การเปรียบเทียบตัวแบบลอจิตและการใช้ ROC Curve ในการวิเคราะห์ปัจจัยที่ส่งผลต่อระดับค่าจ้าง
A Comparison of Logit Models and Use of ROC Curve in the Analysis of Factors Affecting Wages' Levels

กันยาพร หาญกล้า และ วีรานันท์ พงศาภักดี*
ภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยศิลปากร
Kanyaphorn Hankla and Veeranun Pongsapukdee*
Department of Statistics, Faculty of Science, Silpakorn University, Snam-Chandra Campus

บทคัดย่อ

ความหลากหลายและระดับของค่าจ้างที่ลูกจ้างได้รับนั้น อาจเป็นผลมาจากปัจจัยต่างๆ ที่ส่งผลกระทบต่อระดับค่าจ้าง งานวิจัยนี้ ศึกษาเปรียบเทียบตัวแบบด้วยวิธีการวิเคราะห์ทางสถิติต่างๆ เพื่อหาตัวแบบที่เหมาะสม โดยใช้ตัวอย่างของข้อมูลจริงชุดหนึ่งเกี่ยวกับระดับค่าจ้างและปัจจัยต่างๆ ที่อาจส่งผลต่อระดับค่าจ้าง ด้วยการสร้างตัวแบบลอจิตสองกลุ่ม (Dichotomous logit models) เมื่อตัวแปรตอบสนองมี 2 ระดับ (สูงและต่ำ) และตัวแบบลอจิตสะสม (Cumulative logit models) กรณี Proportional odds models เมื่อตัวแปรตอบสนองมี 3 ระดับ (สูง ปานกลาง และต่ำ) โดยตัวแปรอธิบายหรือปัจจัยที่เกี่ยวข้องต่างๆ ของลูกจ้าง เช่น การศึกษา อาชีพ ประสบการณ์การทำงาน การเปรียบเทียบตัวแบบพิจารณาจาก ตัวสถิติวาลด์ (Wald), AIC (Akaike's Information Criterion), SC (Schwarz Criterion), อัตราส่วนภาวะน่าจะเป็น (G^2, -2logL) และการตรวจสอบการพยากรณ์ของตัวแบบให้ถูกต้องด้วยเส้นโค้ง ROC (Receiver Operating Characteristic) ข้อมูลจริงจากเว็บไซต์ http://lib.stat.cmu.edu เป็นตัวแทนจากการสำรวจตัวอย่างลูกจ้าง จำนวน 534 คน ผลการวิจัยพบว่า ปัจจัยที่ส่งผลต่อระดับค่าจ้างมี 6 ปัจจัย คือ การศึกษา ประสบการณ์การทำงานอาชีพ แผนก เพศ และ การเข้าร่วมเป็นสหภาพแรงงาน อย่างมีนัยสำคัญทางสถิติที่ 0.05 โดยตัวแบบที่เหมาะสมกับข้อมูลมากกว่าคือ ตัวแบบลอจิตสองกลุ่ม ซึ่งให้ค่าสถิติ AIC, SC และอัตราส่วนภาวะน่าจะเป็นที่ดีกว่าตัวแบบอื่นที่ใช้ และมีร้อยละการจำแนกกลุ่มได้ถูกต้องในภาพรวมร้อยละ 73.6 ภายใต้ ROC Curve ซึ่งแสดงความไวได้ค่อนข้างสูง รวมทั้งการตรวจสอบส่วนเหลือพบว่า ตัวแบบลอจิตสองกลุ่มให้ค่าส่วนเหลือที่มีการ กระจายตัวได้ดีกว่า ดังนั้นตัวแบบจึงมีประโยชน์และอาจนำไปพยากรณ์ระดับค่าจ้างและตรวจสอบปัจจัยที่ส่งผลกระทบต่อระดับค่าจ้างต่อไป

คำสำคัญ : ตัวแบบลอจิตสองกลุ่ม ROC Curve ตัวแบบ Proportional odds ระดับค่าจ้าง

***Corresponding author.** E-mail: veeranun@su.ac.th

Abstract

Different employee wages' levels are probably associated to various factors. This research aims to compare two logit models that are demonstrated through the analysis of factors affecting wage levels under several goodness-of-fit statistics such that Wald, AIC (Akaike's Information Criterion), SC (Schwarz Criterion), likelihood ratio statistics (G^2, -2logL) and that the use of ROC (Receiver Operating Characteristic) curve. The analysis models are the logit model for dichotomous response categories: high and low wages and the proportional odds model for ordinal trichotomous response categories: high, medium and low wages. All real data are from website http://lib.stat.cmu.edu. These data consist of a random sample of 534 persons. The response variable is wage and the explanatory variables include several characteristics of the workers. The research results show that the effective factors at the 0.05 level of significance are the number of years of education, work experience, occupational status, sex and region of residence, sector, and union membership. The dichotomous logit model gives better fitss due to AIC, SC and Likelihood ratio estimates compared to those of proportional odds model. In conclusion, the dichotomous logit model has an adequate of fit with the overall percent correct classification of 73.6%, or a high sensitivity under the ROC curve. Moreover, the residual plots of the dichotomous logit model are more spread out than others. Therefore, it may be a very helpful tool to predict the employee wages levels and that to assess the essential factors affecting the wages levels.

Keyword : Dichotomous Logit model, Proportional odds model, Wage levels, ROC Curve

บทนำ

ปัจจุบันตัวแบบลอจิต (Logit models) สำหรับตัวแปรตอบสนองจำแนกประเภท 2 ระดับ และตัวแบบ Proportional odds models (McCullagh, 1980) สำหรับตัวแปรตอบสนองอันดับตั้งแต่ 3 ระดับขึ้นไป ซึ่งเป็นตัวแบบในกลุ่มของตัวแบบลอจิตสะสม (Cumulative logit models) ได้รับความสนใจและนำไปใช้อย่างกว้างขวาง เช่น ด้านการแพทย์ ด้านความเสี่ยง ด้านสถานภาพของลูกจ้าง ซึ่งรวมถึงลูกจ้างในภาคธุรกิจ รัฐวิสาหกิจ และสหภาพแรงงาน งานวิจัยนี้สนใจศึกษาเปรียบเทียบตัวแบบเชิงสถิติทั้งสองแบบข้างต้น ด้วยวิธีการวิเคราะห์ทางสถิติภายใต้ตัวสถิติต่างๆ เพื่อหาตัวแบบที่เหมาะสม โดยใช้ตัวอย่างของข้อมูลจริงชุดหนึ่งเกี่ยวกับระดับค่าจ้างและปัจจัยต่างๆ ที่อาจส่งผลต่อระดับค่าจ้าง เป็นกรณีศึกษา และศึกษาปัจจัยที่ส่งผลกระทบต่อระดับค่าจ้างอย่างมีนัยสำคัญทางสถิติ การตรวจสอบและเปรียบเทียบตัวแบบอาศัยตัวสถิติภาวะสารูปดีต่างๆ เช่น วาลด์ (Wald), AIC (Akaike's Information Criterion), SC (Schwarz Criterion), อัตราส่วนภาวะน่าจะเป็น (G^2, -2logL) และการพล็อตส่วนเหลือ โดยเฉพาะมีการตรวจสอบการพยากรณ์ได้ถูกต้องของตัวแบบด้วยสัดส่วนการพยากรณ์ถูกต้อง ภายใต้เส้นโค้ง ROC (Receiver Operating Characteristic) หรือที่เรียกทั่วไปว่า ROC Curve ซึ่งใช้สำหรับการวิเคราะห์ความไว (Sensitivity) ของตัวแบบในการพยากรณ์ได้ถูกต้อง สรุปตัวแบบที่ใช้ในการวิจัยดังต่อไปนี้

ตัวแบบลอจิตสองกลุ่ม (Dichotomous logit model)

การวิเคราะห์ข้อมูลสำหรับตัวแปรตอบสนองเชิงกลุ่ม 2 กลุ่ม อาศัยฟังก์ชันที่แปลงค่าเฉลี่ยของตัวแปรตอบสนอง Y แบบสองกลุ่ม หรือค่าความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจคือ $P_i(x)$ เมื่อกำหนดค่าของตัวแปร X โดยตัวแบบลอจิตที่อยู่ในรูปแบบการแปลงลอจิต (logit transformation) (Agresti, 1990; วีรานันท์ พงศาภักดี, 2544, 2555) ซึ่งมีรูปแบบดังนี้

$$Q(x) = \log\left[\frac{P(x)}{1 - P(x)}\right] = \beta_0 + \boldsymbol{\beta}'\boldsymbol{x} \qquad (1)$$

เมื่อ $Q(x)$ แทนฟังก์ชันเชิงเส้นในเทอมของพารามิเตอร์ ซึ่งมีค่าต่อเนื่องหรืออาจมีค่าช่วงจาก $-\infty$ ถึง ∞ โดยขึ้นอยู่กับค่าของเมทริกซ์ X, $\left[\frac{P(x)}{1 - P(x)}\right]$ แทน อัตราส่วน Odds ซึ่งหมายถึงอัตราส่วนระหว่างความน่าจะเป็นที่จะเกิดเหตุการณ์ที่สนใจกับความน่าจะเป็นที่ไม่เกิดเหตุการณ์ที่สนใจ และ $\log\left[\frac{P(x)}{1 - P(x)}\right]$ แทนลอจิตของ $P(x)$ และสมการ (1) ข้างต้นเรียกว่าตัวแบบลอจิตสองกลุ่ม (Dichotomous logit model) ซึ่งเป็นฟังก์ชันเชิงเส้นแสดงความเกี่ยวพันระหว่างตัวแปรอธิบายต่างๆ กับตัวแปรตอบสนองที่มีค่า สองค่า เช่น ค่าจ้างระดับสูง (1) และระดับต่ำ (2)

ตัวแบบลอจิตสะสมภายใต้ตัวแบบ Proportional odds models

การวิเคราะห์ข้อมูลที่มีตัวแปรตอบสนองจำแนกประเภทมีลำดับ เมื่อตัวแปรตอบสนองมีมากกว่า 2 กลุ่ม สามารถสร้างจากฟังก์ชันที่เรียกว่า Cumulative logits ภายใต้ตัวแบบ Proportional odds models (McCullagh, 1980) ความเกี่ยวพันระหว่างตัวแปรอธิบาย $X = (X_1, X_2, ..., X_p)'$ กับความน่าจะเป็นสะสมของการแจกแจงของ Y ภายใต้ตัวแบบลอจิตสะสม ในรูปแบบดังนี้

$$\text{logit}\left[P(Y \le j)\,|x\right] = \alpha_j + \boldsymbol{\beta}'\boldsymbol{x},\ j = 1, ..., J - 1. \qquad (2)$$

เมื่อ $\alpha_1 \le \alpha_2 \le ... \le \alpha_{J-1}$ และจุดตัด $(\alpha_j : j = 1, ..., J - 1)$ คือ log odds ของ Y ที่น้อยกว่าหรือเท่ากับ j เมื่อเวกเตอร์ x มีค่าเป็นศูนย์ นั่นคือ $P(Y \le j) = \frac{e^{\alpha_j}}{1 + e^{\alpha_j}}$ เมื่อ x มีค่าเป็นศูนย์ เวกเตอร์สัมประสิทธิ์ $\boldsymbol{\beta}$ แทน log odds ของ Y ที่น้อยกว่าหรือเท่ากับ j เมื่อ X_i เปลี่ยนแปลงไป 1 หน่วย หรือเปลี่ยนจากกลุ่มหนึ่งเป็นอีกกลุ่มหนึ่งที่ใช้เปรียบเทียบ ขณะที่ตัวแปรอธิบายตัวอื่นๆ มีค่าคงที่ ทำให้ได้ว่า $\boldsymbol{\beta}$ ไม่ขึ้นกับ j ดังนั้นตัวแบบนี้มีข้อสมมติว่า ความเกี่ยวพันของ X และ Y นั้นเป็นอิสระกับ j หรือเรียกข้อสมมติของความเท่ากันของ log odds ratio ในทุกๆ จุดตัดดังกล่าวว่า Proportional odds assumption จึงเรียกตัวแบบนี้ว่า Proportional odds model (McCullagh, 1980)

การประมาณค่าพารามิเตอร์ $\boldsymbol{\beta}$ ด้วย $\hat{\boldsymbol{\beta}}$ โดยใช้วิธีฟิชเชอร์-สกอริง (Fisher-scoring method) ในการแก้สมการหาตัวประมาณภาวะน่าจะเป็นสูงสุด (maximum likelihood method) และประมาณค่า odds ratio ได้จาก $e^{\hat{\beta}}$

การทดสอบสมมติฐาน $H_0 : \beta_i = 0$ คู่กับ $H_1 : \beta_i \ne 0$ จะปฏิเสธสมมติฐานว่างถ้าค่า P- value น้อยกว่าระดับนัยสำคัญภายใต้ตัวสถิติ Wald (W) หรือตัวสถิติผลต่างของ Deviances ระหว่างตัวแบบที่มีและไม่มีพารามิเตอร์ β_i หมายความว่า ตัวแปรอธิบายนั้นจะมีความเกี่ยวพันกับตัวแปรตอบสนองแบบมีลำดับอย่างมีนัยสำคัญทางสถิติที่ระดับ α

เมื่อ $W = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$ ซึ่งมีการแจกแจงแบบปกติมาตรฐาน หรือ $W^2 = \left(\frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}\right)^2$ มีการแจกแจงใกล้เคียงแบบไค-สแควร์ ที่มี

องศาอิสระเท่ากับ 1 ภายใต้ $H_0 : \beta_j = 0$

ตัวสถิติทดสอบอัตราส่วนภาวะน่าจะเป็น หรือ ผลต่าง Deviances คือ -2 log L (null model) - [-2 log L (fitted or alternative model)] ซึ่งมีการแจกแจงแบบไคสแควร์ด้วยองศาอิสระเท่ากับ 1 เมื่อ log L แทนค่าของ log – likelihood (Agresti, 2002; ผึ้งพร ลาภส่งผล และคณะ, 2551)

การตรวจสอบความเหมาะสมของตัวแบบ ตัวสถิติที่ตรวจสอบความเหมาะสมหรือภาวะสารูปดีของตัวแบบ ได้แก่ AIC (Akaike's Information Criterion), SC (Schwarz Criterion) และ -2 Log L โดยที่ AIC = -2 log L + 2p เมื่อ p แทน จำนวนพารามิเตอร์ทั้งหมดในตัวแบบ ซึ่งในตัวแบบลอจิต p = k(s+1) โดยที่ในตัวแบบลอจิตสะสม p = k + s, k แทน จำนวนระดับของตัวแปรตอบสนอง และ s แทน จำนวนตัวแปรอธิบายของตัวแบบ และค่า AIC ที่น้อยกว่า หมายความว่า ตัวแบบมีความเหมาะสมสำหรับข้อมูลมากกว่า (Agresti, 2007)

ตัวสถิติ SC = -2 log L + p log n เมื่อ n คือขนาดตัวอย่าง ค่า SC ที่น้อยที่สุด หมายความว่า ตัวแบบมีความเหมาะสมสำหรับข้อมูลมากที่สุด (Schwarz, 1978)

ตัวสถิติ -2 Log L พิจารณาเป็น 2 กรณี ดังนี้

กรณีที่ 1 : ภาวะน่าจะเป็นของตัวแบบลอจิตสะสมและลอจิตสองกลุ่ม (Cumulative and Dichotomous logits)

$$L = \sum_{i=1}^{n}\left\{\sum_{j=1}^{J-1} y_{ij}\left(\alpha_j + \beta_j x\right) - log\left[1 + \sum_{j=1}^{J-1} exp\left(\alpha_j + \beta_j x\right)\right]\right\}$$

$$= \sum_{j=1}^{J-1}\left[\alpha_j\left(\sum_{i=1}^{n} y_{ij}\right) + \sum_{k=1}^{p}\beta_{jk}\left(\sum_{i=1}^{n} x_{ik}y_{ij}\right)\right] + \sum_{i=1}^{n} log\left[1 + \sum_{j=1}^{J-1} exp\left(\alpha_j + \beta_j x\right)\right]$$

กรณีที่ 2 : ภาวะน่าจะเป็นของตัวแบบ Proportional odds model

$$L = \prod_{i=1}^{n}\left[\prod_{j=1}^{J}\left(P\left(Y_i \le j|x_i\right) - P\left(Y \le j-1|x_i\right)\right)^{y_{ij}}\right]$$

$$= \prod_{i=1}^{n}\left\{\prod_{j=1}^{J}\left(\frac{\exp\left(\alpha_j + \beta' x_i\right)}{1 + \exp\left(\alpha_j + \beta' x_i\right)} - \frac{\exp\left(\alpha_{j-1} + \beta' x_i\right)}{1 + \exp\left(\alpha_{j-1} + \beta' x_i\right)}\right)^{y_{ij}}\right\}$$

การทดสอบตัวแบบที่มีความเหมาะสมกับข้อมูลมากที่สุด คือ ตัวแบบที่มีค่า -2 Log L น้อยที่สุด (Agresti, 2002; ประภัสสร มีสกุล และคณะ (2552).

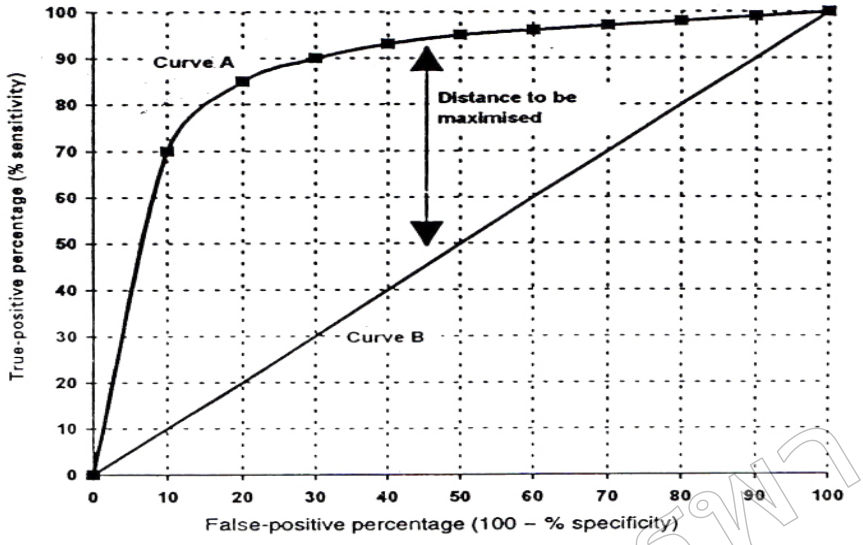
ประสิทธิภาพของการพยากรณ์ อาจวัดด้วยเทอม Sensitivity และ Specificity ภายใต้ตัวอย่างขนาด n โดยพิจารณาผลของการพยากรณ์ด้วยจำนวนเหตุการณ์ในตารางต่อไปนี้ (Agresti, 2007)

เหตุการณ์	ผลของการพยากรณ์		รวม
	ถูก ($\hat{y}$ = 1)	ไม่ถูก ($\hat{y}$ = 0)	
สนใจ (y = 1)	a	b	a+b
ไม่สนใจ (y = 0)	c	d	c+d
รวม	a+c	b+d	a+b + c+d

เมื่อ Sensitivity $= P(\hat{y} = 1|y = 1) = \frac{a}{a+b}$

Specificity $= P(\hat{y} = 0|y = 0) = \frac{d}{c+d}$

ผลของการพยากรณ์สามารถแสดงได้โดย ROC curve ซึ่งเป็นวิธีแสดงอำนาจการจำแนกที่สำคัญในการพยากรณ์ ณ จุดตัด (Cut point) ของตัวแปรตอบสนอง ถ้าเปลี่ยนจุดตัดมีผลให้ ค่า Sensitivity และ Specificity จะเปลี่ยนไปด้วย จุดตัดที่ดีที่สุดควรจะอยู่ที่จุดวกกลับของเส้นโค้ง ซึ่งยิ่งชันยิ่งดี เพราะจะให้พื้นที่ใต้เส้นโค้งด้านซ้ายของเส้นทแยงมุมหรือการพยากรณ์ถูกต้องเพิ่มขึ้น นอกจากนี้ การใช้ ROC curve ยังเป็นวิธีการที่ดีในการวัดอำนาจการพยากรณ์ของตัวแบบ Multiple logistic regression ด้วย เมื่อเพิ่มตัวแปรอธิบายที่มีความสำคัญเข้าในตัวแบบ เส้นโค้งจะชันมากขึ้นและมีพื้นที่การพยากรณ์ได้ถูกต้องใต้เส้นโค้งมากขึ้นด้วย ดังภาพที่ 1



ภาพที่ 1 ROC curve

จากภาพที่ 1 แสดงการวิเคราะห์ความไวของตัวแบบด้วย ROC curve โดยแกนตั้งแสดงค่า Sensitivity และแกนนอนแสดงค่า 1- Specificity ถ้าจุดตัดน้อยไปหรือมากไป จะมีผลต่อ Sensitivity และ Specificity ถ้า Sensitivity มากขึ้น Specificity จะลดลง และถ้า Sensitivity ลดลง Specificity จะเพิ่มขึ้น นอกจากนี้พื้นที่ภายใต้ ROC Curve ยังใช้เป็นตัววัดความสอดคล้องระหว่างค่าสังเกตกับค่าความน่าจะเป็นของการพยากรณ์ ที่เรียกว่า ดัชนีความสอดคล้อง (Concordant index) หรือ ตัวสถิติ c ในภาพที่ 1 เส้นในแนวตั้งที่ต้องการคือเส้นที่ให้ Curve A ห่างจากเส้นทแยงมุมมากที่สุด (Distance to be maximized) นั่นคือจุดที่แสดงว่าตัวแบบเหมาะสมคือจุดภายในเส้น Curve A ที่อยู่ใกล้จุด Sensitivity = 100% มากที่สุด

วิธีการวิจัย

ขั้นตอนวิธีการวิจัย แบ่งเป็น 2 ส่วนคือ ส่วนของข้อมูล และส่วนของเทคนิคการวิเคราะห์ข้อมูล คือ

ส่วนแรก เป็นส่วนของข้อมูลที่นำมาวิเคราะห์ เป็นข้อมูลจริงที่เผยแพร่จากเว็บไซต์ http://lib.stat.cmu.edu เรื่อง Determinants of Wages from the Current Population Survey ซึ่งสำรวจในปี ค.ศ. 1991 เป็นข้อมูลเกี่ยวกับค่าจ้าง และปัจจัยต่างๆ ที่เกี่ยวข้องอื่นๆ ของลูกจ้างจำนวน 534 คน ตัวแปรตอบสนอง (Wage) คือ ตัวแปรจำแนกตามกลุ่มของระดับค่าจ้างจำนวน 3 กลุ่ม ได้แก่ 1 (ค่าจ้างระดับสูง) 2 (ค่าจ้างระดับปานกลาง) และ 3 (ค่าจ้างระดับต่ำ) ตัวแปรอธิบายจำนวน 10 ตัวแปร ที่อาจส่งผลต่อตัวแปรตอบสนอง ได้แก่

1. Edu: จำนวนปีที่ศึกษา ได้แก่ น้อยกว่าหรือเท่ากับ 12 ปี และมากกว่า 12 ปี

2. South: พื้นที่พำนัก ได้แก่ อาศัยอยู่ภาคอื่นๆ และอาศัยอยู่ภาคใต้

3. Sex: เพศ ได้แก่ ชาย และหญิง

4. Exp: ประสบการณ์การทำงาน ได้แก่ น้อยกว่าหรือเท่ากับ 10 ปี, 11-20 ปี, 21-30 ปี และมากกว่า 31 ปีขึ้นไป

5. Union: การเข้าร่วมเป็นสมาชิกสหภาพแรงงาน ได้แก่ ไม่ได้ร่วมเป็นสมาชิกสหภาพแรงงาน และเข้าร่วมเป็นสมาชิกสหภาพแรงงาน

6. Age: อายุ ได้แก่ น้อยกว่าหรือเท่ากับ 20 ปี, 21-39 ปี, 40-60 ปี และ 61 ปีขึ้นไป

7. Race: เชื้อชาติ ได้แก่ อื่นๆ, ลาตินอเมริกัน และอเมริกัน

8. Occ: อาชีพ ได้แก่ นักบริหาร, พนักงานขาย, พนักงานธุรการ, พนักงานบริการ, ผู้เชี่ยวชาญ และอื่นๆ

9. Sector: แผนก ได้แก่ ฝ่ายผลิต ฝ่ายก่อสร้าง และอื่นๆ

10. Marital: สถานภาพการแต่งงาน ได้แก่ โสด และแต่งงาน

ส่วนหลังเป็นส่วนของเทคนิคการวิเคราะห์ข้อมูล ประกอบด้วยการวิเคราะห์ข้อมูลของตัวแปรตอบสนองเชิงกลุ่ม 2 กลุ่ม และการวิเคราะห์ข้อมูลของตัวแปรตอบสนองเชิงกลุ่ม 3 กลุ่ม โดยใช้ตัวแบบลอจิตสองกลุ่มและตัวแบบ Proportional odds model ตามลำดับ ตัวแปรอธิบายจำนวน 10 ตัวเป็น

แบบเชิงกลุ่ม ในการวิเคราะห์ข้อมูล วัดความเหมาะสมของตัวแบบด้วยตัวสถิติ AIC, SC และ -2 log L หรือ G^2 และแผนภาพการกระจายของส่วนเหลือ เพื่อหาตัวแบบที่เหมาะสมมากที่สุด โดยตัวแบบที่เหมาะสมจะเป็นตัวแบบที่มีประสิทธิภาพในการพยากรณ์กลุ่มได้ถูกต้องมากด้วย หรือมีความไวสูงภายใต้ ROC Curve

ผลการวิจัย

ผลการตรวจสอบตัวแบบ Proportional odds model ว่ามีความสอดคล้องกับข้อสมมติของตัวแบบหรือไม่ พบว่าข้อมูลในการศึกษาปัจจัยที่มีผลกระทบต่อระดับค่าจ้างเป็นไปตามข้อสมมติของตัวแบบ Proportional odds model ที่ระดับนัยสำคัญทางสถิติ 0.05 (ตัวสถิติไคสแควร์ 18.7213, p-value = 0.2834) และผลการทดสอบภาวะสารูปดีของตัวแบบ พบว่า ตัวแบบมีภาวะสารูปดีกับข้อมูลระดับค่าจ้างภายใต้ตัวแปรอธิบายต่างๆ ที่ระดับนัยสำคัญทางสถิติที่ 0.05 (p-value = 0.9784) ส่วนค่าของตัวสถิติ AIC, SC และ -2 log L หรือ G^2 เท่ากับ 688.305, 752.511 และ 658.305 ตามลำดับ ดังนั้นตัวแบบมีความเหมาะสมกับข้อมูล ณ α = 0.05 โดยสอดคล้องกับการทดสอบพารามิเตอร์ (β = 0) ของตัวแบบ Proportional odds model ด้วยตัวสถิติ Likelihood Ratio, Score, Wald ดังตารางที่ 1

ตารางที่ 1 ตัวสถิติ Likelihood Ratio, Score และ Wald

Test Statistics	Chi-Square	DF	Pr-ChiSq
Likelihood Ratio	179.4031	13	<0.0001
Score	157.8922	13	<0.0001
Wald	129.6448	13	<0.0001

จากตารางที่ 1 แสดงผลของการทดสอบด้วยตัวสถิติ Likelihood Ratio, Score และ Wald ในตาราง 1 ให้ผลลัพธ์เหมือนกัน นั่นคือ ชี้ให้เห็นว่าค่าสัมประสิทธิ์อย่างน้อยหนึ่งตัวที่ไม่เท่ากับศูนย์อย่างมีนัยสำคัญที่ 0.05 (p-value < 0.0001) และสามารถสรุปว่า ตัวแปรที่เลือกจากตัวแปรทั้งหมดด้วยวิธี Stepwise ซึ่งประกอบด้วย ตัวแปร Edu, Exp, Occ, Sex, Sector และ Union จากตัวแบบนั้น มีความเหมาะสมที่จะนำไปใช้ในการพยากรณ์ตัวแปรตอบสนองเชิงกลุ่มแบบมีลำดับ 3 กลุ่ม อย่างมีนัยสำคัญทางสถิติที่ระดับ 0.05 (p-value < 0.05) ดังตารางที่ 2

ตารางที่ 2 การทดสอบเกี่ยวกับค่าประมาณสัมประสิทธิ์ในตัวแบบ Proportional odds model

Effect	DF	Wald Chi-Square	Pr-ChiSq
Edu	1	26.6136	<0.0001
Exp	3	24.9276	<0.0001
Occ	5	45.9287	<0.0001
Sector	2	7.3722	0.0029
Sex	1	8.8694	0.0029
Union	1	15.2399	<0.0001

จากตารางที่ 2 แสดงค่าประมาณของพารามิเตอร์ในตัวแบบแต่ละค่าที่สอดคล้องกับตารางที่ 2 องศาอิสระ ค่าความคลาดเคลื่อนมาตรฐาน และค่าของตัวสถิติทดสอบแสดงไว้ในตารางที่ 3

ตารางที่ 3 ค่าประมาณของพารามิเตอร์ของตัวแบบ Proportional odds model

Parameter	DF	$\hat{\beta}$	$S.E.\hat{\beta}$	Wald	P-value
Intercept	1	-2.8793	0.2889	99.3382	<.0001
Intercept	1	0.0651	0.2267	0.0826	0.7739
Edu(1)	1	-0.6683	0.1296	26.6136	<0.0001
Exp(1)	1	-0.9362	0.1885	24.6711	<0.0001
Exp(2)	1	0.1987	0.1638	1.4711	0.2252
Exp(3)	1	0.3469	0.2090	2.7537	0.0970
Occ(1)	1	1.3369	0.2633	25.7759	<0.0001
Occ(2)	1	-0.2637	0.3456	0.5824	0.4454
Occ(3)	1	-0.2842	0.2627	1.1698	0.2794
Occ(4)	1	-1.2222	0.3356	13.2625	0.0003
Occ(5)	1	0.8781	0.2160	16.5216	<0.0001
Sector (0)	1	-0.5512	0.2123	6.7399	0.0094
Sector (1)	1	0.1260	0.2150	0.3434	0.5578
Sex(0)	1	0.3503	0.1176	8.8694	0.0029
Union(0)	1	-0.5094	0.1305	15.2399	<0.0001

จากตารางที่ 3 แสดงสัมประสิทธิ์ของตัวแปรอธิบาย Edu, Exp, Occ, Sex, Sector และ Union มีนัยสำคัญทางสถิติที่ 0.05 (p-value < 0.05) สามารถเขียนเป็นสมการกรณีของตัวแปรตอบสนอง 3 กลุ่มได้ 2 สมการ ดังนี้

$$\text{logit}\,[P(Y \le 1|x_i)] = -2.8793 - 0.6683\text{Edu}(1) - 0.9362\text{Exp}(1) + 0.1987\text{Exp}(2) + 0.3469\text{Exp}(3) + 1.3369\text{Occ}(1) - 0.2637\text{Occ}(2) - 0.2842\text{Occ}(3) - 1.2222\text{Occ}(4) + 0.8781\text{Occ}(5) - 0.5512\text{Sector}(0) + 0.1260\text{Sector}(1) + 0.3503\text{Sex}(0) - 0.5094\text{Union}(0) \quad (3)$$

$$\text{logit}\,[P(Y \le 1|x_i)] = 0.0651 - 0.6683\text{Edu}(1) - 0.9362\text{Exp}(1) + 0.1987\text{Exp}(2) + 0.3469\text{Exp}(3) + 1.3369\text{Occ}(1) - 0.2637\text{Occ}(2) - 0.2842\text{Occ}(3) - 1.2222\text{Occ}(4) + 0.8781\text{Occ}(5) - 0.5512\text{Sector}(0) + 0.1260\text{Sector}(1) + 0.3503\text{Sex}(0) - 0.5094\text{Union}(0) \quad (4)$$

การวิเคราะห์ข้อมูลของตัวแปรตอบสนองกรณีตัวแบบลอจิตเมื่อทำการยุบกลุ่มของตัวแปรตอบสนองเชิงกลุ่มแบบมีลำดับ 3 กลุ่มในระดับที่อยู่ติดกันเป็น 2 กลุ่มแรก เป็นกลุ่มค่าจ้างระดับสูง (1) และที่เหลือเป็นกลุ่มค่าจ้างระดับต่ำ (2) ผลการทดสอบภาวะสารูปดีของตัวแบบ พบว่าตัวแบบมีภาวะสารูปดีกับข้อมูลระดับค่าจ้างกับตัวแปรอธิบายภายใต้ระดับนัยสำคัญทางสถิติที่ 0.05 (p-value = 0.5548) โดยมีค่า AIC, SC และ -2 log L หรือ G^2 เท่ากับ 545.744, 605.670 และ 517.744 ตามลำดับ ดังนั้นตัวแบบมีความเหมาะสมกับข้อมูล โดยสอดคล้องกับการทดสอบพารามิเตอร์ ($\beta = 0$) ของตัวแบบลอจิตด้วยตัวสถิติ Likelihood Ratio, Score, Wald ณ $\alpha = 0.05$ ภายใต้สถิติไคสแควร์ ณ (p-value < 0.0001) ดังตารางที่ 4

ตารางที่ 4 ตัวสถิติ Likelihood Ratio, Score และ Wald

Test	Chi-Square	DF	p-value
Likelihood Ratio	170.0690	13	<0.0001
Score	151.3492	13	<0.0001
Wald	113.5502	13	<0.0001

จากตารางที่ 4 สามารถสรุปได้ว่า ตัวแปรที่ได้รับการคัดเลือกจากตัวแปรทั้งหมดด้วยวิธี Stepwise ซึ่งประกอบด้วยตัวแปร Edu, Exp, Occ, Sector, Sex และ Union นั้น มีความเหมาะสมที่จะนำไปใช้ในการพยากรณ์ตัวแปรตอบสนองเชิงกลุ่ม 2 กลุ่ม อย่างมีนัยสำคัญทางสถิติที่ระดับ 0.05 (p-value < 0.0001)

ตารางที่ 5 การทดสอบเกี่ยวกับค่าประมาณสัมประสิทธิ์ในตัวแบบลอจิต

Effect	DF	Wald Chi-Square	Pr-Chisq
Edu	1	22.3678	<0.0001
Exp	3	22.5811	<0.0001
Occ	5	40.1785	0.0005
Sector	2	9.9781	0.0190
Sex	1	8.1723	0.0032
Union	1	17.9671	<0.0001

ค่าประมาณของพารามิเตอร์ในตัวแบบแต่ละค่าที่สอดคล้องกับตารางที่ 5 องศาอิสระ ค่าความคลาดเคลื่อนมาตรฐาน และค่าของตัวสถิติทดสอบ แสดงไว้ในตารางที่ 6

ตารางที่ 6 ค่าประมาณพารามิเตอร์ของตัวแบบลอจิตด้วยวิธีภาวะน่าจะเป็นสูงสุด

Parameter	DF	$\hat{\beta}$	$S.E.\hat{\beta}$	Wald	P-value
Intercept	1	0.2260	0.2407	0.8814	0.3478
Edu(1)	1	-0.6301	0.1332	22.3678	<0.0001
Exp(1)	1	-0.9098	0.1962	21.4988	<0.0001
Exp(2)	1	0.2839	0.1737	2.6714	0.1022
Exp(3)	1	0.2908	0.2239	1.6878	0.1939
Occ(1)	1	1.2783	0.2873	19.8019	<0.0001
Occ(2)	1	-0.1259	0.3477	0.1310	0.7173
Occ(3)	1	-0.2235	0.2658	0.7067	0.4005
Occ(4)	1	-1.2413	0.3416	13.2061	0.0003
Occ(5)	1	0.8601	0.2329	13.6385	0.0002
Sector(0)	1	-0.7029	0.2269	9.5974	0.0019
Sector(1)	1	0.0992	0.2274	0.1902	0.6628
Sex(0)	1	0.3542	0.1239	8.1723	0.0043
Union(0)	1	-0.6017	0.1419	17.9671	<0.0001

ผลลัพธ์จากตารางที่ 6 พบว่า ค่าประมาณพารามิเตอร์ของตัวแบบลอจิตของตัวแปรอธิบาย 6 ตัว คือ Edu, Exp, Occ, Sex, Sector และ Union มีนัยสำคัญทางสถิติที่ 0.05 (p-value < 0.0044) เขียนเป็นสมการในกรณีตัวแปรตอบสนอง 2 กลุ่ม ดังใน (5)

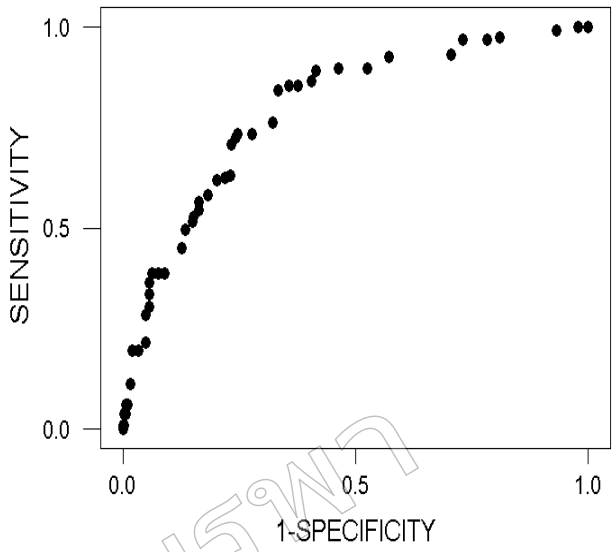
$$\begin{aligned}\text{logit}\,[P(Y \le 1|x_i)] &= 0.2260 - 0.6301\text{Edu}(1) - 0.9098\text{Exp}(1) \\ &\quad + 0.2839\text{Exp}(2) + 0.2908\text{Exp}(3) \\ &\quad + 1.2783\text{Occ}(1) - 0.1259\text{Occ}(2) \\ &\quad - 0.2235\text{Occ}(3) - 1.2413\text{Occ}(4) \\ &\quad - 0.8601\text{Occ}(5) - 0.7029\text{Sector}(0) \\ &\quad + 0.0992\text{Sector}(1) + 0.3542\text{Sex}(0) \\ &\quad - 0.6017\text{Union}(0) \qquad (5)\end{aligned}$$

นอกจากนี้ตัวแบบลอจิต ยังให้ประสิทธิภาพในการพยากรณ์ค่าจ้างระดับสูง (Wage = 1) ได้ถูกต้องร้อยละ 54.3 และการพยากรณ์ค่าจ้างระดับต่ำ (Wage=2) ได้ถูกต้องร้อยละ 83.7 ส่วนประสิทธิภาพในการพยากรณ์ในภาพรวมของตัวแบบลอจิตของตัวแปรตอบสนองเชิงกลุ่ม 2 กลุ่มสามารถพยากรณ์กลุ่มได้ถูกต้องร้อยละ 73.6 ได้ค่อนข้างสูง โดยเส้นโค้ง ROC ชิดแกนตั้งมาก ดังตารางที่ 7 และภาพที่ 2

ตารางที่ 7 ประสิทธิภาพในการพยากรณ์ได้ถูกต้องของตัวแบบลอจิตสองกลุ่ม

ค่าสังเกต (Observed)		ค่าพยากรณ์ (Predicted)		
		Wage		Percentage Correct
		1	2	
Wage	1	100	84	54.3
	2	57	293	83.7
ภาพรวม (Overall Percentage)				73.6

การวิเคราะห์ประสิทธิภาพของการพยากรณ์ด้วยตัวแบบลอจิต เพื่อใช้ในการวิเคราะห์ความไว (Sensitivity) ในการพยากรณ์ถูกของตัวแบบลอจิตสองกลุ่ม พบว่า ค่อนข้างสูงและดีมาก โดยเส้นโค้งชิดแกนตั้งมาก แกนตั้งแสดงค่า Sensitivity และแกนนอนแสดงค่า 1- Specificity โดยเส้นกราฟมีลักษณะโค้งเหนือเส้นทแยงมุม โดยมีลักษณะโค้งติดต่อกันจากจุด (0, 0) ถึงจุด (1, 1) ที่ชิดกับแกนตั้ง (Sensitivity) ดังนั้นตัวแบบลอจิตมีความไวในการพยากรณ์ถูกต้องได้ค่อนข้างสูง ดังภาพที่ 2



ภาพที่ 2 เส้นโค้งของ ROC Curve เมื่อตัวแปรตอบสนอง (Wage) จำแนกเป็น 2 กลุ่ม

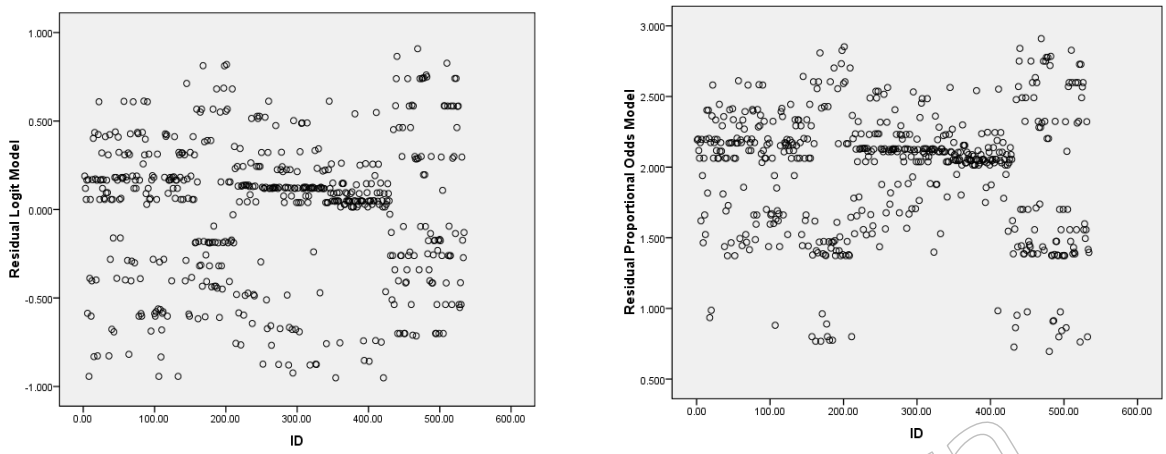
ตารางที่ 8 การเปรียบเทียบความเหมาะสมของตัวแบบ ด้วยตัวสถิติ AIC, SC, และ -2 Log L

ตัวแบบ	AIC	SC	-2 Log L
ตัวแบบลอจิต	545.744	605.670	517.744
ตัวแบบ Proportional odds model	688.305	752.511	658.305

จากตารางที่ 8 แสดงการเปรียบเทียบความเหมาะสมของตัวแบบ พบว่า ตัวแบบลอจิตให้ค่าสถิติ AIC, SC และ -2 Log L ที่ต่ำกว่า เท่ากับ 540.146, 591.511 และ 516.146 ตามลำดับ เมื่อเทียบกับตัวแบบ Proportional odds model ที่ให้ค่า AIC, SC และ -2 Log L เท่ากับ 688.305, 752.511 และ 658.305 ตามลำดับ ดังนั้นตัวแบบลอจิตสองกลุ่มจึงเหมาะสมกว่า

นอกจากนี้ การเปรียบเทียบลักษณะการกระจายของส่วนเหลือ (residuals plots) ของตัวแบบ พบว่า ตัวแบบลอจิตมีการกระจายของส่วนเหลืออย่างไม่มีรูปแบบได้มากกว่าและดีกว่าตัวแบบ Proportional odds model ดังภาพที่ 3

ดังนั้นตัวแบบที่มีความเหมาะสมกับข้อมูล คือ ตัวแบบลอจิต โดยพบว่ามีปัจจัยที่มีผลกระทบต่อระดับค่าจ้าง คือการศึกษา (Edu), ประสบการณ์การทำงาน (Exp), อาชีพ (Occ), แผนก (Sector), เพศ (Sex) และการเข้าร่วมเป็นสมาชิกสหภาพแรงงาน (Union) อย่างมีนัยสำคัญทางสถิติ (p-value < 0.0044)



ภาพที่ 3 การกระจายของส่วนเหลือของตัวแบบลอจิตสองกลุ่ม และตัวแบบ Proportional odds model

สรุปผลการวิจัย

ผลการวิจัยพบว่า ปัจจัยที่มีผลกระทบต่อระดับค่าจ้างอย่างมีนัยสำคัญทางสถิติที่ 0.05 ประกอบด้วย 6 ปัจจัย ได้แก่ การศึกษา (Edu), ประสบการณ์การทำงาน (Exp), อาชีพ (Occ), แผนก (Sector), เพศ (Sex) และการเข้าร่วมเป็นสมาชิกสหภาพแรงงาน (Union) โดยพบจากตัวแบบลอจิตสองกลุ่ม และทำนองเดียวกันกับของตัวแบบ Proportional odds model แต่จากการเปรียบเทียบความเหมาะสมของตัวแบบทั้งสองด้วยตัวสถิติ AIC, SC และ -2 Log L พบว่า ตัวแบบลอจิตสองกลุ่มมีความเหมาะสมกับข้อมูลที่ศึกษามากกว่า เพราะมีค่าสถิติ AIC, SC และ -2 Log L น้อยกว่าของตัวแบบ Proportional odds model และแผนภาพการกระจายส่วนเหลือของตัวแบบลอจิตมีการกระจายตัวดีกว่า ดังนั้น ตัวแบบที่เหมาะสมกับข้อมูลชุดนี้ คือ ตัวแบบลอจิตสองกลุ่ม นอกจากนี้จากการวิเคราะห์ความไวของตัวแบบลอจิตสองกลุ่มพบว่า มีประสิทธิภาพในการพยากรณ์ได้ถูกต้องร้อยละ 73.6 โดยเส้นกราฟที่พบมีลักษณะเส้นโค้งเหนือเส้นทแยงมุม และมีลักษณะโค้งชิดกับแกนตั้งที่แสดงความไว (Sensitivity) ค่อนข้างสูง ดังภาพที่ 1-2 ดังนั้นโดยภาพรวมจึงพบว่าตัวแบบลอจิตสองกลุ่ม ให้ผลการพยากรณ์กลุ่มที่ถูกต้องได้ค่อนข้างสูง จึงอาจนำไปใช้ในการพยากรณ์กลุ่มหรือระดับค่าจ้าง และสามารถตรวจสอบปัจจัยที่ส่งผลกระทบต่อระดับค่าจ้างต่อไป

ข้อเสนอแนะ และอภิปรายผล

เนื่องจากตัวแปรอธิบายในงานวิจัยนี้ ส่วนใหญ่เป็นปัจจัยที่เป็นลักษณะเฉพาะของลูกจ้างแต่ละคน ถ้าสามารถหาปัจจัยที่อาจส่งผลกระทบจากด้านสิ่งแวดล้อมที่เกี่ยวข้องอื่นๆ อาจนำไปสู่การสร้างตัวแบบที่มีความไวมากขึ้นอีก หากต้องการศึกษาเพิ่มเติมในเชิงการเปรียบเทียบตัวแบบในภาพรวมอาจใช้เทคนิคการจำลองแบบ หรือเพิ่มจำนวนตัวอย่างให้มากขึ้น นอกจากนี้การวิเคราะห์ข้อมูลของตัวแปรตอบสนองเชิงกลุ่มแบบมีลำดับมีอยู่หลายตัวแบบ ซึ่งอาจมีความเหมาะสมภายใต้เงื่อนไขทางทฤษฎีที่อาจนำไปสู่การประยุกต์ในงานวิจัยครั้งต่อไป

สำหรับการตรวจสอบการพยากรณ์ของตัวแบบในงานวิจัยครั้งต่อไป อาจใช้วิธีแยกข้อมูลเดิมออกเป็น 2 ชุดๆ ที่ 1 ใช้สร้างตัวแบบเพื่อวิเคราะห์ประสิทธิภาพและผลการพยากรณ์ของตัวแบบด้วยเทคนิคเชิงสถิติต่างๆ ส่วนชุดที่ 2 ใช้ตรวจสอบผลการพยากรณ์ด้วยตัวแบบและวิเคราะห์ความแม่นของตัวแบบจากการใช้ข้อมูลคนละชุดกับที่ใช้ในการสร้างตัวแบบ และในปัจจุบันเป็นที่นิยมใช้ทางด้าน Data mining ที่มีการวิเคราะห์ข้อมูลโดยใช้ข้อมูล 2 ชุดดังกล่าว

เอกสารอ้างอิง

ผึ้งพร ลาภส่งผล, ณัฐภาภรณ์ รอดรัตษะ, วีรานันท์ พงศาภักดี และทิติยา จิตติหรรษา. (2551). การใช้ตัวแบบโลจิทสะสมในการศึกษาข้อมูลการชะลอการเจริญของเส้นใยรา Ascophaera apis. วารสารวิทยาศาสตร์มหาวิทยาลัยนเรศวร, 5(2), 176-189.

ประภัสสร มีสกุล, อิทธิ ถิรนันท์ไพโรจน์ และวีรานันท์ พงศาภักดี. (2552). การใช้ตัวแบบโลจิทสะสมในการศึกษาปัจจัยที่มีผลกระทบต่อน้ำหนักข้าวสวยหลังหุงในกระบวนการผลิตข้าวกล่องแช่แข็ง. วารสารวิทยาศาสตร์ มหาวิทยาลัยนเรศวร, 6(1), 40-54.

วีรานันท์ พงศาภักดี. (2544). การวิเคราะห์ข้อมูลเชิงกลุ่ม. พิมพ์ครั้งที่ 2. นครปฐม: โรงพิมพ์มหาลัยศิลปากร.

วีรานันท์ พงศาภักดี. (2555). การวิเคราะห์ข้อมูลจำแนกประเภท : ทฤษฎีและการประยุกต์ด้วย GLIM, SPSS, SAS, และ MTB. พิมพ์ครั้งที่ 3. นครปฐม: โรงพิมพ์มหาลัยศิลปากร.

Agresti, A. (1990). Categorical Data Analysis. New York: John Wiley & Sons.

_____ (2002). Categorical Data Analysis. 2nd ed. New York: John Wiley & Sons.

_____ (2007). Analysis Introduction to Categorical Data Analysis. New York: John Wiley & Sons.

McCullagh, P. (1980). Regression Models for Ordinal Data. J. Royal, Statist. Soc. Ser, B42, 109-142.

Schwarz, G. (1978). Estimating the dimensions of a model. Annals of statistics, 6, 461-464.

StatLib. Determinants of Wage from the Current Population Survey. Retrieved January, 2010, from http://lib.stat.cmu.edu/.