีผลให้ OR ลดลงร้อยละ 15 (คือ 100*0.85-100 = 15) ซึ่งก็คือโอกาสที่จะประสบผลสำเร็จลดลงร้อยละ 15

การประมาณค่าด้วย Excel Solver

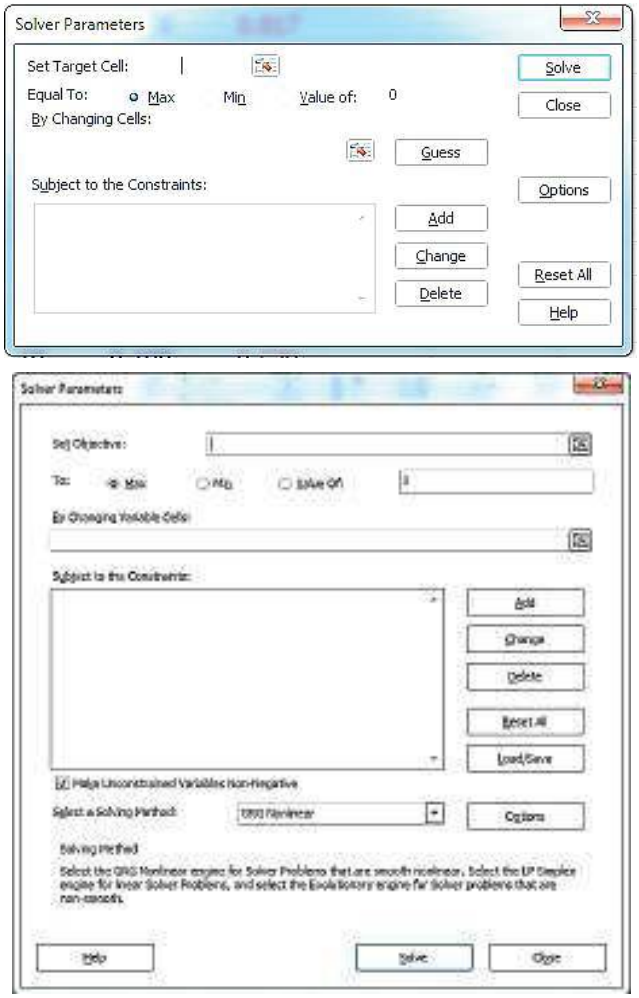
ในที่นี้จะแสดงการประมาณค่าพารามิเตอร์ด้วย Excel Solver โดยจะแสดงตัวอย่างให้ดู 1 ตัวอย่าง ก่อนอื่นให้เพิ่มโปรแกรม solver เข้ามาในเมนูของเอกเซลดังนี้ (Rodríguez, 2007; Zaiontz, 2016)

การเพิ่ม Solver

การเพิ่ม (add in) Solver ให้ไปที่เมนู file แล้วเรียกโปรแกรมเพิ่มดังนี้

File > option > Add-in > go แล้วเลือกโปรแกรมที่ต้องการในที่นี้คือ Solver ซึ่งจะมีโปรแกรม Solver เพิ่มเข้ามาในเมนู Data พร้อมให้เรียกใช้การเรียกใช้กระทำได้ดังนี้คือ

Data > Solver จะปรากฏไดอะล็อกดังภาพที่ 1



ภาพที่ 1 ไดอะล็อกของ Solver ใน Excel 7 (ภาพบน) และ Excel 10 (ภาพล่าง)

การตอบไดอะล็อกให้เตรียมข้อมูลพารามิเตอร์เพื่อหาค่าสูงสุดหรือค่าต่ำสุดหรือค่าตามกำหนด (target value) ค่านี้เรียกว่า Solver objectiveและต้องกำหนดค่าของพารามิเตอร์เป็นค่าลำลอง (ค่าคำตอบชั่วคราว) ซึ่งกำหนดได้เองตามความเหมาะสมเพื่อใช้คำนวณหาค่า Solver objective เรียกว่า Decision variable ในที่นี้คือ ค่าของ β_0 , β_1, β_2, ..., β_k คำว่าสามารถกำหนดได้ตามความเหมาะสมหมายถึงผู้ใช้สามารถกำหนดค่าได้เองตามใจจะกำหนดเท่าไรก็ได้ แต่ขอแนะนำให้ใช้ค่าตัวเลขง่ายเพื่อประโยชน์ในการตรวจสอบคำตอบก่อนสั่ง Solve สำหรับไดอะล็อกของ Solver ใน Excel 10 จะมีตัวเลือกหาคำตอบแบบ nonlinear คือ Generalized Reduced Gradient (GRG) การใช้ Solver กรณีของ Binary Logistic Regression ไม่ต้องมี constraint จึงไม่ต้องตอบไดอะล็อกเรื่องเงื่อนไข (constraint) และไม่ต้องเช็คเครื่องหมายถูกที่เชคบ๊อกซ์หน้า Make Unconstraint Variable Non-Negative ตัวอย่างการใช้งาน Solverปรากฏดังนี้

ตัวอย่าง จากเครื่องจักรจำนวน 20 เครื่องผู้บริหารฝ่ายการผลิตสนใจศึกษาว่าเครื่องจักรจะยังทำการผลิตสินค้าได้ตรงตามข้อกำหนด คือ ผิดพลาดไม่เกินร้อยละ 5 หรือไม่ ข้อมูลที่จดบันทึกจากเครื่องจักรแต่ละเครื่อง คือ ผลการผลิต (1=ผลิตตรงตามข้อกำหนด 0 =ผลิตไม่ตรงตามข้อกำหนด) หากถ้าไม่ตรงตามข้อกำหนดผู้บริหารการผลิตจะได้เรียกผู้ขายมาบำรุงรักษา ปัจจัยที่เป็นตัวกำหนด คือ อายุเครื่องจักร (สัปดาห์) และจำนวนกะการทำงาน (1 กะ = 3 ชั่วโมง) โดยเป็นที่เข้าใจกันโดยทั่วไปว่าเครื่องจักรเก่าและใช้งานผลิตต่อเนื่องนานหลายกะน่าจะก่อให้เกิดความผิดเพี้ยน คือ ผลผลิตคลาดเคลื่อนจากข้อกำหนด ข้อมูลผลการทำงานเครื่องจักรปรากฏดังตารางที่ 1

ตารางที่ 1 ข้อมูลผลการทำงานของเครื่องจักร อายุเครื่องจักร และความถี่การใช้งาน

ผลการทำงานผลิต	อายุเครื่องจักร (สัปดาห์)	จำนวนกะต่อสัปดาห์
0	73	5
0	59	4
0	68	5
0	49	5
0	27	7
0	78	8
0	57	7
0	73	8
0	71	7
0	35	4
1	57	4
1	22	5
1	15	4
1	36	2
1	59	3
1	10	6
1	22	6
1	36	4
1	38	5
1	44	5

การวิเคราะห์ให้ดำเนินการดังนี้

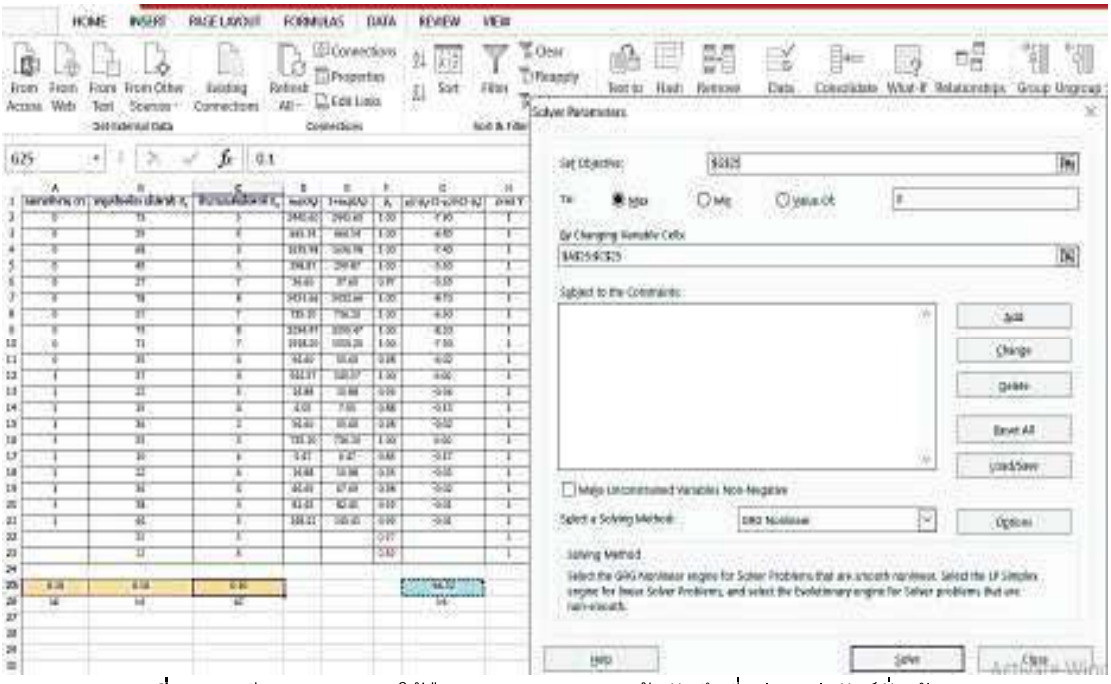
1. คำนวณหาค่า P_iตามสมการ (6) โดยที่

$$P_i=\frac{e^{\hat{\beta}_0+\hat{\beta}_1 X_{1i}+\hat{\beta}_2 X_{2i}+\cdots+\hat{\beta}_k X_{ki}}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X_{1i}+\hat{\beta}_2 X_{2i}+\cdots+\hat{\beta}_k X_{ki}}};i = 1, 2, 3, ..., 20$$

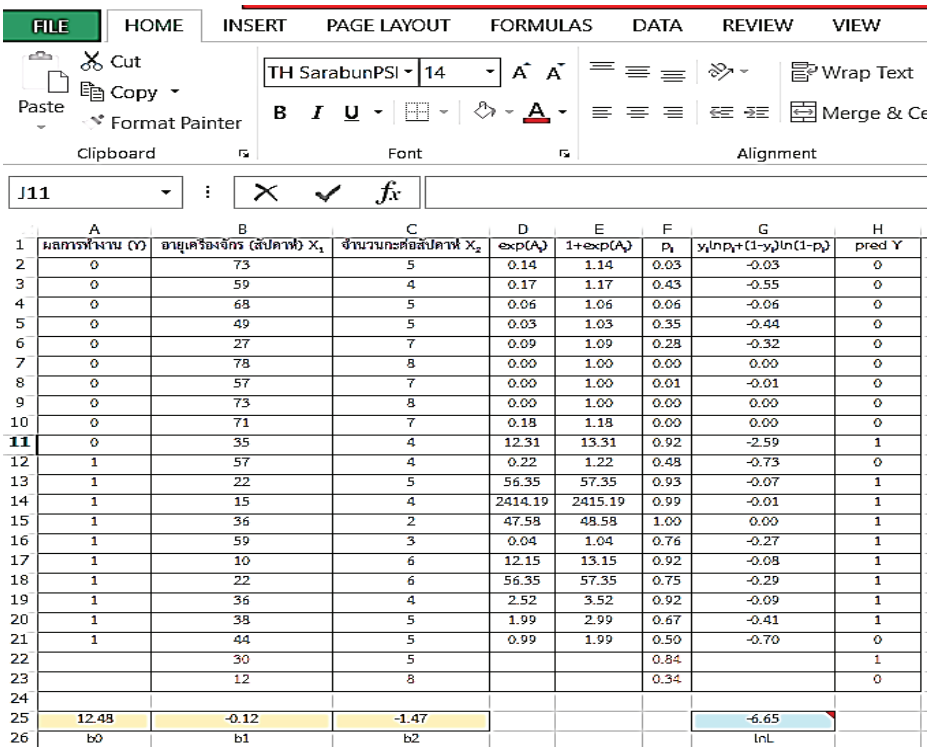
ในที่นี้ n = 20 กำหนดเซลไว้ 3 เซลสำหรับแสดงค่า decision variable คือ β_0 ,β_1 และ β_2 โดยในตอนแรกให้กำหนดค่าให้แก่ β_0 ,β_1และβ_2 เป็นค่าใด ๆ ก็ได้ เรียกว่าค่าลำลอง ค่านี้เอกเซลจะใช้สำหรับเริ่มต้นกระบวนการวนเวียนค้นหาคำตอบซึ่งหากสามารถหาคำตอบได้แล้วค่านี้ก็จะเปลี่ยนไปเพราะถูกแทนที่ด้วยค่าจากการคำนวณดังกล่าว ในที่นี้ขอกำหนดเป็นค่าง่าย ๆ และมีค่าเท่ากันคือ β_0= β_1=β_2= 0.10 สังเกตว่าผู้เขียนกำหนดด้วยค่ากลม ทั้งนี้เพื่อสังเกตง่ายว่าผลเปลี่ยนแปลงไปเพียงใด และหากประสงค์จะตรวจสอบคำตอบด้วยการคำนวณเอง (สำหรับบางเรื่องที่คำนวณไม่ยาก) ก็สามารถกระทำได้

2. คำนวณหาค่า lnL= $\sum_i^n Y_i \ln P_i + \sum_i^n (1-Y_i)\ln(1-P_i)$ ยอดรวมนี้เรียกว่า solver objectiveโดยให้เตรียมที่ไว้อีก 1 เซลสำหรับแสดงค่ายอดรวมตามสูตรในสมการ (7)

3. สั่ง solveการตอบไดอะล็อกต้องเลือก maximize เพราะเป็นการหาค่าตามเกณฑ์ของ maximum likelihood estimation (MLE)



ภาพที่ 2 การเรียก Solver มาใช้คือ Data > Solver แล้วบันทึกที่อยู่ของค่าฟังก์ชั่นเป้าหมาย และที่อยู่ของตัวแปรตัดสินใจลงในเมนูโต้ตอบ



	A	B	C	D	E	F	G	H
1	ผลการทำงาน (Y)	อายุเครื่องจักร (สัปดาห์) X_1	จำนวนกะต่อสัปดาห์ X_2	$\exp(A_i)$	$1+\exp(A_i)$	p_i	$y_i\ln p_i+(1-y_i)\ln(1-p_i)$	pred Y
2	0	73	5	0.14	1.14	0.03	-0.03	0
3	0	59	4	0.17	1.17	0.43	-0.55	0
4	0	68	5	0.06	1.06	0.06	-0.06	0
5	0	49	5	0.03	1.03	0.35	-0.44	0
6	0	27	7	0.09	1.09	0.28	-0.32	0
7	0	78	8	0.00	1.00	0.00	0.00	0
8	0	57	7	0.00	1.00	0.01	-0.01	0
9	0	73	8	0.00	1.00	0.00	0.00	0
10	0	71	7	0.18	1.18	0.00	0.00	0
11	0	35	4	12.31	13.31	0.92	-2.59	1
12	1	57	4	0.22	1.22	0.48	-0.73	0
13	1	22	5	56.35	57.35	0.93	-0.07	1
14	1	15	4	2414.19	2415.19	0.99	-0.01	1
15	1	36	2	47.58	48.58	1.00	0.00	1
16	1	59	3	0.04	1.04	0.76	-0.27	1
17	1	10	6	12.15	13.15	0.92	-0.08	1
18	1	22	6	56.35	57.35	0.75	-0.29	1
19	1	36	4	2.52	3.52	0.92	-0.09	1
20	1	38	5	1.99	2.99	0.67	-0.41	1
21	1	44	5	0.99	1.99	0.50	-0.70	0
22		30	5			0.84		1
23		12	8			0.34		0
24								
25	12.48	-0.12	-1.47				-6.65	
26	b0	b1	b2				lnL	

ภาพที่ 3 ผลการคำนวณหาค่า P ค่าสมมุติของตัวแปรตัดสินใจ และค่า log likelihoodและค่าพยากรณ์

หมายเหตุ ถ้าประสงค์จะใช้ Solver คำนวณหาค่าของสัมประสิทธิ์การถดถอยของสมการถดถอยพหุก็กระทำได้ในทำนองเดียวกัน ต่างกันที่สมการเป้าหมาย คือในกรณี MRA จะใช้วิธี OLS คือการ minimize $\sum_i^n e_i^2$ โดยที่ $e_i = Y_i - \hat{Y}_i$และ $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_j(X_{ji} + 1) + \cdots + \hat{\beta}_k X_{ki}$; i = 1, 2, ... , n

ผลการรันด้วยการหาค่าสูงที่สุดของฟังก์ชั่นเป้าหมายเพื่อเปลี่ยนแปลงค่าตัวแปรตัดสินใจพบว่าค่าสูงที่สุดคือ -6.65 ค่านี้เป็นค่าที่เกิดจากการปรับค่าตัวแปรตัดสินใจจากที่กำหนดไว้เดิมตามความสะดวก b_o= 0.1, b_1= 0.1, b_2= 0.1 เป็นค่าใหม่คือ b_o= 12.48, b_1= -0.117, b_2= -1.469 ทำให้ได้รับค่าความน่าจะเป็นที่เครื่องจักรจะผลิตสินค้าได้ตรงตามข้อกำหนดและและได้ odd ratio ดังต่อไปนี้

$$Pr(Y_i = 1) = P_i = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X_{1i}+\hat{\beta}_2 X_{2i}}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X_{1i}+\hat{\beta}_2 X_{2i}}}; i = 1, 2, 3, \ldots, n$$

$$= \frac{e^{12.48-0.117\, X_{1i}-1.469\, X_{2i}}}{1+e^{12.48-0.117\, X_{1i}-1.69\, X_{2i}}}; i = 1, 2, 3, \ldots, n$$

$$e^{\hat{\beta}_1} = e^{-0.117} = 0.89 \text{และ } e^{\hat{\beta}_2} = e^{-1.469} = 0.230$$

ค่า odd ratio =0.89 หมายความว่าถ้าอายุการทำงานของเครื่องจักรเพิ่มขึ้น 1 หน่วยสัปดาห์จะมีผลให้เครื่องจักรผลิตสินค้าผลิตสินค้าผิดข้อกำหนดเพิ่มขึ้นร้อยละ 11 (คือ 1.00-0.89) และ odd ratio =0.230 หมายความว่าถ้าเครื่องจักรทำงานเพิ่มขึ้นสัปดาห์ละ 1 กะจะมีผลให้เครื่องจักรผลิตสินค้าผลิตสินค้าผิดข้อกำหนดเพิ่มขึ้นร้อยละร้อยละ 77

การพยากรณ์ความน่าจะเป็น Pr(Y=1) ปรากฏดังตารางที่ 2 เมื่อเปรียบเทียบค่าความน่าจะเป็นกับเกณฑ์ตัดสินใจ (cut point) โดยปกติจะกำหนดให้มีค่าเท่ากับ 0.50 โดยที่ถ้าค่าความน่าจะเป็นต่ำกว่า 0.50 ให้ถือว่า Y=0 ถ้าค่าความน่าจะเป็นมีค่าตั้งแต่ 0.50 เป็นต้นไปให้ถือว่า Y=1 ค่าพยากรณ์ของ Y และร้อยละของการพยากรณ์ถูกปรากฏดังนี้

ตารางที่ 2 ค่าพยากรณ์ความน่าจะเป็นที่เครื่องจักรจะทำงานไม่ตรงข้อกำหนด ค่าพยากรณ์ของ Y เทียบกับค่าจริงของ Y

predicted Pr(Y=1)	predicted Y	ค่า Yจริง
0.032	0	0
0.426	0	0
0.056	0	0
0.355	0	0
0.277	0	0
0.000	0	0
0.011	0	0
0.000	0	0
0.002	0	0
0.925	1	0
0.484	0	1
0.928	1	1
0.992	1	1
0.995	1	1
0.763	1	1
0.924	1	1
0.749	1	1
0.916	1	1
0.666	1	1
0.497	0	1

ตารางที่ 3 ผลการพยากรณ์และคุณภาพโดยรวมของสมการพยากรณ์

	พยากรณ์เป็น 0	พยากรณ์เป็น 1
ค่าจริงคือ 0	9	1
ค่าจริงคือ 1	2	8
ร้อยละพยากรณ์ถูก	85	
ร้อยละพยากรณ์ผิด	15	

การนำสมการไปใช้ประโยชน์

1. ด้านการพยากรณ์ การพยากรณ์กระทำได้โดยการแทนที่ค่าตัวแปรอิสระในสมการ (6) ด้วยค่าที่สนใจ เช่นแทนค่า X_1 และ X_2 ในสมการพยากรณ์ $Pr(Y_i=1) = P_i = \frac{e^{\hat{\beta}_0+\hat{\beta}_1X_{1i}+\hat{\beta}_2X_{2i}}}{1+e^{\hat{\beta}_0+\hat{\beta}_1X_{1i}+\hat{\beta}_2X_{2i}}}$;i = 1, 2, 3, ..., n ด้วยค่าที่กำหนดตามความสนใจเช่น สนใจว่าเครื่องจักรอายุ 10 สัปดาห์ (X_1 = 10) ทำงานสัปดาห์ละ 5 กะ (X_2 = 5) จะมีโอกาสผลิตสินค้าได้ตรงข้อกำหนดเท่ากับเท่าไร พบว่าสมการสามารถพยากรณ์โอกาสที่เครื่องจักรจะทำการผลิตได้ตรงข้อกำหนดคือ

$Pr(Y_i=1) = P_i = \frac{e^{12.48-0.117*10-1.469*5}}{1+e^{12.48-0.117*10-1.469*5}}$

= 0.981ซึ่งพบว่าเป็นค่าความน่าจะเป็นที่มีค่าสูงกว่า 0.50 จึงสรุปว่า Y = 1

2. ด้านการระบุตัวกำหนด การระบุว่าตัวแปรใดเป็นตัวกำหนด (determinants) ของโอกาสที่จะเกิดเหตุการณ์ที่สนใจพิจารณาได้จาก

1) นัยสำคัญของสัมประสิทธิ์การถดถอย ถ้าสัมประสิทธิ์การถดถอยของตัวแปร X_j มีนัยสำคัญแสดงว่า X_j เป็นตัวกำหนดความน่าจะเป็นที่จะเกิดเหตุการณ์คือ Pr(event)

สำหรับการทดสอบสมมุติฐานเราจำเป็นต้องจัดข้อมูลเป็นกลุ่มตามตัวแปรกลุ่มทำให้ได้ข้อมูลเป็นกลุ่มๆ รวม k กลุ่ม ขนาดกลุ่มคือ n_i ; i = 1,2,..., k ดังนั้น $w_{ii} = n_i\hat{P}_i\hat{Q}_i$โดยที่ $\hat{Q}_i = 1 - \hat{P}_i$คือ

$$W = \begin{bmatrix} n_1\hat{P}_1\hat{Q}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & n_k\hat{P}_k\hat{Q}_k \end{bmatrix}$$

และ $\hat{V}(\hat{\beta}) = (X'WX)^{-1}$ที่เมื่อตัวอย่างมีขนาดใหญ่เราสามารถทดสอบสมมุติฐาน Ho: β_j = 0 ด้วย z = $\frac{\hat{\beta}_j}{\sqrt{\hat{V}(\hat{\beta}_j)}}$

2) odd ratio ของตัวแปรอิสระ คือ e^{β_j} ถ้าตัวแปรใดมีค่า odd ratioสูงกว่าแสดงว่าตัวแปรนั้นมีอิทธิพลต่อ Pr(event)มากกว่า

สรุป

การถดถอยลอจิสติก คือ กรณีหนึ่งของการถดถอยพหุที่ตัวแปรตามเป็นตัวแปรกลุ่ม ถ้าตัวแปรตามมี 2 กลุ่ม คือ มีค่า 2 ค่าคือ 0 กับ 1 เรียกว่า Binary logistic regression ถ้าตัวแปรตามมีมากกว่า 2 กลุ่มคือค่ามากกว่า 2 ค่า คือ 0, 1, 2, ..., r เรียกว่า Multinomial logistic regression การที่มีชื่อว่า logistic ปนอยู่เนื่องจากการพัฒนาทฤษฎีอาศัย logistic distribution เป็นฟังก์ชันช่วยแปลงค่า $\hat{\beta}_0 + \hat{\beta}_1X_{1i} + \hat{\beta}_2X_{2i} + \cdots + \hat{\beta}_kX_{ki}$มาเป็น Pr(Y = 1) แทนที่จะเป็น Y โดยตรง หลังจากนั้นจึงค่อยเปลี่ยนค่าความน่าจะเป็นมาเป็นค่า Y ด้วยเกณฑ์ตัดสินใจ

การถดถอยลอจิสติกถูกนำไปใช้ในหลายบริบททั้งในทางสังคมศาสตร์ วิทยาศาสตร์สุขภาพ การเงินการบัญชี และอื่น ๆ โดยมีทั้งการนำไปใช้กับกรณีของข้อมูลตามเวลา (time series) และกรณีของการศึกษากลุ่มเป็นระยะเวลายาว (Longitudinal Studies, panel analysis)

การวิเคราะห์ข้อมูลอาจใช้ซอฟท์แวร์ทางสถิติได้มากมาย เช่น SPSS SAS EVIEWSMINITAB XLSTAT ซึ่งนักวิจัยสามารถเลือกใช้ตัวใดก็ได้เพราะได้ผลตรงกัน ในที่นี้ผู้เขียนแนะนำการวิเคราะห์ด้วยเอกเซล เพราะเห็นว่าอาจมีความจำเป็นต้องพัฒนาระบบสนับสนุนการตัดสินใจที่ทำงานด้วย excel VBA จะได้เรียกใช้ส่งต่อข้อมูลกันง่ายอีกทั้งยังเป็นการแสดงศักยภาพของผู้ใช้ด้วยว่าสามารถเข้าใจเรื่องราวของการถดถอยลอจิสติกในทางลึก

เอกสารอ้างอิง

มนตรี พิริยะกุล. (2544). **เทคนิคการวิเคราะห์สมการถดถอย**. กรุงเทพฯ: มหาวิทยาลัยรามคำแหง.

Hosmer, Jr., D. W. , Lemeshow, S. , & Sturdivant, R. X. (2013). **Applied Logistic Regression**. (3rd ed.). New York: Wiley.

Kleinbaum, D. G. , & Mitchel ,K. M. (2010). **Logistic Regression: A Self-Learning Text**. New York: Springer.

Menard, S. W. (2001). **Applied logistic regression analysis (quantitative applications in the social sciences)** (2nd ed.). Thousand Oaks, CA: Sage Publications.

Rodríguez, G. (2007). **Lecture Notes on Generalized Linear Models**. [Online] Available: http:// data.princeton.edu/wws509/notes/. [2016, April 20].

Zaiontz, C. (2016). **Finding Logistic Regression Coefficients using Excel's Solver**. [Online] Available: http://www.real-statistics.com/logistic-regression/finding-logistic-regression-coefficients-using-excels-solver/ [2016, March 12].