ตัวแบบการถดถอยลอจิสติกสำหรับการวิเคราะห์การขาดอายุของกรมธรรม์ประกันชีวิต

Logistic Regression Model for Lapse Analysis of Life Insurance Policy

อริญชย์ บุญมีคำ*, วินัย โพธิ์สุวรรณ์ และ ธิดาพร ศุภภากร

Arin Boonmeekham*, Winai Bodhisuwan and Thidaporn Supapakorn

ภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์

Statistics Department, Faculty of Science, Kasetsart University

Received : 28 December 2018

Revised : 21 February 2019

Accepted : 9 May 2019

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อสร้างตัวแบบการถดถอยลอจิสติกสำหรับพยากรณ์การขาดอายุกรมธรรม์ของกรมธรรม์ประกันชีวิตที่มีตัวแปรตาม คือ การขาดอายุกรมธรรม์ซึ่งเป็นตัวแปรเชิงคุณภาพทวิภาค และมีตัวแปรอิสระซึ่งเป็นตัวแปรเชิงคุณภาพ หรือตัวแปรเชิงปริมาณรวมทั้งหมด 18 ตัว โดยแบ่งข้อมูลที่ศึกษาออกเป็น 2 ชุด ได้แก่ ชุดข้อมูลฝึกหัดสำหรับสร้างสมการพยากรณ์จำนวน 1,864 กรมธรรม์ และชุดข้อมูลทดสอบจำนวน 466 กรมธรรม์ ผลการวิจัยพบว่าสมการพยากรณ์ประกอบด้วยตัวแปรอิสระทั้งหมด 6 ตัว ได้แก่ อายุ จำนวนเงินเอาประกันชีวิตระหว่าง 50,001 - 100,000 บาท ระยะเวลาการชำระเบี้ยประกันภัยมากกว่า 3 ปี รากที่สองของระยะเวลาความคุ้มครอง ระดับชั้นอาชีพแบบอาชีพชั้น 3 และอาชีพชั้น 4 ซึ่งจากสมการที่ได้นี้สามารถพยากรณ์ได้ว่าผู้เอาประกันภัยเกิดการขาดอายุกรมธรรม์ร้อยละ 31.76 และมีประสิทธิภาพในการพยากรณ์ได้ถูกต้องร้อยละ 66.95

คำสำคัญ: การขาดอายุกรมธรรม์, การประกันชีวิต, การถดถอยลอจิสติก, ฟังก์ชันตอบสนองลอจิต

*Corresponding author. E-mail : arin.knotztu.1412@gmail.com

Abstract

The purpose of this research is to create the logistic regression model for forecasting life insurance lapse policy. The dependent variable is lapse class that is dichotomous qualitative variable. There are 18 independent variables which are qualitative or quantitative. The data set is divided into 2 parts; training set of 1,864 policies for building the predicting equation and testing set of 466 policies. The results show that the predicting equation consist of 6 independent variables which are age, face amount (between 50,001 – 100,000), duration of payment (more than 3 years), square root of duration of protection and occupation class (class 3 and class 4). The logistic regression equation predict that the lapse rate of insured is 31.76% and the accuracy performance of forecasting is 66.95%.

Keywords: policy lapse, life insurance, logistic regression, logit response function

บทนำ

ธุรกิจการประกันภัย ได้แบ่งการประกันภัยออกเป็นสองประเภท ได้แก่ การประกันวินาศภัย และการประกันชีวิต จากการคาดการณ์ถึงแนวโน้มการเติบโตของธุรกิจประกันชีวิต ว่าจะยังคงมีอัตราการเติบโตอย่างต่อเนื่องประมาณร้อยละ 4-6 นั้น เป็นผลมาจากการคาดการณ์เศรษฐกิจภายในประเทศที่จะขยายตัวดีขึ้นประมาณร้อยละ 3.6-4.6 จากปัจจัยเศรษฐกิจโลกที่มีแนวโน้มขยายตัว และแรงขับเคลื่อนภายในประเทศจากภาคการส่งออก การท่องเที่ยว การลงทุนของภาครัฐและเอกชน ตลอดจนการบริโภคของภาคเอกชนทำให้เศรษฐกิจมีการหมุนเวียนดีขึ้นส่งผลให้เกิดการกระจายรายได้ ประชาชนมีกำลังซื้อเพิ่มมากขึ้น และตระหนักถึงความสำคัญของการวางแผนทางการเงินตลอดจนการบริหารความเสี่ยงของอนาคต (Thai Life Assurance Association, 2018)

ในปัจจุบันประเทศไทยนั้นได้แบ่งการประกันชีวิตประเภทสามัญ (Ordinary insurance) ออกเป็น 4 ประเภท ได้แก่ การประกันชีวิตแบบตลอดชีพ (Whole life insurance) การประกันชีวิตแบบชั่วระยะเวลา (Term life insurance) การประกันชีวิตแบบเงินรายปี (Annuity life insurance) และการประกันชีวิตแบบสะสมทรัพย์ (Endowment life insurance) (Office of Insurance Commission, 2016) ซึ่งผลิตภัณฑ์ประกันชีวิตแต่ละแบบนั้นมีระยะเวลาคุ้มครองและผลประโยชน์ที่แตกต่างกันไป ทำให้ผลกำไรของผลิตภัณฑ์ประกันชีวิตแต่ละแบบแตกต่างกันด้วย ในช่วงแรกของการออกกรมธรรม์ประกันภัย บริษัทประกันชีวิตจะต้องแบกรับภาระค่าใช้จ่ายสูงกว่าจำนวนเบี้ยประกันภัยที่ได้รับ เช่น ค่าบำเหน็จ (Commission) ซึ่งกรมธรรม์บางแบบมีค่าบำเหน็จสูงถึงร้อยละ 40 และลดลงในปีถัดไป ค่าใช้จ่ายในการดำเนินงาน (Administrative expense) และค่าใช้จ่ายของการดำเนินการกรมธรรม์ (Policy expense) เป็นต้น แต่ในบางครั้งผู้เอาประกันชีวิตได้รับกรมธรรม์ประกันชีวิตที่ไม่ตรงกับความต้องการ เช่น เงื่อนไขของกรมธรรม์ไม่เป็นไปตามที่ตกลงในการขอทำประกันภัย หรือเหตุผลอื่น ๆ ที่ทำให้ผู้เอาประกันภัยจำเป็นต้องยกเลิกกรมธรรม์ไป ซึ่งการขาดอายุกรมธรรม์ (Lapse) ของผู้เอาประกันภัยเกิดขึ้นจากการที่ผู้เอาประกันภัยมีการยกเลิกกรมธรรม์ประกันภัยโดยการเวนคืนกรมธรรม์ประกันภัยก่อนที่กรมธรรม์จะครบกำหนดอายุ ดังนั้นหากบริษัทประกันชีวิตมีอัตราการเวนคืนกรมธรรม์ประกันภัยในช่วงปีแรกสูง ทำให้บริษัทประกันชีวิตอาจมีรายรับไม่เพียงพอต่อรายจ่ายที่เกิดขึ้นข้างต้น และทำให้บริษัทประกันชีวิตสูญเสียผลกำไร ในวงการธุรกิจในปัจจุบันมีการแข่งขันและมีความสลับซับซ้อนมากขึ้น

ดังนั้นเทคนิคการพยากรณ์จึงเป็นอีกหนึ่งเทคนิคที่สำคัญต่อการพัฒนาและส่งผลต่อความก้าวหน้าของบริษัทประกันชีวิต ทำให้เกิดความต้องการในการพยากรณ์เพื่อนำผลมาประกอบการตัดสินใจทางธุรกิจต่าง ๆ ซึ่งมีงานวิจัยที่ได้นำตัวแบบการถดถอยลอจิสติก (Logistic regression model) มาใช้ในการพยากรณ์การขาดอายุกรมธรรม์ (Zian *et al*., 2016) พบว่าตัวแปรที่เหมาะสมสำหรับการสร้างตัวแบบการถดถอยลอจิสติกได้แก่ เพศ รูปแบบของกรมธรรม์ เบี้ยประกันภัย รูปแบบการชำระเบี้ยประกันภัย ระยะเวลาความคุ้มครอง ผลิตภัณฑ์มวลรวมภายในประเทศ การว่างงาน วิกฤติทางการเงิน อัตราดอกเบี้ย รวมไปถึงปีปฏิทิน เป็นต้น

เนื่องจากการขาดอายุกรมธรรม์ประกันชีวิตเป็นสาเหตุหลักในการกำหนดราคาผลิตภัณฑ์ประกันชีวิต อีกทั้งยังมีผลกระทบโดยตรงกับระยะเวลาของการชำระเบี้ยประกันภัยและผลกำไรของบริษัทประกันชีวิต ดังนั้นผู้วิจัยสนใจที่จะสร้างตัวแบบเพื่อพยากรณ์การขาดอายุกรมธรรม์ประกันชีวิตโดยใช้ตัวแบบการถดถอยลอจิสติก

การวิเคราะห์การถดถอยลอจิสติก

การวิเคราะห์การถดถอยลอจิสติกเป็นเทคนิคการวิเคราะห์สถิติเชิงคุณภาพ (Qualitative statistical analysis) มีวัตถุประสงค์เพื่อศึกษาความสัมพันธ์ระหว่างตัวแปรตาม ซึ่งเป็นตัวแปรเชิงคุณภาพที่มีค่าอย่างน้อย 2 ค่า กับตัวแปรอิสระ ซึ่งอาจเป็นตัวแปรเชิงปริมาณทุกตัว หรือตัวแปรเชิงคุณภาพทุกตัว หรือมีทั้งตัวแปรเชิงปริมาณและเชิงคุณภาพก็ได้ และทำนายโอกาสที่จะเกิดเหตุการณ์ที่สนใจ โดยตัวแบบการถดถอยลอจิสติกสามารถแบ่งออกได้เป็น 2 ประเภท (Sinsomboonthong, 2016) ได้แก่ 1) การถดถอยลอจิสติกทวิภาค (Binary logistic regression) มีตัวแปรตาม Y เป็นตัวแปรเชิงคุณภาพที่มีค่าเพียง 2 ค่า (Dichotomous variable) 2) การถดถอยลอจิสติกอเนกนาม (Multinomial logistic regression) ตัวแปรตาม Y เป็นตัวแปรเชิงคุณภาพ มีค่ามากกว่า 2 ค่าขึ้นไป ส่วนตัวแปรอิสระ X เป็นตัวแปรเชิงปริมาณหรือเชิงคุณภาพหรือมีทั้งตัวแปรเชิงปริมาณและเชิงคุณภาพ

การถดถอยลอจิสติกทวิภาค

การถดถอยลอจิสติกทวิภาคเป็นการศึกษาความสัมพันธ์ของตัวแปรตาม ซึ่งเป็นตัวแปรเชิงคุณภาพที่มีค่าเพียง 2 ค่า กำหนดให้ตัวแปร Y และ X มีความสัมพันธ์ภายใต้ตัวแบบการถดถอยลอจิสติกทวิภาค โดย Y มีการแจกแจงแบร์นูลลี (Bernoulli Distribution) และมีตัวแปรอิสระจำนวน p ตัว ซึ่งความน่าจะเป็นที่จะการเกิดเหตุการณ์ที่สนใจ คือ π ดังนี้

$$Y=\begin{cases} 1\ ; \text{ เมื่อเกิดเหตุการณ์ที่สนใจ ด้วยความน่าจะเป็น } \pi \\ 0\ ; \text{ เมื่อไม่เกิดเหตุการณ์ที่สนใจ ด้วยความน่าจะเป็น } 1-\pi \end{cases} \qquad (1)$$

ดังนั้น P (เกิดเหตุการณ์ที่สนใจ) $= P(Y=1)$

$$= E(Y)$$
$$= \pi$$
$$= \frac{e^{(\beta_0+\beta_1 X_1+...+\beta_p X_p)}}{1+e^{(\beta_0+\beta_1 X_1+...+\beta_p X_p)}}$$
$$= \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+...+\beta_p X_p)}} \quad (2)$$

จากสมการที่ (2) พบว่าความสัมพันธ์ระหว่างโอกาสเกิดเหตุการณ์ที่สนใจกับตัวแปรอิสระไม่ได้อยู่ในรูปเชิงเส้น จึงปรับให้อยู่ในรูปเชิงเส้น (Kleinbaum, 2002) โดยกำหนดให้

$$Odds\ Ratio = OR$$
$$= \frac{P(Y=1)}{P(Y=0)}$$
$$= \frac{\pi}{1-\pi}$$
$$= e^{\beta_0+\beta_1 X_1+...+\beta_p X_p} \quad (3)$$

ถ้า OR มีค่ามากกว่า 1 แสดงถึงโอกาสเกิดเหตุการณ์ที่สนใจมากกว่าโอกาสที่จะไม่เกิดเหตุการณ์ที่สนใจ จากสมการที่ (3) หาค่าฟังก์ชันตอบสนองลอจิต (Logit response function: $Logit$) พบว่า $Logit$ อยู่ในรูปเชิงเส้นดังสมการที่ (4)

$$Logit = \log_e(OR)$$
$$= \ln(OR)$$
$$= \ln(e^{\beta_0+\beta_1 X_1+...+\beta_p X_p})$$
$$= \beta_0 + \beta_1 X_1 + ... + \beta_p X_p \quad (4)$$

ข้อสมมติเบื้องต้นของการถดถอยลอจิสติกทวิภาค

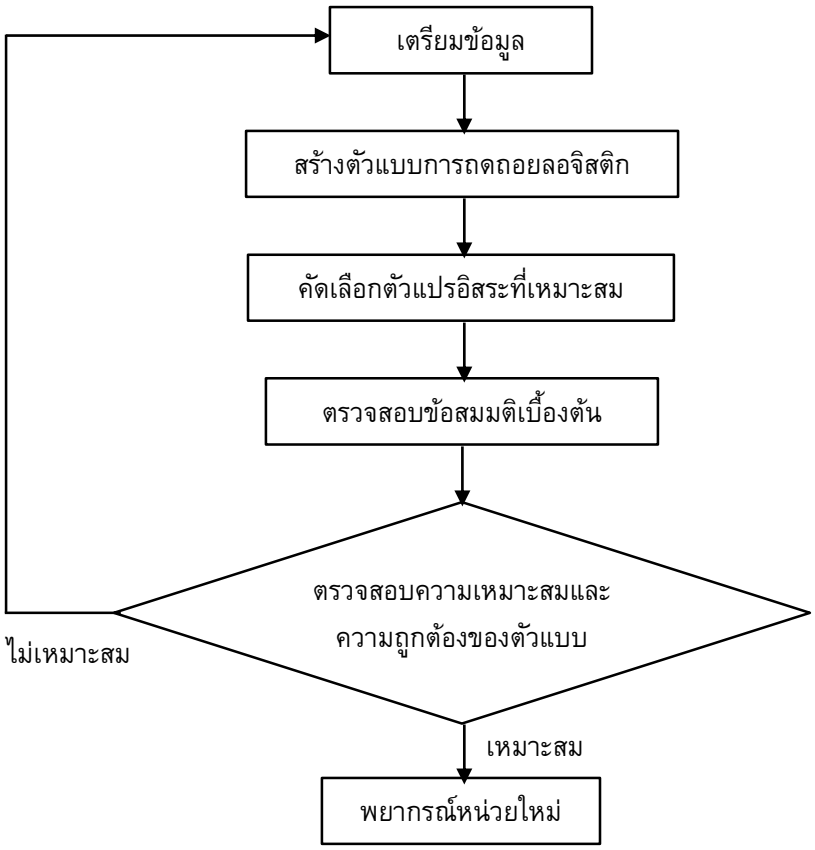
การวิเคราะห์การถดถอยลอจิสติกทวิภาคมีข้อสมมติเบื้องต้น ทั้งหมด 2 ข้อ (Hilbe, 2019; Tabachnick, 2013) ดังนี้

1) ตัวแปรอิสระแบบต่อเนื่องมีความสัมพันธ์เชิงเส้น (Linearity) กับฟังก์ชันตอบสนองลอจิต

2) ตัวแปรอิสระไม่มีความสัมพันธ์เชิงเส้นแบบพหุ (Multicollinearity)

วิธีดำเนินการวิจัย

การดำเนินงานในวิจัยนี้ สามารถเขียนในรูปของผังงาน (Flow chart) ดังภาพที่ 1



ภาพที่ 1 แผนผังการดำเนินงานวิจัย

การเตรียมข้อมูล

เตรียมข้อมูลที่ใช้ในการวิเคราะห์ โดยใช้ข้อมูลจากบริษัทประกันชีวิตปี พ.ศ. 2560 จำนวน 2,330 กรมธรรม์ มีตัวแปรอิสระทั้งหมด 18 ตัว แบ่งเป็น

1) ตัวแปรเชิงปริมาณจำนวน 2 ตัว คือ อายุ $\left(X_1\right)$ และระยะเวลาความคุ้มครอง $\left(X_4\right)$

2) ตัวแปรเชิงคุณภาพจำนวน 16 ตัว คือ จำนวนเงินเอาประกันชีวิต $\left(X_{2A}-X_{2B}\right)$ ระยะเวลาการชำระเบี้ยประกันภัย $\left(X_3\right)$ รูปแบบการชำระเบี้ยประกันภัย $\left(X_{5A}-X_{5D}\right)$ ระดับชั้นอาชีพ $\left(X_{6A}-X_{6D}\right)$ เพศ $\left(X_7\right)$ เบี้ยประกันภัย $\left(X_8\right)$ ช่องทางการจำหน่าย $\left(X_9\right)$ และรูปแบบกรมธรรม์ $\left(X_{10A}-X_{10B}\right)$ รายละเอียดดังตารางภาคผนวกที่ 1 ซึ่งตัวแปรเชิงคุณภาพนั้นผู้วิจัยได้กำหนดค่าเป็นตัวแปรหุ่น (Dummy variable) โดยที่แต่ละตัวมีค่าเป็นไปได้แค่ 0 และ 1 เท่านั้น กล่าวคือถ้าตัวแปรเชิงคุณภาพตัวนั้นแบ่งเป็น k กลุ่ม จะต้องกำหนดตัวแปรหุ่นเป็นจำนวน $k-1$ ตัว เช่น จำนวนเงินเอาประกันชีวิตแบ่งเป็น 3 กลุ่ม คือ กลุ่มที่ 1 น้อยกว่าหรือเท่ากับ 50,000 บาท กลุ่มที่ 2 50,001-100,000 บาท และกลุ่มที่ 3

มากกว่า 100,000 บาท ดังนั้นผู้วิจัยจึงสร้างตัวแปรหุ่น 2 ตัว คือ X_{2A} และ X_{2B} ซึ่งทั้ง X_{2A} และ X_{2B} มีค่าเป็นไปได้แค่ 0 และ 1 เท่านั้น โดยกำหนดให้กลุ่มที่ 1 น้อยกว่าหรือเท่ากับ 50,000 บาท เป็นกลุ่มอ้างอิงแล้ว จะได้ว่า

ถ้า $X_{2A}=0$ และ $X_{2B}=0$ แทน จำนวนเงินเอาประกันชีวิตน้อยกว่าหรือเท่ากับ 50,000 บาท

$X_{2A}=1$ และ $X_{2B}=0$ แทน จำนวนเงินเอาประกันชีวิต 50,001 - 100,000 บาท

และ $X_{2A}=0$ และ $X_{2B}=1$ แทน จำนวนเงินเอาประกันชีวิตมากกว่า 100,000 บาท

การใช้ข้อมูลทั้งหมดมาสร้างตัวแบบการถดถอยและใช้ตัวแบบการถดถอยนั้นไปพยากรณ์ข้อมูลชุดเดิม อาจทำให้สัดส่วนการพยากรณ์ได้ถูกต้องมีค่าสูง และทำให้ผู้ศึกษามีความมั่นใจที่จะนำตัวแบบไปพยากรณ์ข้อมูลหน่วยใหม่ ๆ ทั้ง ๆ ที่เมื่อนำไปพยากรณ์หน่วยใหม่แล้วอาจจะไม่ถูกต้องก็ได้ ดังนั้นจึงควรแบ่งข้อมูลออกเป็น 2 ส่วน ดังนี้

1) ข้อมูลฝึกหัด (Training data) เป็นข้อมูลส่วนที่นำมาสร้างตัวแบบการพยากรณ์ โดยให้ m_1 แทน จำนวนข้อมูลฝึกหัด

2) ข้อมูลทดสอบ (Testing data) เป็นข้อมูลส่วนที่นำมาใช้ตรวจสอบหรือทดสอบ โดยนำตัวแบบที่ได้จากข้อมูลฝึกหัดมาพยากรณ์หน่วยใหม่ที่อยู่ในข้อมูลทดสอบนี้ โดยให้ m_2 แทน จำนวนข้อมูลทดสอบ และ n แทน จำนวนข้อมูลทั้งหมด ในที่นี้ $n=m_1+m_2$ โดยที่ในงานวิจัยนี้ได้แบ่งอัตราส่วนระหว่างชุดข้อมูลฝึกหัดกับข้อมูลทดสอบเท่ากับ $80:20$ นั่นคือ $m_1=1,864$ และ $m_2=466$

การสร้างตัวแบบการถดถอยลอจิสติก

เมื่อเตรียมข้อมูลเรียบร้อยแล้ว นำตัวแปรเหล่านั้นเข้าสู่ตัวแบบการถดถอยลอจิสติก จะได้ฟังก์ชันตอบสนองลอจิตของตัวแบบเต็ม (full model) ดังสมการที่ (5)

$$\begin{aligned} Logit = {} & \beta_0+\beta_1X_1+\beta_{2A}X_{2A}+\beta_{2B}X_{2B}+\beta_3X_3+\beta_4X_4+\beta_{5A}X_{5A}+\beta_{5B}X_{5B}+\beta_{5C}X_{5C} \\ & +\beta_{5D}X_{5D}+\beta_{6A}X_{6A}+\beta_{6B}X_{6B}+\beta_{6C}X_{6C}+\beta_{6D}X_{6D}+\beta_7X_7+\beta_8X_8+\beta_9X_9 \\ & +\beta_{10A}X_{10A}+\beta_{10B}X_{10B}+\varepsilon \end{aligned} \quad (5)$$

การคัดเลือกตัวแปรอิสระที่เหมาะสม

สมการฟังก์ชันตอบสนองลอจิตที่ดีควรจะประกอบไปด้วยตัวแปรอิสระที่เหมาะสม ไม่เกิดการ Under-fit หรือ Over-fit เพื่อทำให้ค่าพยากรณ์โอกาสที่จะเกิดเหตุการณ์ที่สนใจมีความใกล้เคียงกับค่าจริงมากที่สุด ซึ่งการพิจารณาว่าตัวแปรอิสระใดบ้างที่มีอิทธิพลต่อการเกิดเหตุการณ์ที่สนใจมีอยู่หลายวิธี โดยงานวิจัยนี้ได้เลือกใช้วิธีการ Backward stepwise โดยในงานวิจัยนี้พิจารณาการคัดเลือกตัวแปรอิสระจากเกณฑ์สารสนเทศอะกะอิเกะ (เอไอซี) เกณฑ์นี้จะเลือกตัวแบบที่ประกอบไปด้วยตัวแปรอิสระที่ให้ค่าเอไอซีต่ำที่สุด กล่าวคือตัวแบบที่ได้มีแนวโน้มให้ค่าพยากรณ์ที่ใกล้เคียงกับค่าจริง โดยขั้นตอนการคัดเลือกตัวแปรอิสระจะเริ่มด้วยการสร้างตัวแบบเต็ม หลังจากนั้นจะนำตัวแปรอิสระออกจากสมการถดถอยครั้งละ 1 ตัว จากตารางที่ 1 พบว่าตัวแปรอิสระ 10 ตัว ได้แก่ รูปแบบการชำระเบี้ยประกันภัยแบบรายปีและรายครึ่งปี เบี้ยประกันภัยมากกว่า 10,000 บาท ช่องทางการจำหน่ายทางบริษัท รูปแบบกรมธรรม์แบบชั่วระยะเวลาและแบบสะสมทรัพย์ เพศหญิง จำนวนเงิน

เอาประกันชีวิตมากกว่า 100,000 บาท และระดับชั้นอาชีพแบบอาชีพชั้น 1 และอาชีพชั้น 2 ถูกนำออกจากสมการฟังก์ชันตอบสนองลอจิต แล้วประมาณค่าสัมประสิทธิ์การถดถอยของตัวแปรอิสระที่เหลือ รายละเอียดแสดงดังตารางที่ 2 ได้ฟังก์ชันตอบสนองลอจิตหลังคัดเลือกตัวแปรอิสระที่เหมาะสม ดังสมการที่ (6)

$$Logit = -16.9881 - 0.0182X_1 - 0.3465X_{2A} - 0.9448X_3 - 0.0643X_4 + 18.1759X_{5A} + 17.9908X_{5B} + 0.4202X_{6C} + 0.7785X_{6D} \quad (6)$$

ตารางที่ 1 ขั้นตอนการคัดเลือกตัวแปรอิสระ

ขั้นตอนที่	การดำเนินการ	องศาเสรี	เอไอซี
1	ตัวแบบเต็ม		1,937.57
2	นำ X_{5D} ออกจากฟังก์ชันตอบสนองลอจิต	1	1,935.57
3	นำ X_{5C} ออกจากฟังก์ชันตอบสนองลอจิต	1	1,933.57
4	นำ X_8 ออกจากฟังก์ชันตอบสนองลอจิต	1	1,931.57
5	นำ X_9 ออกจากฟังก์ชันตอบสนองลอจิต	1	1,929.57
6	นำ X_{10B} ออกจากฟังก์ชันตอบสนองลอจิต	1	1,927.66
7	นำ X_7 ออกจากฟังก์ชันตอบสนองลอจิต	1	1,925.75
8	นำ X_{2B} ออกจากฟังก์ชันตอบสนองลอจิต	1	1,923.84
9	นำ X_{6B} ออกจากฟังก์ชันตอบสนองลอจิต	1	1,921.99
10	นำ X_{6A} ออกจากฟังก์ชันตอบสนองลอจิต	1	1,920.21
11	นำ X_{10A} ออกจากฟังก์ชันตอบสนองลอจิต	1	1,918.92

ตารางที่ 2 ค่าสัมประสิทธิ์ของการถดถอยลอจิสติกเมื่อมีตัวแปรอิสระ 8 ตัว (1)

ตัวแปรอิสระ	$\hat{\beta}$	SE	Z	ค่าพี
ค่าคงที่	-16.9881	346.6896	-0.05	0.96
X_1	-0.0182	0.0056	-3.22	0.00
X_{2A}	-0.3465	0.1541	-2.25	0.02
X_3	-0.9448	0.1634	-5.78	0.00
X_4	-0.0643	0.2248	-2.86	0.00
X_{5A}	18.1759	346.6895	0.05	0.96
X_{5B}	17.9908	346.6897	0.05	0.96
X_{6C}	0.4202	0.2217	1.90	0.06
X_{6D}	0.7785	0.5050	1.54	0.12

การตรวจสอบข้อสมมติเบื้องต้นของการวิเคราะห์การถดถอยลอจิสติก

1) การทดสอบความสัมพันธ์เชิงเส้นของตัวแปรอิสระแบบต่อเนื่องกับฟังก์ชันตอบสนองลอจิต ใช้การทดสอบของ Box-Tidwell (Hilbe, 2009) ทำได้โดยการเพิ่มปฏิสัมพันธ์ (Interaction) หรือ $ln(X_i)\times X_i$ และสร้างตัวแบบการถดถอยด้วยตัวแปรอิสระและพจน์ปฏิสัมพันธ์ของตัวแปรอิสระตัวนั้นหรือ $Logit = \beta_0 + \beta_1 X_i + \beta_2 \left[ln(X_i)\times X_i\right]$ ถ้าพจน์ปฏิสัมพันธ์มีนัยสำคัญทางสถิติแล้วแสดงว่าตัวแปรอิสระนั้นไม่มีความสัมพันธ์เชิงเส้นกับฟังก์ชันตอบสนองลอจิต พบว่าในตัวแบบการถดถอยลอจิสติกมีตัวแปรอิสระแบบต่อเนื่อง 2 ตัว คือ อายุ และระยะเวลาความคุ้มครอง

ตารางที่ 3 การทดสอบความสัมพันธ์เชิงเส้นของตัวแปรอายุและระยะเวลาความคุ้มครอง

การทดสอบ	$\hat{\beta}$	SE	Z	ค่าพี
<u>อายุ</u>				
- ค่าคงที่	2.4741	1.1936	2.07	0.04
- X_1	-0.2088	0.1487	-1.40	0.16
- $ln(X_1)\times X_1$	0.0352	0.0320	1.10	0.27
<u>ระยะเวลาความคุ้มครอง</u>				
- ค่าคงที่	1.0898	0.2001	5.45	0.00
- X_4	-0.5479	0.0834	-6.57	0.00
- $ln(X_4)\times X_4$	0.1221	0.0250	4.89	0.00

จากตารางที่ 3 พบว่าปฏิสัมพันธ์ของอายุมีค่าพี (p-value) มากกว่าค่าระดับนัยสำคัญ (α) เท่ากับ 0.05 ดังนั้นจึงไม่มีนัยสำคัญทางสถิติ นั่นคือ มีความสัมพันธ์เชิงเส้นระหว่างตัวแปรอิสระกับฟังก์ชันตอบสนองลอจิต ส่วนปฏิสัมพันธ์ของระยะเวลาความคุ้มครองมีค่าพีน้อยกว่าค่าระดับนัยสำคัญ 0.05 ดังนั้นจึงไม่มีความสัมพันธ์เชิงเส้นระหว่างตัวแปรอิสระกับฟังก์ชันตอบสนองลอจิต จึงต้องทำการแปลงข้อมูลด้วยวิธีการแปลงของ Box-Tidwell (Hilbe, 2009) ดังนี้

$$X_4' = \sqrt{X_4} \qquad (7)$$

เมื่อทำการทดสอบ Box-Tidwell พบว่า ปฏิสัมพันธ์ของระยะเวลาความคุ้มครองมีค่าพีเท่ากับ 0.80 ซึ่งมีค่ามากกว่าค่าระดับนัยสำคัญ 0.05 ดังนั้น จึงไม่มีนัยสำคัญทางสถิติ นั่นคือ มีความสัมพันธ์เชิงเส้นระหว่างตัวแปรอิสระกับฟังก์ชันตอบสนองลอจิต รายละเอียดแสดงดังตารางที่ 4

ตารางที่ 4 การทดสอบความสัมพันธ์เชิงเส้นของตัวแปรระยะเวลาความคุ้มครองครั้งที่ 2

การทดสอบ	$\hat{\beta}$	SE	Z	ค่าพี
ระยะเวลาความคุ้มครอง				
- ค่าคงที่	3.7333	1.2919	2.89	0.00
- X_4'	-2.7406	1.0183	-2.69	0.01
- $ln(X_4') \times X_4'$	0.9204	0.5188	1.77	0.08

2) การทดสอบความสัมพันธ์เชิงเส้นแบบพหุ (Multicollinearity) ทำได้โดยการตรวจสอบค่าความคลาดเคลื่อนมาตรฐาน (SE) ของสัมประสิทธิ์การถดถอย $(\hat{\beta})$ (Josephat, 2018) จากตารางที่ 5 พบว่าตัวแปรอิสระรูปแบบการชำระเบี้ยประกันภัยแบบรายเดือนและแบบรายไตรมาส มีค่าความคลาดเคลื่อนมาตรฐานมากกว่า 2.0 ดังนั้นจึงต้องนำตัวแปรอิสระรูปแบบการชำระเบี้ยประกันภัยแบบรายเดือนและแบบรายไตรมาสออกจากตัวแบบการถดถอยลอจิสติก (Tabachnick, 2013) พบว่าค่าความคลาดเคลื่อนมาตรฐานของตัวแปรอิสระแต่ละตัวมีค่าไม่เกิน 2.0 ดังนั้นตัวแปรอิสระแต่ละตัวไม่มีความสัมพันธ์เชิงเส้นแบบพหุ รายละเอียดแสดงดังตารางที่ 6 และสามารถสร้างฟังก์ชันตอบสนองลอจิตดังนี้

$$Logit = 1.6004 - 0.0316X_1 - 0.2396X_{2A} - 0.8553X_3 - 0.2785X_4' + 0.4924X_{6C} + 1.0946X_{6D} \quad (8)$$

ตารางที่ 5 ค่าสัมประสิทธิ์ของการถดถอยลอจิสติกเมื่อมีตัวแปรอิสระ 8 ตัว (2)

ตัวแปรอิสระ	$\hat{\beta}$	SE	Z	ค่าพี
ค่าคงที่	-16.5211	346.4035	-0.05	0.96
X_1	-0.0182	0.0056	-3.22	0.00
X_{2A}	-0.3456	0.1541	-2.24	0.02
X_3	-0.8713	0.1793	-4.86	0.00
X_4'	-0.3892	0.1328	-2.93	0.00
X_{5A}	18.1904	346.4033	0.05	0.96
X_{5B}	18.0051	346.4035	0.05	0.96
X_{6C}	0.4188	0.2217	1.89	0.06
X_{6D}	0.7769	0.5048	1.54	0.12

ตารางที่ 6 ค่าสัมประสิทธิ์ของการถดถอยลอจิสติกเมื่อมีตัวแปรอิสระ 6 ตัว

ตัวแปรอิสระ	$\hat{\beta}$	SE	Z	ค่าพี	OR	ช่วงความเชื่อมั่นของ OR	
						ขอบล่าง	ขอบบน
ค่าคงที่	1.6004	0.2855	5.61	0.00	4.9551	2.8523	8.7393
X_1	-0.0316	0.0051	-6.21	0.00	0.9689	0.9592	0.9786
X_{2A}	-0.2396	0.1486	-1.61	0.11	0.7870	0.5863	1.0504
X_3	-0.8553	0.1720	-4.97	0.00	0.4252	0.3029	0.5947
X_4'	-0.2785	0.1232	-2.26	0.02	0.7569	0.5910	0.9579
X_{6C}	0.4924	0.2053	2.40	0.02	1.6363	1.0956	2.4546
X_{6D}	1.0946	0.5077	2.16	0.03	2.9879	1.1572	8.7520

ผลการวิจัย

ผลการวิจัยจะประกอบไปด้วย 3 ส่วน ได้แก่ การตรวจสอบความเหมาะสมของตัวแบบการถดถอยลอจิสติก การตรวจสอบความถูกต้องของตัวแบบการถดถอยลอจิสติก และการพยากรณ์หน่วยใหม่ ดังนี้

การตรวจสอบความเหมาะสมของตัวแบบการถดถอยลอจิสติก

ตรวจสอบความเหมาะสมของตัวแบบการถดถอยลอจิสติกจากค่าสถิติไคกำลังสอง $\left(\chi^2\right)$ การประเมินตัวแบบโดยรวม (Overall model evaluation) ด้วยวิธี Likelihood ratio test มีค่าพีเท่ากับ 0.00 ซึ่งมีค่าน้อยกว่าระดับนัยสำคัญ 0.05 (Ae, 2013) แสดงว่าตัวแบบที่มีตัวแปรอิสระมีความเหมาะสมกว่าตัวแบบที่มีแค่เพียงค่าคงที่ และการทดสอบภาวะสารูปดี (Goodness of fit test) ด้วยวิธีฮอสเมอร์และลีเมสโชว์ มีค่าพีเท่ากับ 0.39 ซึ่งมีค่ามากกว่าระดับนัยสำคัญ 0.05 (Peng, 2002; Allison, 2014) แสดงว่าตัวแบบการถดถอยลอจิสติกมีความเหมาะสม รายละเอียดแสดงดังตารางที่ 7

ตารางที่ 7 การตรวจสอบความเหมาะสมของตัวแบบการถดถอยลอจิสติก

วิธีการ	องศาเสรี	ค่าไคกำลังสอง	ค่าพี
การทดสอบอัตราส่วนภาวะน่าจะเป็น	1	256.29	0.00
ฮอสเมอร์และลีเมสโชว์	8	8.46	0.39

การตรวจสอบความถูกต้องของตัวแบบการถดถอยลอจิสติก

ตรวจสอบความถูกต้องโดยพิจารณาจากค่าสถิติ R^2 ของค็อกซ์และสเนล (Cox and Snell) และนาเจลเคอร์คี (Nagelkerki) แสดงร้อยละความผันแปรของตัวแปรตามที่อธิบายได้ด้วยตัวแปรอิสระ พบว่าตัวแปรอิสระสามารถใช้อธิบายตัวแปรตามอยู่ในช่วงระหว่างร้อยละ 12.85 ถึง 17.60 รายละเอียดแสดงดังตารางที่ 8

ตารางที่ 8 การทดสอบความถูกต้องของตัวแบบการถดถอยลอจิสติก

วิธีการ	R^2
ค็อกซ์และสเนล	0.1285
นาเจลเคอร์คิ	0.1760

การพยากรณ์หน่วยใหม่

การพยากรณ์หน่วยใหม่ ทำได้โดยนำข้อมูลในส่วนที่ 2 หรือข้อมูลทดสอบมาพยากรณ์ว่าจะเกิดผู้เอาประกันภัยจะขอยกเลิกกรมธรรม์หรือไม่ กำหนดให้

$P(Y=1) < 0.5$ จะพยากรณ์ว่าไม่เกิดการขาดอายุกรมธรรม์

$P(Y=1) \geq 0.5$ จะพยากรณ์ว่าเกิดการขาดอายุกรมธรรม์

จากตารางที่ 9 พบว่าตัวแบบพยากรณ์นี้ พยากรณ์จำนวนผู้เอาประกันภัยที่ขาดอายุกรมธรรม์จำนวน 148 คน คิดเป็น ร้อยละ 31.76 และผู้เอาประกันภัยไม่ขาดอายุกรมธรรม์จำนวน 318 คน คิดเป็นร้อยละ 68.24 ซึ่งพยากรณ์ได้ถูกต้อง $\left[(81+231)/466\right]\times 100 = 66.95\%$

ตารางที่ 9 การพยากรณ์การขาดอายุกรมธรรม์

ค่าสังเกต	ค่าพยากรณ์		
	ขาดอายุกรมธรรม์	ไม่ขาดอายุกรมธรรม์	ผลรวม
ขาดอายุกรมธรรม์	81	87	168
ไม่ขาดอายุกรมธรรม์	67	231	298
ผลรวม	148	318	466

วิจารณ์ผลการวิจัย

จากการสร้างตัวแบบพยากรณ์การขาดอายุกรมธรรม์ด้วยตัวแบบการถดถอยลอจิสติก โดยมีตัวแปรอิสระที่เหมาะสมทั้งหมด 6 ตัว ได้แก่ อายุ จำนวนเงินเอาประกันชีวิตระหว่าง 50,001 - 100,000 บาท ระยะเวลาการชำระเบี้ยประกันภัยมากกว่า 3 ปี รากที่สองของระยะเวลาความคุ้มครอง และระดับชั้นอาชีพแบบอาชีพชั้น 3และอาชีพชั้น 4 ซึ่งมีความสอดคล้องกับงานวิจัยของ Xu (2015) สำหรับตัวแปรอิสระที่คล้ายคลึงกัน โดยตัวแบบพยากรณ์สามารถพยากรณ์การขาดอายุกรมธรรม์ได้ถูกต้องร้อยละ 66.95 ซึ่งอาจมีความถูกต้องมากขึ้นโดยการใช้ตัวแปรอิสระอื่น ๆ ที่เกี่ยวข้องกับการขาดอายุกรมธรรม์ในการสร้างตัวแบบพยากรณ์ เช่น ข้อมูลการตรวจสุขภาพของผู้เอาประกันภัย อัตราดอกเบี้ย รวมถึงปัจจัยทางด้านเศรษฐศาสตร์ เป็นต้น

สรุปผลการวิจัย

จากข้อมูลพื้นฐานของผู้เอาประกันภัย พบว่าร้อยละ 36.14 ขาดอายุกรมธรรม์ โดยส่วนใหญ่เป็นเพศหญิง จำนวน 452 คน คิดเป็นร้อยละ 19.40 จำนวนเงินเอาประกันชีวิตอยู่ในช่วงน้อยกว่าหรือเท่ากับ 50,000 บาท จำนวน 465 คน คิดเป็นร้อยละ 19.96 ระยะเวลาการชำระเบี้ยประกันภัยอยู่ในช่วงระหว่าง 1-3 ปี จำนวน 629 คน คิดเป็นร้อยละ 27.00 รูปแบบการชำระเบี้ยประกันภัยแบบรายเดือนจำนวน 818 คน คิดเป็นร้อยละ 35.11 รูปแบบกรมธรรม์แบบสะสมทรัพย์ จำนวน 838 คน คิดเป็นร้อยละ 35.97 ผ่านช่องทางจำหน่ายทางโทรศัพท์ จำนวน 842 คน คิดเป็นร้อยละ 36.14 เบี้ยประกันภัยน้อยกว่าหรือเท่ากับ 10,000 บาท จำนวน 842 คน คิดเป็นร้อยละ 36.14 และระดับชั้นอาชีพชั้นที่ 1 จำนวน 328 คน คิดเป็นร้อยละ 14.08

จากการคัดเลือกตัวแปรอิสระ และนำตัวแปรอิสระมาสร้างตัวแบบพยากรณ์การขาดอายุกรมธรรม์ ได้ตัวแบบดังนี้

$$Logit = 1.6004 - 0.0316X_1 - 0.2396X_{2A} - 0.8553X_3 - 0.2785X_4' + 0.4924X_{6C} + 1.0946X_{6D}$$

ซึ่งตัวแบบที่ได้พยากรณ์ว่าผู้เอาประกันภัยเกิดการขาดอายุกรมธรรม์จำนวน 148 คน คิดเป็นร้อยละ 31.76 และมีความสามารถในการพยากรณ์ชุดข้อมูลทดสอบได้ถูกต้องร้อยละ 66.95

กิตติกรรมประกาศ

การศึกษาสำเร็จลุล่วงไปได้ด้วยความกรุณาและการอนุเคราะห์จากบริษัทประกันชีวิต โครงการพัฒนากำลังคนด้านวิทยาศาสตร์ (ทุนเรียนดีวิทยาศาสตร์แห่งประเทศไทย) รวมทั้งภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ ที่เอื้อเฟื้อสถานที่ให้ใช้ทรัพยากรต่าง ๆ ในการวิเคราะห์ข้อมูลจนสำเร็จ

เอกสารอ้างอิง

Ae, H.P. (2013). An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain. *Journal of Korean Academy of Nursing*, *43*(2), 154-164.

Allison, D.P. (2014). Measures of Fit for Logistic Regression. *SAS Global Forum,* 1-12.

Hilbe, M.J. (2009). *Logistic Regression Models.* (1). United States of America: CRC Press.

Josephat, K.P. & Ame, A. (2018). Effect on Testing Logistic Regression Assumption on the Improvement of the Propensity Scores. *International Journal of Statistics and Applications, 8*(1), 9-17.

Kaiyawan, Y. (2012). Principle and Using Logistic Regression Analysis for Research. *Journal of Rajamongala University of Technology Srivijaya, 4*(1), 1-12. (in Thai)

Kleinbaum, G.D. & Klein, M. (2002) *Logistic Regression: A Self-Learning Text*. (2). United States of America: Springer.

Office of Insurance Commission. (2016). *Life insurance*. Retrieved August 1, 2018, from http://www.oic.or.th/en/consumer/insurance/about/life.

Peng, J., Lee L.K. & Ingersoll M.G. (2002). An Introduction to Logistic Regression Analysis and Reporting. *Journal of Educational Research, 96*(1), 3-14.

Sinsomboonthong, S. (2016). *Multivariate Analysis.* (1). Bangkok: Chulalongkorn University Press. (in Thai)

Tabachnick, G.B. & Fidell, S.L. (2013). *Using Multivariate Statistics*. (6). United States of America: Courier Companies.

Thai Life Assurance Association, (2018). Overview of Thai Life Assurance business in 2017 and Trend of Thai Life Assurance business in 2018. *Journal of Life Assurance*, *38*(1), 1-32. (In Thai)

Xu, R., Lai, D., Cao, M., Rushing, S. & Rozar, T. (2015). Lapse Modeling for the Post-Level Period: A Practical Application of Predictive Modeling. *Society of Actuaries,* 1-22.

Zian, J., Miller, A. & Ducuroir, F. (2016). Lapse Rate Models in Life Insurance and a Practical Method to Foresee Interest Rate Dependencies. *A Reacfin White Paper on Life Insurance*, 1-20.

ภาคผนวก

ตารางภาคผนวกที่ 1 ข้อมูลทั่วไปของผู้เอาประกันภัย

ข้อมูล	ตัวแปร	ขาดอายุกรมธรรม์ จำนวน	(ร้อยละ)	ไม่ขาดอายุกรมธรรม์ จำนวน	(ร้อยละ)	รวม	(ร้อยละ)
อายุ	X_1	842	(36.14)	1,488	(63.86)	2,330	(100.00)
<u>จำนวนเงินเอาประกันชีวิต (บาท)</u>							
1) น้อยกว่าหรือเท่ากับ 50,000	กลุ่มอ้างอิง	465	(19.96)	744	(31.93)	1,209	(51.89)
2) 50,001 - 100,000	X_{2A}	109	(4.68)	354	(15.19)	463	(19.87)
3) มากกว่า 100,000	X_{2B}	268	(11.50)	390	(16.74)	658	(28.24)
<u>ระยะเวลาการชำระเบี้ยประกันภัย (ปี)</u>							
1) 1 - 3	กลุ่มอ้างอิง	629	(27.00)	629	(27.00)	1,258	(53.99)
2) มากกว่า 3	X_3	213	(9.14)	859	(36.87)	1,072	(46.01)
ระยะเวลาความคุ้มครอง (ปี)	X_4	842	(36.14)	1,488	(63.86)	2,330	(100.00)
<u>รูปแบบการชำระเบี้ยประกันภัย</u>							
1) ชำระครั้งเดียว	กลุ่มอ้างอิง	0	(0.00)	67	(2.88)	67	(2.88)
2) แบบรายเดือน	X_{5A}	818	(35.11)	1,075	(46.14)	1,893	(81.24)
3) แบบรายไตรมาส	X_{5B}	24	(1.03)	28	(1.20)	52	(2.23)
4) แบบรายครึ่งปี	X_{5C}	0	(0.00)	163	(7.00)	163	(7.00)
5) แบบรายปี	X_{5D}	0	(0.00)	155	(6.65)	155	(6.65)
<u>ระดับชั้นอาชีพ</u>							
1) ไม่ระบุ	กลุ่มอ้างอิง	274	(11.76)	754	(32.36)	1,028	(44.12)

ข้อมูล	ตัวแปร	ขาดอายุกรมธรรม์ จำนวน	(ร้อยละ)	ไม่ขาดอายุกรมธรรม์ จำนวน	(ร้อยละ)	รวม	(ร้อยละ)
2) อาชีพชั้น 1	X_{6A}	328	(14.08)	474	(20.34)	802	(34.42)
3) อาชีพชั้น 2	X_{6B}	146	(6.27)	190	(8.15)	336	(14.42)
4) อาชีพชั้น 3	X_{6C}	76	(3.26)	63	(2.70)	139	(5.97)
5) อาชีพชั้น 4	X_{6D}	18	(0.77)	7	(0.30)	25	(1.07)
<u>เพศ</u>							
1) ชาย	กลุ่มอ้างอิง	390	(16.74)	666	(28.58)	1,056	(45.32)
2) หญิง	X_7	452	(19.40)	822	(35.28)	1,274	(54.68)
<u>เบี้ยประกันภัย (บาท)</u>							
1) น้อยกว่าหรือเท่ากับ 10,000	กลุ่มอ้างอิง	842	(36.14)	1,329	(57.04)	2,171	(93.18)
2) มากกว่า 10,000	X_8	0	(0.00)	159	(6.82)	159	(6.82)
<u>ช่องทางการจำหน่าย</u>							
1) ทางโทรศัพท์	กลุ่มอ้างอิง	842	(36.14)	1,412	(60.60)	2,254	(96.74)
2) ทางบริษัท	X_9	0	(0.00)	76	(3.26)	76	(3.26)
<u>รูปแบบกรมธรรม์</u>							
1) แบบตลอดชีพ	กลุ่มอ้างอิง	1	(0.04)	4	(0.17)	5	(0.21)
2) แบบชั่วระยะเวลา	X_{10A}	3	(0.13)	3	(0.13)	6	(0.26)
3) แบบสะสมทรัพย์	X_{10B}	838	(35.97)	1,481	(63.56)	2,319	(99.53)
รวม		842	(36.14)	1,488	(63.86)	2,330	(100.00)

<u>หมายเหตุ</u> อาชีพในการรับประกันภัยจะแบ่งเป็น 4 ชั้น ได้แก่

อาชีพชั้น 1 ลักษณะงานส่วนใหญ่ทำงานประจำในสำนักงาน และทำงานที่ไม่ได้ใช้เครื่องจักร เช่น ผู้บริหาร พนักงานบริษัท แพทย์ เภสัชกร พยาบาล ข้าราชการ เป็นต้น

อาชีพชั้น 2 ลักษณะงานที่ทำส่วนใหญ่อยู่นอกสำนักงาน หรือต้องทำงานกลางแจ้งตลอดเวลา เป็นกลุ่มช่างฝีมือที่มีความชำนาญและทักษะ บางครั้งอาจจะมีการใช้เครื่องจักร เช่น ตัวแทน/นายหน้า วิศวกร ช่างไม้ เจ้าของกิจการขนาดเล็ก เป็นต้น

อาชีพชั้น 3 ลักษณะงานที่ทำส่วนใหญ่มีการใช้เครื่องจักรกลหนัก หรือเป็นผู้ใช้แรงงาน หรือทำงานนอกสำนักงานเป็นประจำ เช่น ผู้ที่ปฏิบัติงานด้านช่าง ด้านการผลิต การขนส่ง พนักงานขาย นักแสดง มัคคุเทศก์ นักข่าว พนักงานขับรถ เป็นต้น

อาชีพชั้น 4 ลักษณะงานที่มีโอกาสเกิดอุบัติเหตุมากที่สุด มีความเสี่ยงสูงมากกว่าอาชีพชั้นอื่นๆ เป็นพิเศษ เช่น นักแสดงผาดโผน คนงานก่อสร้าง พนักงานรักษาความปลอดภัย พนักงานรับส่งเอกสาร เป็นต้น