บทที่ 2

วิธีการทางสถิติและงานวิจัยที่เกี่ยวข้อง

ในปัจจุบันการวิจัยทางด้านต่างๆ ได้มีการนำเอาวิธีการทางสถิติเข้ามาเพื่อช่วยในการหาผลสรุปของการวิจัย ไม่ว่าจะเป็นการวิจัยทางด้านการแพทย์ วิศวกรรม หรืออุตสาหกรรม โดยวิธีส่วนใหญ่ที่ศึกษา จะเป็นการศึกษาเกี่ยวกับเวลาเริ่มต้นศึกษาสิ่งที่สนใจไปจนกระทั่งเหตุการณ์ที่สนใจเกิดขึ้น หรือที่เรียกว่าเวลาการคงอยู่ (Survival time: T) ยกตัวอย่าง กรณีศึกษาเกี่ยวกับระยะเวลาการสิ้นผลบังคับลงของกรมธรรม์ประกันชีวิต เพื่อพิจารณาปัจจัยต่างๆ ที่คาดว่าจะมีผลต่อระยะเวลาการคงอยู่ของกรมธรรม์ประกันชีวิต ในที่นี้กรมธรรม์ประกันชีวิตที่ขาดอายุลงคือ ตัวแปรที่สนใจศึกษา ระยะเวลาการคงอยู่ของกรมธรรม์จนกระทั่งสิ้นผลบังคับลงคือ ค่าของตัวแปรที่สนใจศึกษา โดยการวิเคราะห์ข้อมูลการคงอยู่เหล่านี้ จะเกี่ยวกับช่วงเวลาการคงอยู่ (life time) และช่วงเวลาล้มเหลว (failure time) ซึ่งอาจพบกับประเภทของข้อมูลต่างๆ อันเนื่องมาจากข้อจำกัดของเวลา งบประมาณในการศึกษา และข้อจำกัดอื่นๆ จากการวิเคราะห์การคงอยู่ของข้อมูล (Survival analysis of data)

ในบทนี้จะกล่าวถึงประเภทของข้อมูลที่มีค่าไม่สมบูรณ์ หรือมีค่าถูกตัดทิ้ง (Type of data censoring or truncation) ฟังก์ชันการคงอยู่ (Survival function) ฟังก์ชันความหนาแน่น (Density function) ฟังก์ชันภาวะภัย (Hazard function) วิธีประมาณฟังก์ชันการคงอยู่ การทดสอบสมมติฐาน และตัวแบบเชิงพหุของข้อมูลการคงอยู่ต่างๆ ที่เกี่ยวข้อง

2.1 ประเภทของข้อมูลที่มีค่าสังเกตไม่สมบูรณ์หรือมีค่าถูกตัดทิ้ง

1. ค่าสังเกตไม่สมบูรณ์ประเภทที่ 1 (Type I censoring)

เป็นลักษณะของข้อมูลที่มีการกำหนดเวลาของการเกิดค่าสังเกตเอาไว้ล่วงหน้า ตัวอย่างของค่าสังเกตที่ไม่สมบูรณ์ประเภทที่ 1 เช่น การทำประกันชีวิตแบบชั่วระยะเวลา เป็นการทำประกันชีวิตที่มีการกำหนดระยะเวลาที่แน่นอนของสัญญาประกันชีวิต โดยกำหนดระยะเวลาการคุ้มครองไว้ 5 ปี ถ้าผู้เอาประกันภัยเสียชีวิตภายในระยะเวลาที่สัญญายังมีผลคุ้มครองอยู่ ก็จะได้รับความคุ้มครองจากการเสียชีวิต ซึ่งค่าสังเกตในลักษณะนี้เป็นค่าสังเกตที่สมบูรณ์ (uncensored data) เพราะทราบเวลาที่ผู้เอาประกันภัยเสียชีวิตที่แน่นอน แต่ในกรณีเมื่อครบกำหนดสัญญาประกันชีวิต ยังมีผู้เอาประกันภัยบางคนที่ยังคงมีชีวิตอยู่ ค่าสังเกตในลักษณะนี้ จะเป็นค่าสังเกตที่ไม่สมบูรณ์ (censored data) เพราะไม่สามารถทราบเวลาที่ผู้เอาประกันภัยจะเสียชีวิตที่แน่นอนได้ หากพ้นจากระยะเวลาคุ้มครองของสัญญาไป (เจษฎา สุทธิอุดม, 2539: 7; Klien and Moeschberger, 1997: 58)

2. ค่าสังเกตไม่สมบูรณ์ประเภทที่ 2 (Type II censoring)

เป็นลักษณะของข้อมูลที่มีการกำหนดจำนวนค่าสังเกตที่สมบูรณ์ไว้ล่วงหน้า นั่นคือ เมื่อจำนวนค่าสังเกตที่สมบูรณ์เกิดขึ้นครบตามจำนวนที่กำหนดไว้ จะหยุดทำการทดลอง ตัวอย่างของค่าสังเกตไม่สมบูรณ์ประเภทที่ 2 เช่น ในการทำประกันชีวิตของผู้เอาประกันภัยจำนวน n คน ต้องการศึกษาระยะเวลาที่ผู้เอาประกันภัยจะเสียชีวิต โดยข้อมูลที่เก็บคือ ระยะเวลาที่ผู้เอาประกันภัยจะเสียชีวิตจำนวน r คน (เมื่อ $r \leq n$) และเมื่อเก็บข้อมูลครบตามที่กำหนดไว้ล่วงหน้า จึงหยุดการเก็บข้อมูล (เจษฎา สุทธิอุดม, 2539: 8; Klien et al., 1997: 58)

3. ค่าสังเกตไม่สมบูรณ์แบบสุ่ม (Random censoring)

เป็นลักษณะของค่าสังเกตที่มีการตัดข้อมูลในระหว่างการทดลอง และหลังสิ้นสุดการทดลอง ซึ่งเกิดจากหน่วยสังเกตประสบกับเหตุการณ์อื่นที่อยู่นอกเหนือภายใต้การศึกษา หรือค่าสังเกตมีความไม่สมบูรณ์ คือ มีค่าไม่เท่ากับค่าสังเกตใดๆ เลย ซึ่งกรณีนี้อาจเกิดจากการที่หน่วยสังเกตขาดการติดต่อ หรือหน่วยสังเกตถอนตัวจากการทดสอบ เช่น ในงานด้านประกันชีวิต ต้องการศึกษาถึงระยะเวลาการคงอยู่ของกรมธรรม์ประกันชีวิตในช่วงระยะเวลา 5 ปี ซึ่งอาจจะมีผู้เอาประกันภัยบางคนเสียชีวิตภายในระยะเวลาที่กำหนด ค่าสังเกตในลักษณะนี้ จะเป็นค่าสังเกตที่สมบูรณ์ แต่เมื่อมีกรณีที่ผู้เอาประกันภัยบางคนมีการเวนคืนเงินสด หรือขาดส่งเบี้ยประกันภัย มีผลทำให้กรมธรรม์ประกันชีวิตขาดอายุลง ค่าสังเกตในลักษณะนี้ จะเป็นค่าสังเกตไม่สมบูรณ์ (Klien et al., 1997: 61)

4. ค่าสังเกตไม่สมบูรณ์แบบอื่น (Other type of censoring)

นอกจากค่าสังเกตไม่สมบูรณ์ประเภทที่ 1 ค่าสังเกตไม่สมบูรณ์ประเภทที่ 2 และค่าสังเกตไม่สมบูรณ์แบบสุ่ม ซึ่งเป็นข้อมูลไม่สมบูรณ์ทางขวาแล้ว ยังมีค่าสังเกตไม่สมบูรณ์ทางซ้าย (left censoring) ค่าสังเกตไม่สมบูรณ์ทางซ้ายและทางขวา (doubly censoring) และค่าสังเกตไม่สมบูรณ์ที่เป็นช่วง (interval censoring) ยกตัวอย่างข้อมูลที่ไม่สมบูรณ์ทางซ้าย เช่น ในการรับประกันภัยรถยนต์ จะมีการกำหนดความเสียหายส่วนแรก (deductible) ที่ผู้เอาประกันภัยจะต้องรับผิดชอบ มีผลให้ข้อมูลการจ่ายค่าสินไหมทดแทนที่บริษัทเก็บไว้ จะเป็นข้อมูลเฉพาะกรณีค่าเสียหายที่เกินมากกว่าค่าเสียหายส่วนแรก และถ้าบริษัทประกันภัยนี้ได้ส่งต่อให้กับบริษัทรับประกันภัยต่อ ซึ่งบริษัทประกันภัยจะกำหนดความรับผิดสูงสุดของบริษัทเอาไว้ ความสูญเสียส่วนที่เกินจากที่กำหนดนี้บริษัทรับประกันภัยต่อจะเป็นผู้รับผิดชอบ ดังนั้น ข้อมูลที่บริษัทประกันภัยมีอยู่ จึงเป็นข้อมูลความสูญเสียที่มากกว่าความเสียหายส่วนแรก และไม่เกินจำนวนความรับผิดสูงสุดที่บริษัทประกันภัยกำหนดเอาไว้

นอกเหนือจากลักษณะของข้อมูลที่มีค่าสังเกตไม่สมบูรณ์แล้ว ข้อมูลการคงอยู่ยังมีลักษณะอีกประเภทหนึ่ง คือ การตัดออก (truncation) การตัดออกจะเป็นเหมือนกับการกำหนดเงื่อนไขในการคัดเลือกหน่วยสังเกตที่จะตกอยู่ในกลุ่มตัวอย่าง ว่าจะต้องมีเหตุการณ์ตามเงื่อนไขที่กำหนดไว้ก่อนเหตุการณ์ที่สนใจ ซึ่งแบ่งได้เป็น

1. การตัดออกทางซ้าย (Left truncation)

เหตุการณ์สำคัญของสิ่งที่สนใจที่กล่าวว่าเป็นการตัดออกทางซ้ายคือ ถ้า Y เป็นเวลาของเหตุการณ์ที่ค่าสังเกตถูกตัดออก ดังนั้น สำหรับข้อมูลที่มีการตัดออกทางซ้ายจะมีเพียงข้อมูลที่ $t \geq Y$ เท่านั้นที่ถูกสังเกต เมื่อ t คือ ช่วงเวลาการคงอยู่ การตัดออกทางซ้ายส่วนมาก จะเกิดขึ้นเมื่อตัวอย่างที่เข้าไปในการศึกษาที่อายุสุ่ม (ไม่จำเป็นว่าต้องเป็นสาเหตุที่ทำให้เหตุการณ์ที่สนใจเกิดขึ้น) และติดตามจากเวลาที่เข้ามาในกลุ่มการศึกษาจนกระทั่งเกิดเหตุการณ์ที่สนใจ หรือกระทั่งข้อมูลเป็นค่าสังเกตไม่สมบูรณ์ทางขวา ยกตัวอย่างข้อมูลที่ถูกตัดออกทางซ้าย เช่น การศึกษาเกี่ยวกับสาเหตุการสิ้นผลบังคับลงของกรมธรรม์ประกันชีวิตตั้งแต่ปีพ.ศ. 2545 เป็นต้นไป กรมธรรม์ประกันชีวิตที่จะเข้ามาเป็นหน่วยสังเกตภายใต้การศึกษาได้ จึงต้องมีเงื่อนไขที่ว่า กรมธรรม์ประกันชีวิตจะต้องมีผลบังคับอยู่จนถึงอย่างน้อย ณ สิ้นปีพ.ศ. 2544 จึงจะมาสิ้นผลบังคับลงตั้งแต่ปีพ.ศ. 2545 ได้ เป็นต้น ซึ่งจากตัวอย่างนี้จะเห็นว่าทุกๆ หน่วยสังเกตจะต้องมีเงื่อนไขที่สนใจก่อนเข้ามาในกลุ่มการศึกษา

2. การตัดออกทางขวา (Right truncation)

เกิดขึ้นเมื่อมีเพียงหน่วยสังเกตที่เคยประสบกับเหตุการณ์เท่านั้น ที่รวมอยู่ในกลุ่มศึกษา แต่มีหน่วยสังเกตอื่นๆ ที่อาจจะประสบกับเหตุการณ์ที่สนใจเหมือนกัน แต่ไม่ได้ตกอยู่ในกลุ่มการศึกษา ตัวอย่างเช่น ต้องการศึกษาเกี่ยวกับการเสียชีวิตของผู้ป่วยโรคเอดส์ ซึ่งเมื่อทำการเก็บรวบรวมข้อมูล ก็จะมีหน่วยสังเกตที่เป็นผู้ป่วยโรคเอดส์เท่านั้นที่รวมอยู่ในกลุ่มการศึกษา แต่ก็อาจมีผู้ป่วยบางคนที่อยู่ในระยะการติดเชื้อ HIV+ ซึ่งอาจจะประสบกับเหตุการณ์ที่สนใจเหมือนกัน แต่ไม่ตกอยู่ในกลุ่มการศึกษาครั้งนี้ด้วย

3. การตัดออกทางซ้ายและทางขวา (Left and right truncation)

เกิดจากกรณีที่ข้อมูลมีการตัดออกทั้งทางซ้ายและทางขวา ตัวอย่างเช่น การศึกษาผู้ป่วยด้วยโรคเอดส์ในผู้สูงอายุ ซึ่งมีเงื่อนไขว่า หน่วยสังเกตที่จะตกอยู่ในกลุ่มการศึกษาได้นั้น ต้องมีอายุมากกว่า 60 ปีขึ้นไป แต่ก็อาจมีผู้สูงอายุบางคนที่อยู่ในระยะการติดเชื้อ HIV+ ที่ไม่ตกอยู่ในกลุ่มการศึกษาครั้งนี้ด้วย (Klien et al., 1997: 65)

2.2 ฟังก์ชันที่เกี่ยวข้อง

ในการอธิบายการแจกแจงการคงอยู่นั้น การทราบฟังก์ชันของเวลาการคงอยู่ ซึ่งได้แก่ ฟังก์ชันการคงอยู่ ฟังก์ชันความหนาแน่น และฟังก์ชันภาวะภัย จะสามารถบอกให้ทราบถึงตัวแบบการแจกแจงของข้อมูลการคงอยู่ได้ นอกจากนี้แล้ว การทราบฟังก์ชันใดฟังก์ชันหนึ่งใน 3 ฟังก์ชัน ก็จะสามารถหาฟังก์ชันที่เหลือได้จากความสัมพันธ์ของฟังก์ชัน

การวิเคราะห์การคงอยู่

ให้ตัวแปรสุ่ม T แทนระยะเวลาที่กรมธรรม์สิ้นผลบังคับลง (โดยใช้หน่วยวัดเป็นปี) ในการประมาณฟังก์ชันการแจกแจงสะสม (Cumulative distribution function) หรือ F(t) ของ T สามารถทำได้ 2 วิธี คือ วิธีอิงพารามิเตอร์ (Parametric method) และวิธีไม่อิงพารามิเตอร์ (Nonparametric method) รวมทั้งฟังก์ชันการแจกแจงสะสมนี้ อาจขึ้นอยู่กับตัวแปรอิสระอื่นๆ (Z_i โดยที่ $i = 1,2,\ldots,n$) เช่น ขึ้นอยู่กับอายุที่เริ่มต้นทำประกันชีวิต เพศผู้เอาประกันภัย แบบกรมธรรม์ประกันชีวิต เป็นต้น

(1) ฟังก์ชันการคงอยู่ (Survival function : S(t)) ฟังก์ชันการคงอยู่นี้ คือ ฟังก์ชันที่แสดงค่าความน่าจะเป็นที่หน่วยสังเกต จะมีการคงอยู่นานกว่าหรือเท่ากับเวลา t นั่นคือ

$$S(t) = \Pr ob(T \geq t)$$

เมื่อ T คือ ตัวแปรสุ่มแบบต่อเนื่อง ฟังก์ชันการคงอยู่จะเป็นส่วนเติมเต็ม (complement) ของฟังก์ชันการแจกแจงสะสม นั่นคือ

$$= 1 - F(t)$$

$$= \int_t^{\infty} f(x)\,dx$$

เมื่อ S(t) ขึ้นอยู่กับตัวแปรอิสระ Z_i สามารถเขียน S(t) ได้เป็น $S(t,z_1,z_2\ldots,z_n)$

คุณสมบัติของฟังก์ชันการคงอยู่

ก) S(t) เป็นฟังก์ชันไม่เพิ่ม

ข) S(t) เป็นฟังก์ชันต่อเนื่อง

ค) S(t) = 1 เมื่อ t = 0

ง) S(t) = 0 เมื่อ t เข้าสู่อนันต์

(2) ฟังก์ชันความหนาแน่น (Density function : f(t)) เมื่อตัวแปรสุ่ม T เป็นตัวแปรสุ่มแบบต่อเนื่อง ดังนั้น ฟังก์ชันความหนาแน่นของ T จึงมีค่าเท่ากับลิมิตของความน่าจะเป็นที่หน่วยสังเกตจะสูญเสียในช่วงเวลาสั้นๆ (Δt) หรือความน่าจะเป็นของการสูญเสียในช่วงเวลา t ถึง t +Δt ต่อหน่วยเวลา Δt นั่นคือ (วิชัย มหัตเคชกุล, 2535: 8)

$$f(t) = \lim_{\Delta t \to 0} \mathrm{Pr}ob\,\frac{(t < T \le t + \Delta t)}{\Delta t}$$

$$= \frac{dF(t)}{dt}$$

$$= \frac{-dS(t)}{dt}$$

คุณสมบัติของฟังก์ชันความหนาแน่น

ก) $f(t) \ge 0$ เมื่อ $t \ge 0$ และ $f(t) = 0$ เมื่อ $t < 0$

ข) $\int_0^\infty f(t)dt = 1$

(3) ฟังก์ชันภาวะภัย (Hazard function : h(t)) มีค่าเท่ากับลิมิตของความน่าจะเป็นที่หน่วยสังเกตจะสูญเสียในช่วงเวลาสั้นๆ จาก t ถึง t +Δt ต่อหน่วยเวลา Δt เมื่อหน่วยสังเกตจะคงอยู่ถึงเวลา t นั่นคือ (Parmer and Machin, 1995)

$$h(t) = \lim_{\Delta t \to 0} \frac{\mathrm{Pr}ob[t < T \le t + \Delta t | T \ge t]}{\Delta t}$$

$$= f(t)/(1 - F(t))$$

$$= f(t)/S(t)$$

ฟังก์ชันภาวะภัยเป็นฟังก์ชันที่แสดงถึงความน่าจะเป็นที่หน่วยสังเกตนั้น จะประสบกับเหตุการณ์ที่สนใจขึ้นอย่างทันทีทันใด ณ เวลา t และเนื่องจากเวลาเป็นตัวแปรแบบตัวเนื่อง ความน่าจะเป็นที่เหตุการณ์นั้นจะเกิดขึ้นที่เวลา t พอดี จึงมีค่าเท่ากับศูนย์ การอธิบายความน่าจะเป็นที่เหตุการณ์จะเกิดขึ้น จึงแสดงในช่วงเวลาสั้นๆ (t ,t +Δt)

คุณสมบัติของฟังก์ชันภาวะภัย

ก) $h(t) \ge 0$

ข) $\lim_{t \to \infty} \int_0^t h(t)dt = \infty$

2.3 การประมาณฟังก์ชันการคงอยู่

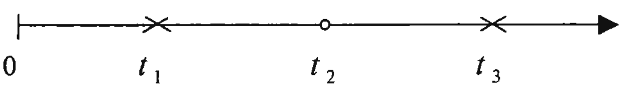
เนื่องจากข้อมูลที่เกี่ยวกับการวิเคราะห์ช่วงเวลาการคงอยู่มักไม่ทราบการแจกแจง ดังนั้น ตัวสถิติที่เหมาะสมกับการนำมาใช้ส่วนใหญ่ จึงเป็นตัวสถิติไม่อิงพารามิเตอร์ เนื่องจากสามารถเข้าใจและนำไปใช้ได้ง่ายเมื่อข้อมูลไม่เป็นไปตามทฤษฎี ในที่นี้ จะอธิบายเฉพาะวิธีการประมาณฟังก์ชันการคงอยู่โดยใช้ตัวสถิติแบบไม่อิงพารามิเตอร์ด้วยวิธีลิมิตผลคูณ (Product-Limit Method)

วิธีลิมิตผลคูณ (Product-Limit Method : PL)

วิธีลิมิตผลคูณ เป็นการประมาณฟังก์ชันการคงอยู่ที่สามารถจัดการได้ทั้งกับข้อมูลที่มีค่าสังเกตสมบูรณ์ และไม่สมบูรณ์ ที่ไม่ได้มีการจัดกลุ่ม มีค่าสังเกตซ้ำเป็นจำนวนไม่มาก และค่าสังเกตมีค่าไม่ต่อเนื่อง สำหรับค่าสังเกตที่มีค่าต่อเนื่อง สามารถทำให้เป็นค่าไม่ต่อเนื่องได้โดยการจัดกลุ่มให้เป็นช่วง (สากันย์ สุวรรณการ, 2530: 17)

นิยามและวิธีการคำนวณ

ในทางปฏิบัติสามารถคำนวณฟังก์ชันการคงอยู่ได้ โดยสมมติว่าค่าสังเกตของเวลาการคงอยู่จำนวน n มีค่าสังเกตเป็น $t_1 < t_2 < \ldots < t_D$



โดยที่เครื่องหมาย x หมายถึง ค่าสังเกตสมบูรณ์

0 หมายถึง ค่าสังเกตไม่สมบูรณ์

(ก) สำหรับแต่ละช่วง (t_{i-1}, t_i) $\quad p_i = \Pr ob(T > t_i | T > t_{i-1})$

(ข) ถ้า t เป็นเวลาที่กำหนด S(t) จะประมาณจากผลคูณของค่าประมาณ p_i ที่ทุกเวลาก่อนถึงเวลา t

$$S(t) = p_1 \times p_2 \times \ldots \times p_t$$

$$= \prod_{i \le t} p_i$$

ค่าประมาณด้วยวิธีลิมิตผลคูณของ S(t) คือ

$$\hat{S}(t) = \begin{cases} 1 & t < t_1 \\ \prod_{t_i \le t} [1 - \frac{d_i}{n_i}] & t_1 \le t \end{cases}$$

เมื่อ n_i คือ จำนวนหน่วยสังเกตหลังเวลา t_{i-1} ที่ยังคงอยู่ ณ เวลา t_i

d_i คือ จำนวนหน่วยสังเกตที่สูญเสีย ณ เวลา t_i

ตัวประมาณลิมิตผลคูณ (Product-Limit Estimator) จะเป็นฟังก์ชันขั้นบันได (step function) ที่มีการกระโดด ณ จุดเวลาที่เหตุการณ์นั้นถูกสังเกต ขนาดของการกระโดดเหล่านี้ จะขึ้นอยู่กับจำนวนเหตุการณ์ที่สังเกตได้ในแต่ละค่าสังเกต t_i และรูปแบบของค่าสังเกตไม่สมบูรณ์ก่อนเวลา t_i

ความแปรปรวนของตัวประมาณลิมิตผลคูณ สามารถประมาณได้โดยใช้ Greenwood's formular

$$\hat{Var}[\hat{S}(t)] = \hat{S}(t)^2 \sum_{t_i \le t} \frac{d_i}{n_i(n_i - d_i)}$$

การประมาณค่าเฉลี่ยและค่ามัธยฐานการคงอยู่โดยใช้ตัวประมาณลิมิตผลคูณ

ค่าเฉลี่ยการคงอยู่สามารถคำนวณได้จาก

$$\hat{\mu} = \sum_{i=0}^{D-1} \hat{S}(t_i)(t_{i+1} - t_i)$$

เมื่อ $t_1 < t_2 < \dots < t_D$ คือ ข้อมูลอันดับของค่าสังเกตสมบูรณ์จำนวน D ค่า และค่าสังเกตที่ใหญ่ที่สุดคือ t_D นั่นคือ $\hat{S}(t_D) = 0$

ค่าประมาณความแปรปรวนของค่าเฉลี่ยการคงอยู่ (Lee, 1980: 86)

$$\hat{Var}(\hat{\mu}) = \sum_{t_r \le t} \frac{A_r^2}{(n-r)(n-r+1)}$$

โดยที่ r คือ เลขจำนวนเต็มสำหรับ t_r เมื่อเหตุการณ์ที่สนใจเกิดขึ้น

เมื่อ $A_r = \int_{t_r}^{\infty} \hat{S}(u)du = \sum_{i=r}^{D} \hat{S}(t_i)(t_{i+1} - t_i)$ สำหรับกรณีไม่มีค่าสังเกตไม่สมบูรณ์

$A_r = \sum_{i=r+1}^{n} (t_i - t_r)/n$ และค่าความแปรปรวนจะลดรูปลงเหลือ $\hat{Var}(\hat{\mu}) = \dfrac{\sum (t_i - \bar{t})^2}{n^2}$

ค่ามัธยฐานการคงอยู่สามารถพิจารณาได้จากเส้นโค้งการคงอยู่ กล่าวคือ ค่าประมาณมัธยฐานของเวลาการคงอยู่ จะมีค่าเท่ากับ t ที่ $\hat{S}(t) = 0.5$ แต่โดยทั่วไปแล้ว จะมีช่วงเวลา (t_i, t_{i+1}) ที่มี $\hat{S}(t) > 0.5$ และ $\hat{S}(t) < 0.5$ ดังนั้น ค่าประมาณมัธยฐานการคงอยู่ คือ

$$\hat{t}_{0.5} = t_i + [(\hat{S}(t_i) - 0.5)/(\hat{S}(t_i) - \hat{S}(t_{i+1}))](t_{i+1} - t_i)$$

และค่าประมาณความแปรปรวนของค่ามัธยฐานการคงอยู่

$$\hat{Var}(\hat{t}_{0.5}) = (t_{i+1} - t_i)^2 / [4n_i(\hat{S}(t_i) - \hat{S}(t_{i+1}))^2]$$

2.4 การทดสอบสมมติฐาน

การทดสอบสมมติฐานเกี่ยวกับการแจกแจงการคงอยู่ของประชากรมากกว่า หรือเท่ากับ 2 ประชากร ในกรณีเกิดและไม่เกิดค่าสังเกตไม่สมบูรณ์นั้น จะถูกพัฒนาขึ้นเพื่อนำไปใช้ในการวิเคราะห์ข้อมูลที่มีการเปลี่ยนแปลงไปตามเวลาในรูปแบบต่างๆ การทดสอบสมมติฐานดังกล่าว มีหลายวิธีที่นำมาใช้ในการพิจารณาความแตกต่างของการแจกแจงการคงอยู่ระหว่างกลุ่มประชากร ซึ่งจะแตกต่างกันไปตามฟังก์ชันถ่วงน้ำหนัก (weight function) ที่กำหนด

โดยทั่วไปแล้ว ฟังก์ชันถ่วงน้ำหนักที่นำมาใช้ทดสอบในโปรแกรมสำเร็จรูปส่วนมาก จะมีอยู่ 2 ฟังก์ชัน คือ ฟังก์ชันที่กำหนดให้ W(t) = 1 ทุกค่า t และฟังก์ชันที่สอง คือ $W(t_i) = n_i$ ในส่วนของฟังก์ชันถ่วงน้ำหนักแบบอื่นๆ อาจกำหนดให้ฟังก์ชันถ่วงน้ำหนักเป็น $W(t_i) = \sqrt{n_i}$ หรือจะเป็นวิธีทดสอบ Mann -Whitney -Wilcoxon โดยเปลี่ยนแปลงค่าประมาณของฟังก์ชันการคงอยู่ให้เป็น

$$\tilde{S}(t) = \prod_{t_i \le t} \left(\frac{n_i + 1 - d_i}{n_i + 1} \right)$$

และน้ำหนักที่ใช้ คือ $W(t_i) = \hat{S}(t_{i-1}) \times \dfrac{n_i}{(n_i + 1)}$

วิธีทดสอบที่เสนอโดย Harrington and Fleming จะใช้ฟังก์ชันถ่วงน้ำหนักเป็น $W(t_i) = \left[\hat{S}(t_{i-1})\right]^{\rho}$ ซึ่งเป็นการรวมลักษณะการทดสอบด้วยวิธี Log Rank และ Mann-Whitney-Wilcoxon เข้าไว้ด้วยกัน โดยถ้า $\rho = 0$ หรือ $W(t_i) = 1$ การทดสอบนี้จะเป็นการทดสอบด้วยวิธี Log Rank และเมื่อ $\rho = 1$ การถ่วงน้ำหนักนี้ จะเป็นเหมือนกับการทดสอบด้วยวิธี Mann-Whitney-Wilcoxon (Klien et al., 1997; Hosmer and Lemeshow, 1998: 61-62)

การวิจัยในครั้งนี้จะกล่าวถึงวิธีทดสอบไม่อิงพารามิเตอร์ 2 วิธี เพื่อทดสอบสมมติฐานเกี่ยวกับการแจกแจงการคงอยู่ของประชากรมากกว่า หรือเท่ากับ 2 ประชากร อันได้แก่ วิธีทดสอบ Gehan's Generalized Wilcoxon และวิธีทดสอบ Log Rank ที่มีอยู่ในโปรแกรมสำเร็จรูป SAS

กำหนดให้ $S_j(t)$ คือ การแจกแจงของฟังก์ชันการคงอยู่ที่มาจากค่าสังเกตของกลุ่มตัวอย่างที่ j

สมมติฐาน $H_0 : S_1(t) = S_2(t) = ... = S_k(t)$

$H_1 : S_j(t) \ne S_g(t)$ เมื่อ $j \ne g$

สำหรับทุกค่า $t \le \tau$ เมื่อ τ คือ เวลามากสุดของการศึกษา และสามารถเขียนฟังก์ชันการคงอยู่ในรูปของฟังก์ชันการแจกแจงสะสมได้ดังนี้

สมมติฐาน $H_0 : F_1(t) = F_2(t) = ... = F_k(t)$

$H_1 : F_j(t) \neq F_g(t)$ เมื่อ $j \neq g$

ในการคำนวณวิธีทดสอบทั้ง 2 วิธี สำหรับ n กลุ่มการทดสอบ ที่ให้ $t_1 < t_2 < ... t_D$ เป็นข้อมูลอันดับของค่าสังเกตสมบูรณ์ที่แตกต่างกัน k กลุ่มตัวอย่าง โดยกำหนดให้

d_{ij} คือ จำนวนค่าสังเกตสมบูรณ์ที่มีค่าเท่ากับ t_i ในตัวอย่างที่ j

$$\text{โดย} \quad d_i = \sum_{j=1}^{k} d_{ij} \qquad \text{เมื่อ } i = 1,2,...D$$

n_{ij} คือ จำนวนค่าสังเกตสมบูรณ์ที่มีค่ามากกว่าหรือเท่ากับ t_i ซึ่งหมายถึง เซตความเสี่ยงที่ค่าสังเกต t_i โดย n_{ij} จะประกอบด้วยค่าสังเกตสมบูรณ์ที่มีค่ามากกว่าหรือเท่ากับ t_i เมื่อ $j = 1,2,...K$ และ $i = 1,2,...D$

$$\text{โดย} \quad n_i = \sum_{j=1}^{k} n_{ij}$$

2.4.1 วิธีทดสอบ Gehan's Generalized Wilcoxon (GH)

วิธีทดสอบ Gehan's Generalized Wilcoxon จะขึ้นอยู่กับค่าสถิติ $Z_j(\tau)$ ซึ่งสามารถคำนวณได้จากสูตร

$$Z_j(\tau) = \sum_{i=1}^{D} n_i \left\{ d_{ij} - n_{ij}\left(\frac{d_i}{n_i}\right) \right\}$$

ความแปรปรวนของ $Z_j(\tau)$ คำนวณได้จาก

$$\hat{Var}[Z_j(\tau)] = \sum_{i=1}^{D} n_i^2 \frac{n_{ij}}{n_i}\left(1 - \frac{n_{ij}}{n_i}\right)\left(\frac{n_i - d_i}{n_i - 1}\right) d_i$$

และความแปรปรวนร่วมของ $Z_j(\tau)$ กับ $Z_g(\tau)$ คือ

$$\hat{Cov}[Z_j(\tau), Z_g(\tau)] = -\sum_{i=1}^{D} n_i^2 \frac{n_{ij}}{n_i} \frac{n_{ig}}{n_i}\left(\frac{n_i - d_i}{n_i - 1}\right) d_i \qquad \text{เมื่อ } j \neq g$$

ตัวสถิติทดสอบ คือ

$$\chi^2 = (Z_1(\tau), Z_2(\tau),..., Z_{k-1}(\tau)) \sum^{-1} (Z_1(\tau), Z_2(\tau),..., Z_{k-1}(\tau))'$$

ซึ่งมีการแจกแจงแบบไคสแควร์ ที่องศาความเป็นอิสระเท่ากับ k-1

2.4.2 วิธีทดสอบ Log Rank

วิธีทดสอบ Log Rank จะมีวิธีการคำนวณที่คล้ายคลึงกับวิธีทดสอบ Gehan's Generalized Wilcoxon โดยมีสูตรการคำนวณเป็น

$$Z_j(\tau) = \sum_{i=1}^{D}\left\{d_{ij} - n_{ij}\left(\frac{d_i}{n_i}\right)\right\}$$

ความแปรปรวนของ $Z_j(\tau)$ คำนวณได้จาก

$$\hat{Var}[Z_j(\tau)] = \sum_{i=1}^{D}\frac{n_{ij}}{n_i}\left(1 - \frac{n_{ij}}{n_i}\right)\left(\frac{n_i - d_i}{n_i - 1}\right)d_i$$

และความแปรปรวนร่วมของ $Z_j(\tau)$ กับ $Z_g(\tau)$ คือ

$$\hat{Cov}[Z_j(\tau), Z_g(\tau)] = -\sum_{i=1}^{D}\frac{n_{ij}}{n_i}\frac{n_{ig}}{n_i}\left(\frac{n_i - d_i}{n_i - 1}\right)d_i \qquad \text{เมื่อ } j \neq g$$

ตัวสถิติทดสอบ คือ

$$\chi^2 = (Z_1(\tau), Z_2(\tau), \ldots, Z_{k-1}(\tau))\sum\nolimits^{-1}(Z_1(\tau), Z_2(\tau), \ldots, Z_{k-1}(\tau))'$$

ซึ่งมีการแจกแจงแบบไคสแควร์ ที่องศาความเป็นอิสระเท่ากับ k-1

2.5 ตัวแบบเชิงพหุของข้อมูลการคงอยู่ (Multivariate Model of Survival Data)

การใช้ตัวสถิติไม่อิงพารามิเตอร์ในส่วนที่ 2.4 สามารถนำมาใช้เปรียบเทียบความแตกต่างระหว่างกลุ่มค่าสังเกตในเหตุการณ์ที่สนใจ (time to event) ได้โดยตรง เมื่อกรณีกลุ่มเหล่านั้นมีลักษณะคล้ายกัน แต่ในทางปฏิบัติแล้ว สิ่งที่สนใจในกลุ่มบางครั้งมีลักษณะที่เพิ่มเข้าไป อันมีอิทธิพลต่อปัญหาที่สนใจ ยกตัวอย่างสิ่งที่สนใจในการวิจัยครั้งนี้ จะประกอบไปด้วยตัวแปรลักษณะต่างๆ ของหน่วยสังเกตที่เก็บบันทึกไว้ ไม่ว่าจะเป็น อายุที่เริ่มทำประกันชีวิต เพศผู้เอาประกันภัย แบบกรมธรรม์ประกันชีวิต หรือจำนวนเงินเอาประกันภัย โดยตัวแปรเหล่านี้เป็นเหมือนตัวแปรอิสระ (independent variable) ในการอธิบายตัวแปรตาม (ระยะเวลาการคงอยู่ของกรมธรรม์ประกันชีวิตที่สิ้นผลบังคับลงก่อนกำหนด) ซึ่งหลังจากที่ได้มีการปรับเพิ่มตัวแปรอิสระเหล่านี้แล้ว การเปรียบเทียบเวลาการคงอยู่ของกรมธรรม์ที่สิ้นผลบังคับลงก่อนกำหนดระหว่างกลุ่มประชากร ก็จะมีความเอนเอียงที่น้อย และมีความแม่นยำมากขึ้น

ปัญหาที่สำคัญอีกประการหนึ่ง คือ การพยากรณ์การแจกแจงของเวลาจนกระทั่งเกิดเหตุการณ์ที่สนใจจากเซตของตัวแปรอิสระ (independent set) สิ่งที่น่าสนใจในที่นี้ คือ การพยากรณ์ตัวประกอบความเสี่ยง (risk factors) สำหรับเหตุการณ์ที่สนใจ โดยใช้วิธีการทางสถิติสำหรับเทคนิคการวิเคราะห์

ตัวแบบในการศึกษาการคงอยู่มาช่วยในการวิจัย ซึ่งตัวแบบที่ถูกเสนอขึ้นมานี้ จะประกอบไปด้วย Cox Proportional Hazard Model (Cox PH model) และ Accelerated Failure-Time Model (AFT model)

2.5.1 Cox Proportional Hazard Model (Cox PH model)

Cox PH model เป็นตัวแบบกึ่งอิงพารามิเตอร์ (Semi-Parametric Model) ตัวแบบหนึ่งที่นิยมนำไปใช้ เนื่องจากเป็นตัวแบบที่ใช้ได้ทั้งกับข้อมูลที่มีค่าสมบูรณ์และไม่สมบูรณ์ โดย Cox PH model จะมีข้อสมมุติกำหนดไว้สำหรับอัตราส่วนภาวะภัยระหว่างกลุ่มที่ต้องการเปรียบเทียบ 2 กลุ่มว่าจะต้องมีค่าคงที่ทุกค่าสังเกต t หรือกล่าวอีกนัยหนึ่งคือ การสูญเสียสำหรับแต่ละกลุ่มที่สนใจนั้น เป็นสัดส่วนกับการสูญเสียของอีกกลุ่มตัวอย่างที่ต้องการเปรียบเทียบ เมื่อค่าสัดส่วนคงที่นั้นเป็นอิสระกับเวลา การอธิบายในเทอมตัวแบบของฟังก์ชันภาวะภัย ณ เวลา t สำหรับสิ่งที่สนใจ เมื่อสมมติให้แต่ละหน่วยสังเกตมีการพิจารณาที่ค่าของตัวแปรต่างๆ จะแทนด้วย $(Z_1, Z_2,..., Z_p)$

กำหนดให้หน่วยสังเกตที่ j มีค่าของตัวแปร p ตัวเป็น $Z_j(t) = (Z_{j1}(t), Z_{j2}(t),..., Z_{jp}(t))'$ ซึ่งค่าของตัวแปรเหล่านี้อาจเป็น อายุที่เริ่มทำประกันชีวิต เพศผู้เอาประกันภัย แบบกรมธรรม์ประกันชีวิต หรืออาจเป็นฟังก์ชันของเวลา ณ เวลา t ที่มีผลต่อการแจกแจงการคงอยู่ของค่าสังเกต

ให้ $h(t|Z)$ เป็นฟังก์ชันภาวะภัยที่เวลา t สำหรับตัวอย่างที่มีตัวประกอบความเสี่ยงเป็น Z ความสัมพันธ์ระหว่างการแจกแจงของเวลาการคงอยู่ และ Z จะเสนอในรูปของฟังก์ชันภาวะภัยเป็น

$$h(t|Z) = h_0(t)C(\beta'Z)$$

เมื่อ $h_0(t)$ เป็นอัตราภาวะภัยพื้นฐานค่าเลือก (arbitrary baseline hazard rate) $\beta = (\beta_1,..., \beta_p)'$ เป็นเวกเตอร์ของพารามิเตอร์ และ $C(\beta'Z)$ เป็นฟังก์ชันที่ทราบค่า โดยรูปแบบปกติสำหรับ$C(\beta'Z)$ คือ

$$C(\beta'Z) = \exp(\beta'Z) = \exp(\sum_{k=1}^{p} \beta_k Z_k)$$

จึงได้ว่า

$$h(t|Z) = h_0(t)\exp(\beta'Z) = h_0(t)\exp(\sum_{k=1}^{p} \beta_k Z_k)$$

ดังนั้น ลอการิทึมของ $h(t|Z)/h_0(t)$ คือ $\sum_{k=1}^{p} \beta_k Z_k$ จะอยู่ในรูปตัวแบบเชิงเส้นทั่วไป (General Linear Model) และเมื่อมีกลุ่มตัวอย่างที่ต้องการศึกษา 2 กลุ่ม โดยมีตัวแปรอิสระเป็น Z และ Z^* อัตราส่วนฟังก์ชันภาวะภัยจะเป็น

$$\frac{h(t|Z)}{h(t|Z^*)} = \frac{h_0(t)\exp[\sum_{k=1}^{p}\beta_k Z_k]}{h_0(t)\exp[\sum_{k=1}^{p}\beta_k Z^*_k]} = \exp[\sum_{k=1}^{p}\beta_k (Z_k - Z_k^*)]$$

ค่าที่ได้จะเป็นค่าคงที่ ที่เรียกว่าอัตราส่วนภาวะภัย (hazard ratio) ของกลุ่มการศึกษา ที่มีตัวประกอบความเสี่ยง Z เทียบกับอีกกลุ่มหนึ่งที่มีตัวประกอบความเสี่ยงเป็น Z^* (Kleinbaum, 1996; Klien et al., 1997: 230-231)

การประมาณค่าพารามิเตอร์ด้วยวิธีความควรจะเป็นบางส่วน (The Partial Likelihood Estimation)

ฟังก์ชันความควรจะเป็นบางส่วน (The partial likelihood function) สำหรับค่าสังเกตที่มีค่าซ้ำ (tied observation) จะกำหนดให้

$t_1 < t_2 < ... < t_D$ คือ ข้อมูลอันดับของค่าสังเกตสมบูรณ์ D ค่า

d_i คือ จำนวนค่าสังเกตสมบูรณ์ที่มีค่าเท่ากับ t_i

D_i คือ เซตของค่าสังเกตสมบูรณ์ที่มีค่าเท่ากับ t_i

S_i คือ ผลรวมของเวกเตอร์ Z_j ทุกค่าสังเกตสมบูรณ์ที่มีค่าเท่ากับ t_i

นั่นคือ

$$S_i = \sum_{j \in D_i} Z_j$$

และ R_i คือ เซตของจำนวนหน่วยสังเกตที่เข้าร่วมความเสี่ยงก่อนค่าสังเกต t_i

วิธีความควรจะเป็นบางส่วนจะพิจารณาแต่ละเหตุการณ์ของ d_i ณ ค่าสังเกตต่างๆ และสร้างการแจกแจงร่วม (contribution) ของจำนวนค่าสังเกตเหล่านั้นเป็นฟังก์ชันความควรจะเป็นบางส่วน โดยการคูณทุกๆ เหตุการณ์ที่ค่าสังเกต t_i วิธีการประมาณความควรจะเป็นบางส่วนที่นิยมกันโดยส่วนใหญ่จะมีด้วยกันอยู่ 3 วิธี คือ

1) การประมาณความควรจะเป็นบางส่วนด้วยวิธี Breslow

$$L(\beta) = \prod_{i=1}^{D} \frac{\exp(\beta' S_i)}{[\sum_{k \in R_i} \exp(\beta' Z_k)]^{d_i}}$$

2) การประมาณความควรจะเป็นบางส่วนด้วยวิธี Efron

$$L(\beta) = \prod_{i=1}^{D} \frac{\exp(\beta' S_i)}{\prod_{j=1}^{d_i}\left[\sum_{k \in R_i} \exp(\beta' Z_k) - \frac{j-1}{d_i}\sum_{k \in D_i} \exp(\beta' Z_k)\right]}$$

3) การประมาณความควรจะเป็นบางส่วนด้วยวิธี Cox*

$$L(\beta) = \prod_{i=1}^{D} \frac{\exp(\beta' S_i)}{\sum_{q \in Q_i} \exp(\beta' S_q^*)}$$

เมื่อ Q_i แสดงถึงเซตของทุกๆ เซตย่อยของแต่ละหน่วยสังเกต d_i ที่สามารถเลือกได้จากเซตของจำนวนหน่วยสังเกตที่เข้าร่วมความเสี่ยง R_i ณ ค่าสังเกต t_i ผลรวมในตัวหาร คือ การรวมทุกเซตของแต่ละหน่วยสังเกต d_i จากเซตความเสี่ยงที่ไม่มีการแทนที่ โดย $q = (q_1,\ldots, q_d)$ เป็นหนึ่งในเหตุการณ์ของ Q_i และกำหนดให้ $S_q^* = \sum_{j=1}^{d} Z_{qj}$ (Collet, 1994: 66)

ในที่นี้จะขอพิจารณาตัวประมาณความควรจะเป็นบางส่วนสูงสุดโดยวิธี Efron ซึ่งพิจารณาได้จากการหาอนุพันธ์บางส่วนของสมการ (2.5.1) กับค่าของ β ต่างๆ

ให้ $LL(\beta) = \ln[L(\beta)]$ หลังจากจัดรูปเทอมต่างๆ แล้ว สามารถเขียน $LL(\beta)$ ได้เป็น

$$LL(\beta) = \sum_{i=1}^{D} \beta' S_i - \sum_{i=1}^{D} \sum_{j=1}^{d} \ln\Big[\sum_{k \in R} \exp(\beta' Z_k) - \frac{j-1}{d_i} \sum_{k \in D} \exp(\beta' Z_k)\Big] \qquad (2.5.1)$$

$$\text{ให้ } U_h(\beta) = \frac{\partial LL(\beta)}{\partial \beta_h} = 0 \qquad , h = 1,2,\ldots p$$

ดังนั้น

$$U_h(\beta) = \sum_{i=1}^{D} S_{hi} - \sum_{i=1}^{D} \sum_{j=1}^{d} \frac{\sum_{k \in R} Z_{hk} \exp(\beta' Z_k) - \frac{j-1}{d_i} \sum_{k \in D} Z_{hk} \exp(\beta' Z_k)}{\sum_{k \in R} \exp(\beta' Z_k) - \frac{j-1}{d_i} \sum_{k \in D} \exp(\beta' Z_k)}$$

เมทริกซ์สารสนเทศ (Information matrix) คือลบของเมทริกซ์อนุพันธ์อันดับที่ 2 ของลอการิทึมความควรจะเป็น ที่กำหนดโดย $I(\beta) = [I_{gh}(\beta)]_{pxp}$ กับ $(g,h)^{th}$

$$I_{gh}(\beta) = \sum_{i=1}^{D} \sum_{j=1}^{d} \frac{\sum_{k \in R} Z_{kg} Z_{kh} \exp(\beta' Z_k) - \frac{j-1}{d_i} \sum_{k \in D} Z_{kg} Z_{kh} \exp(\beta' Z_k)}{\sum_{k \in R} \exp(\beta' Z_k) - \frac{j-1}{d_i} \sum_{k \in D} \exp(\beta' Z_k)}$$

$$- \sum_{i=1}^{D} \sum_{j=1}^{d} \frac{\sum_{k \in R} Z_{kg} \exp(\beta' Z_k) - \frac{j-1}{d_i} Z_{kg} \sum_{k \in D} \exp(\beta' Z_k)}{\sum_{k \in R} \exp(\beta' Z_k) - \frac{j-1}{d_i} \sum_{k \in D} \exp(\beta' Z_k)} \times \frac{\sum_{k \in R} Z_{hk} \exp(\beta' Z_k) - \frac{j-1}{d_i} \sum_{k \in D} Z_{hk} \exp(\beta' Z_k)}{\sum_{k \in R} \exp(\beta' Z_k) - \frac{j-1}{d_i} \sum_{k \in D} \exp(\beta' Z_k)}$$

---

* ฟังก์ชันความควรจะเป็นมีการสมมติว่า ฟังก์ชันภาวะภัยมีการแจกแจงเป็นแบบลอจิสติค (The logistic distribution)

การประมาณค่าพารามิเตอร์ในสมการ จะใช้การวิเคราะห์เชิงตัวเลข (Numerical analysis) โดยวิธีนิวตัน-ราฟสัน (Newton-Raphson Method) และเนื่องจากในสมการไม่ขึ้นอยู่กับ $h_0(t)$ ดังนั้น การสรุปอาจพิจารณาบนฐานของตัวแปรอิสระโดยที่ไม่ต้องทราบค่า $h_0(t)$

สำหรับการทดสอบสมมติฐานเกี่ยวกับค่าพารามิเตอร์การถดถอย β ($H_0 : \beta = \beta_0$) ที่สำคัญในที่นี้มีอยู่ 3 วิธี โดยให้ $I(\beta)$ เป็นเมทริกซ์สารสนเทศจากการคำนวณที่ β (ขนาด pxp)

1) Wald's Test

$$\chi^2_W = (b - \beta_0)' I(b)(b - \beta_0)$$

ถ้าสมมติฐานว่างเป็นจริง และตัวอย่างมีขนาดใหญ่ จะมีการแจกแจงไคสแควร์ที่มีองศาความอิสระ p

2) Likelihood Ratio Test

$$\chi^2_{LR} = 2[LL(b) - LL(\beta_0)]$$

ถ้าสมมติฐานว่างเป็นจริง และตัวอย่างมีขนาดใหญ่ จะมีการแจกแจงไคสแควร์ที่มีองศาความเป็นอิสระ p

3) Score Test

$$\chi^2_{SC} = U(\beta_0)' I^{-1}(\beta_0) U(\beta_0)$$

ถ้าสมมติฐานว่างเป็นจริง และตัวอย่างมีขนาดใหญ่ จะมีการแจกแจงไคสแควร์ที่มีองศาความเป็นอิสระ p

2.5.2 Accelerated Failure-Time Model (AFT model)

นอกจาก Cox PH Model ที่เป็นตัวแบบการถดถอยสำหรับข้อมูลการคงอยู่ที่ไม่ต้องกำหนดสมมติฐานรูปแบบการแจกแจงเกี่ยวกับรูปร่างของฟังก์ชันการคงอยู่ ซึ่งเป็นตัวแบบกึ่งอิงพารามิเตอร์แล้ว ยังมีอีกตัวแบบหนึ่งที่สามารถแสดงความสัมพันธ์ระหว่างตัวแปรอิสระ กับเวลาการคงอยู่ คือ ตัวแบบอิงพารามิเตอร์ (Parametric Model) ในส่วนนี้จะอธิบายเฉพาะตัวแบบอิงพารามิเตอร์แบบต่างๆ ที่พบบ่อยในการวิเคราะห์ข้อมูลการคงอยู่ ซึ่งถ้าสามารถเลือกตัวแบบได้เหมาะสมกับข้อมูลการคงอยู่แล้ว ค่าประมาณพารามิเตอร์ที่ได้ก็จะมีความแม่นยำมากขึ้น ดังนั้น การประมาณค่าพารามิเตอร์ของตัวสถิติอิงพารามิเตอร์ ขั้นแรก จึงควรทราบตัวแบบการแจกแจงของตัวอย่างที่สุ่มขึ้นมาว่ามีการแจกแจงแบบใด โดยทั่วไปตัวแบบการแจกแจงการคงอยู่ที่มักพบอยู่เสมอ ได้แก่ (Klien et al., 1997)

ก) การแจกแจงแบบเลขชี้กำลัง (The Exponential Distribution)

การแจกแจงแบบเลขชี้กำลังจะมีลักษณะฟังก์ชันภาวะภัย (hazard function) เป็นค่าคงที่เท่ากับ λ เมื่อ λ คือ ค่าพารามิเตอร์ของการแจกแจงที่แสดงถึงความเสี่ยงในการสูญเสีย โดย λ ที่มีค่ามาก จะแสดงถึงความเสี่ยงในการสูญเสียสูง นั่นคือ มีเวลาการคงอยู่สั้น (short survival) และเมื่อ λ มีค่าน้อย นั่นหมายถึง ความเสี่ยงในการสูญเสียต่ำ หรือมีเวลาการคงอยู่ยาวนาน

ฟังก์ชันความน่าจะเป็นหนาแน่น

$$f(t) = \lambda \exp(-\lambda t) \qquad t \geq 0, \lambda > 0$$
$$= 0 \qquad t < 0$$

ฟังก์ชันการแจกแจงสะสม

$$F(t) = 1 - \exp(-\lambda t) \qquad t \geq 0$$

ฟังก์ชันการคงอยู่

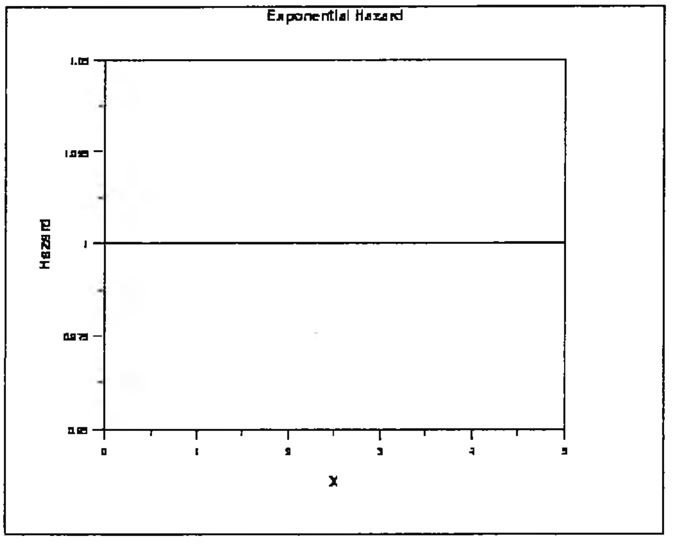
$$S(t) = \exp(-\lambda t) \qquad t \geq 0$$

ฟังก์ชันภาวะภัย

$$h(t) = \lambda \qquad t \geq 0$$

โดยที่ค่าเฉลี่ย $= 1/\lambda$

ค่าความแปรปรวน $= 1/\lambda^2$



รูปที่ 2 ฟังก์ชันภาวะภัยสำหรับการแจกแจงแบบเลขชี้กำลัง

อนึ่ง แม้ว่าการแจกแจงแบบเลขชี้กำลังจะเป็นที่นิยมจากการมีตัวแบบง่ายในการประมาณ แต่เนื่องจากฟังก์ชันภาวะภัยมีค่าคงที่ (λ) จึงมีข้อจำกัดค่อนข้างมากในการประยุกต์ใช้

ข) การแจกแจงแบบไวบูลล์ (The Weibull Distribution)

การแจกแจงแบบไวบูลล์มีการปรับเปลี่ยนมาจากการแจกแจงแบบเลขชี้กำลัง ต่างกันแต่ว่า การแจกแจงแบบไวบูลล์ไม่ได้มีฟังก์ชันภาวะภัยเป็นค่าคงที่เหมือนกับการแจกแจงแบบเลขชี้กำลัง และสามารถนำไปประยุกต์ใช้ได้กว้างขวางมากกว่า นอกจากนี้ การแจกแจงแบบไวบูลล์ยังสามารถแสดงฟังก์ชันภาวะภัยได้หลายค่า ขึ้นอยู่กับค่าพารามิเตอร์ต่างๆ คือ พารามิเตอร์เชิงมาตราส่วน (scale parameter : λ) และพารามิเตอร์เชิงรูปร่าง (shape parameter : α)

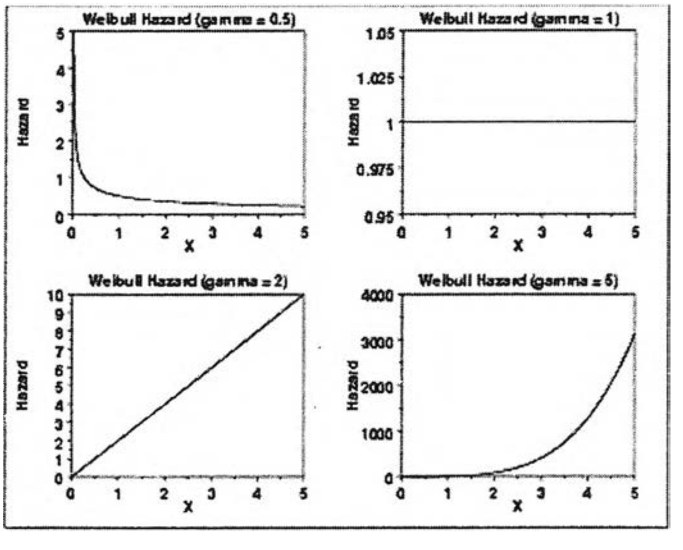
ฟังก์ชันภาวะภัยจะขึ้นอยู่กับค่า α โดยเมื่อ $\alpha > 1$ จะมีความเสี่ยงในการสูญเสียสูงหรือมีเวลาการคงอยู่สั้น และฟังก์ชันภาวะภัยมีค่าลดลงเมื่อ $\alpha < 1$ ดังนั้น การแจกแจงแบบไวบูลล์จึงเหมาะสมกับการนำมาใช้กับตัวแบบการแจกแจงการคงอยู่ของประชากร ที่มีความเสี่ยงระดับคงที่ เพิ่มขึ้น หรือลดลง

ฟังก์ชันความน่าจะเป็นหนาแน่น

$$f(t) = \alpha\lambda t^{\alpha-1} \exp[-\lambda t^{\alpha}] \qquad t \geq 0; \alpha, \lambda > 0$$

ฟังก์ชันการแจกแจงสะสม

$$F(t) = 1 - [\exp - \lambda t^{\alpha}]$$



รูปที่ 3 ฟังก์ชันภาวะภัยสำหรับการแจกแจงแบบไวบูลล์

ฟังก์ชันการคงอยู่

$$S(t) = \exp[-\lambda t^{\alpha}]$$

ฟังก์ชันภาวะภัย

$$h(t) = \alpha\lambda t^{\alpha-1}$$

โดยที่ค่าเฉลี่ย $= \alpha / \lambda^2$

ค่าความแปรปรวน $= \alpha^2 / \lambda$

ค) การแจกแจงแบบลอก-นอร์มอล (The Log-Normal Distribution)

การแจกแจงแบบลอก-นอร์มอลจะมีลักษณะฟังก์ชันภาวะภัยเป็นรูปตัวยูคว่ำ (hump-shaped) โดยพบว่า เมื่อ t เพิ่มมากขึ้น ฟังก์ชันภาวะภัยจะเพิ่มขึ้นเรื่อยๆ จนถึงค่า t ค่าหนึ่ง ที่ฟังก์ชันภาวะภัยมีค่าสูงสุดแล้ว จึงเริ่มลดลงเข้าสู่ศูนย์เมื่อเวลาเข้าใกล้อนันต์ การแจกแจงแบบนี้ จึงเหมาะสมสำหรับตัวแบบของเวลาการคงอยู่ ที่มีฟังก์ชันภาวะภัยเพิ่มขึ้นในช่วงระยะแรก และลดลงในเวลาต่อมา

เมื่อ T มีการแจกแจงเป็นแบบลอก-นอร์มอล ln (T) จะมีการแจกแจงแบบปกติ ที่มีค่าเฉลี่ย μ และความแปรปรวน σ^2 เขียนได้เป็น

$$T \sim \Lambda(\mu, \sigma^2)$$

ฟังก์ชันความน่าจะเป็นหนาแน่น

$$f(t) = \frac{\exp[-\frac{1}{2}(\frac{\ln t - \mu}{\sigma})^2]}{t\sigma\sqrt{2\pi}} \qquad t > 0, \sigma > 0$$

$$= \phi\left(\frac{\ln t - \mu}{\sigma}\right) / t$$

ฟังก์ชันการคงอยู่

$$S(t) = 1 - \Phi[\frac{\ln t - \mu}{\sigma}]$$

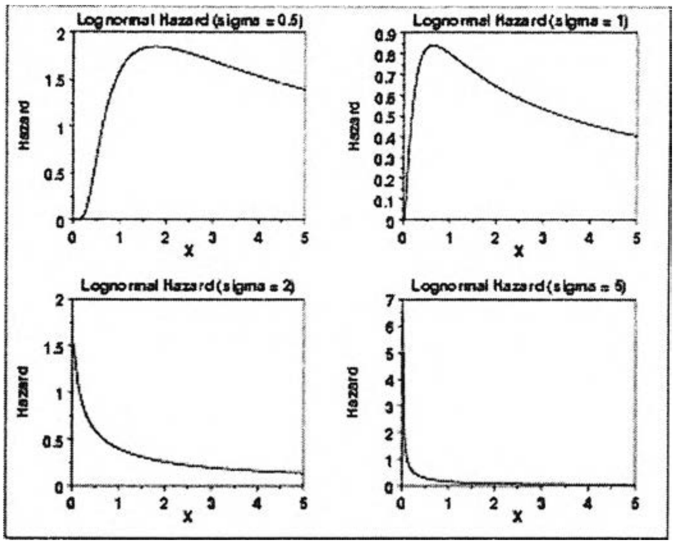
เมื่อ $\Phi(\phi)$ เป็นฟังก์ชันการแจกแจงสะสม (Cumulative distribution function) ของตัวแปรปกติมาตรฐาน

ฟังก์ชันภาวะภัย

$$h(t) \quad = f(t) / S(t)$$

โดยที่ค่าเฉลี่ย $= \exp[\mu + (\sigma^2 / 2)]$

ค่าความแปรปรวน $= [\exp(\sigma^2) - 1][\exp(2\mu + \sigma^2)]$



รูปที่ 4 ฟังก์ชันภาวะภัยสำหรับการแจกแจงแบบลอก-นอร์มอล

โดยส่วนใหญ่ การแจกแจงแบบลอก-นอร์มอลมักเป็นที่ยอมรับและใช้กันทั่วไป เพราะมีความสัมพันธ์กับการแจกแจงแบบปกติ (The normal distribution)

ง) การแจกแจงแบบลอกลอจิสติค (The Log-Logistic Distribution)

การแจกแจงแบบลอกลอจิสติคเป็นอีกหนึ่งการแจกแจง ที่มีรูปร่างของฟังก์ชันภาวะภัยเป็นรูปตัวยูคว่ำเหมือนการแจกแจงแบบลอก-นอร์มอล และมีลักษณะคล้ายกับการแจกแจงแบบปกติ (Cox and Oakes, 1990: 21) โดยมี

ฟังก์ชันความน่าจะเป็นหนาแน่น

$$f(t) = \frac{\alpha\lambda t^{\alpha-1}}{(1+\lambda t^{\alpha})^2}$$

ฟังก์ชันการคงอยู่

$$S(t) = \frac{1}{1+\lambda t^{\alpha}}$$

ฟังก์ชันภาวะภัย

$$h(t) = \frac{\alpha\lambda t^{\alpha-1}}{1+\lambda t^{\alpha}}$$

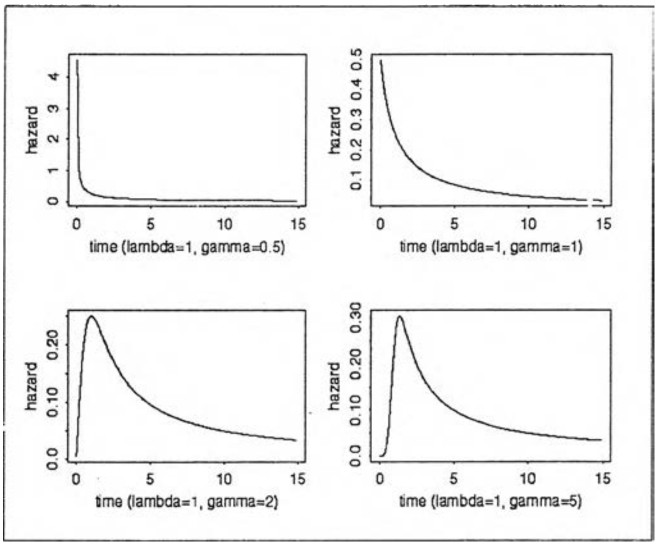
เมื่อ $\alpha = (1/\sigma) > 0$ และ $\lambda = \exp(-\mu/\sigma)$

ตัวเศษของฟังก์ชันภาวะภัยจะมีลักษณะคล้ายกับการแจกแจงแบบไวบูลล์ ส่วนตัวหาร แสดงให้เห็นถึงฟังก์ชันภาวะภัยว่ามีลักษณะลดลงทางเดียว (monotone decreasing) สำหรับ $\alpha = \dfrac{1}{\sigma} \leq 1$ และเป็นรูปตัวยูคว่ำมีจุดสูงสุดเพียงจุดเดียวถ้า $\alpha = \dfrac{1}{\sigma} > 1$

โดยที่ค่าเฉลี่ย $= \pi Csc(\pi/\alpha)/(\alpha\lambda^{1/\alpha})$ ถ้า $\alpha > 1$

ค่าความแปรปรวน $= 2\pi Csc(2\pi/\alpha)/(\alpha\lambda^{2/\alpha}) - E[X]^2$ ถ้า $\alpha > 1$

การแจกแจงแบบลอกลอจิสติคมีลักษณะคล้ายกับการแจกแจงแบบไวบูลล์ และแบบเลขชี้กำลัง



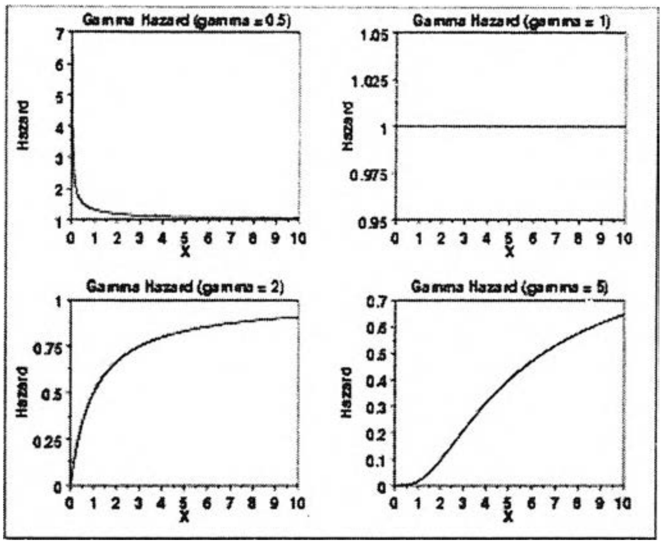
รูปที่ 5 ฟังก์ชันภาวะภัยสำหรับการแจกแจงแบบลอกลอจิสติค

จ) การแจกแจงแบบแกมมา (The Gamma Distribution)

การแจกแจงแบบแกมมา จะมีคุณสมบัติคล้ายกับการแจกแจงแบบไวบูลล์ แต่สูตรโดยทั่วไปจะมีลักษณะค่อนข้างซับซ้อนมากกว่า โดยมีฟังก์ชันความน่าจะเป็นหนาแน่นคือ

$$f(t) = \frac{\lambda^{\beta} t^{\beta-1}\exp(-\lambda t)}{\Gamma(\beta)}$$

เมื่อ λ (scale parameter) > 0, β (shape parameter) > 0, t > 0 และ $\Gamma(\beta)$ คือฟังก์ชันแกมมา ฟังก์ชันภาวะภัยสำหรับการแจกแจงแบบแกมมาจะมีลักษณะเพิ่มขึ้นทางเดียว (monotone increasing) สำหรับค่า $\beta > 1$, h(0) = 0 และ h(t) ลู่เข้าสู่ λ เมื่อ t ลู่เข้าสู่อนันต์ และลดลงทางเดียว (monotone decreasing) สำหรับ $\beta < 1$, h(0) = ∞ และ h(t) ลู่เข้าสู่ λ เมื่อ t ลู่เข้าสู่อนันต์



รูปที่ 6 ฟังก์ชันภาวะภัยสำหรับการแจกแจงแบบแกมมา

ฟังก์ชันการคงอยู่

$$S(t) = 1 - \mathrm{I}(\lambda t, \beta)$$

เมื่อ I คือ ฟังก์ชันแกมมาไม่สมบูรณ์

ฟังก์ชันภาวะภัย

$$h(t) = f(t) / S(t)$$

โดยที่ค่าเฉลี่ย $= \beta / \lambda$

ค่าความแปรปรวน $= \beta / \lambda^2$

การแจกแจงแบบแกมมามีอยู่ 2 ตัวแบบ คือ ตัวแบบมาตรฐาน (The standard model) ที่มี 2 พารามิเตอร์ และตัวแบบทั่วไป (The generalized model) ที่มี 3 พารามิเตอร์ การวิเคราะห์ส่วนใหญ่จะนิยมใช้ตัวแบบแกมมาทั่วไป เพราะตัวแบบแกมมาทั่วไปจะมีพารามิเตอร์มากกว่าตัวแบบอื่นๆ ที่ทำการพิจารณา ส่งผลให้ฟังก์ชันภาวะภัยมีได้หลายลักษณะ (Allison, 1995: 74) และเป็นประโยชน์เมื่อต้องการทดสอบอัตราส่วนความควรจะเป็นสำหรับการคัดเลือกตัวแบบที่ดีที่สุด

อย่างไรก็ตาม แม้ว่าตัวแบบแกมมาทั่วไปจะมีคุณสมบัติพิเศษต่างๆ เมื่อเทียบกับตัวแบบอื่นๆ แต่ก็ไม่สามารถที่จะใช้ตัวแบบแกมมาทั่วไปในการแทนตัวแบบอื่นๆ ได้ เนื่องมาจากเหตุผลสองประการ คือ สูตรฟังก์ชันภาวะภัยของตัวแบบแกมมาทั่วไปค่อนข้างจะซับซ้อน มีความสัมพันธ์ทั้งกับฟังก์ชันแกมมาและฟังก์ชันแกมมาแบบไม่สมบูรณ์ จึงเป็นการยากที่จะพิจารณา

ลักษณะฟังก์ชันภาวะภัยจากค่าประมาณพารามิเตอร์ รวมทั้งฟังก์ชันภาวะภัยสำหรับตัวแบบเฉพาะก็ค่อนข้างที่จะอธิบายได้ง่ายกว่า (Allison, 1995: 74)

ฟังก์ชันความน่าจะเป็นหนาแน่นของตัวแบบแกมมาทั่วไป คือ

$$f(t) = \frac{\alpha\lambda^{\beta} t^{\alpha\beta-1} \exp(-\lambda t^{\alpha})}{\Gamma(\beta)}$$

โดยจะลดลงเป็นตัวแบบเลขชี้กำลังเมื่อ $\alpha, \beta = 1$ ตัวแบบไวบูลล์เมื่อ $\beta = 1$ ตัวแบบแกมมามาตรฐานเมื่อ $\alpha = 1$ และเข้าใกล้ตัวแบบลอก-นอร์มอลเมื่อ β ลู่เข้าสู่อนันต์

จากรูปแบบการแจกแจงที่มักพบอยู่เสมอในการวิเคราะห์ข้อมูลการคงอยู่ ตัวแบบที่จะทำการพิจารณาในส่วนนี้จะประกอบไปด้วย AFT model และตัวแบบลอการิทึมเวลาเชิงเส้น (linear log-time model) โดยสัญลักษณ์ T หมายถึง ค่าสังเกตของเหตุการณ์ที่สนใจ และ Z เป็นเวกเตอร์ของตัวแปรอิสระที่มีค่าคงที่ กล่าวคือ ฟังก์ชันการคงอยู่ของหน่วยสังเกตที่มีตัวแปรอิสระ Z ณ เวลา t จะมีค่าเช่นเดียวกับฟังก์ชันการคงอยู่ของหน่วยสังเกตกับฟังก์ชันการคงอยู่พื้นฐาน (baseline survival function) ที่เวลา $t \exp(\theta' Z)$ เมื่อ $\theta' = (\theta_1, \ldots, \theta_p)$ เป็นเวกเตอร์ของพารามิเตอร์การถดถอย กล่าวคือ AFT model ถูกกำหนดโดยความสัมพันธ์

$$S(t|Z) = S_0[\exp(\theta' Z)t] \qquad \text{ทุกค่า t}$$

$\exp(\theta' Z)$ ถูกเรียกว่าตัวประกอบเร่ง (acceleration factor) ซึ่งเป็นค่าที่บอกให้ทราบถึงการเปลี่ยนแปลงในค่าของตัวแปรอิสระจากฐานมาตราเวลา

สำหรับหน่วยสังเกตที่มีตัวแปรอิสระ Z มีความเกี่ยวข้องกับฟังก์ชันภาวะภัยพื้นฐานโดย

$$h(t|Z) = \exp(\theta' Z) h_0[\exp(\theta' Z)t] \qquad \text{ทุกค่า t}$$

แบบที่สองของการแสดงความสัมพันธ์ระหว่างค่าของตัวแปรอิสระและเวลาการคงอยู่ คือ ความสัมพันธ์เชิงเส้นระหว่างลอการิทึมของเวลาและค่าของตัวแปรอิสระ ในที่นี้จะสมมติตัวแบบเชิงเส้นทั่วไปสำหรับลอการิทึมของเวลาคือ

$$Y = \ln(T) = \mu + \gamma' Z + \sigma W$$

เมื่อ $\gamma' = (\gamma_1, \ldots, \gamma_p)$ เป็นเวกเตอร์ของพารามิเตอร์การถดถอย และ W เป็นการแจกแจงของความคลาดเคลื่อน (The error distribution)

ทั้งสองแบบที่แสดงมานี้มีความสัมพันธ์กันอย่างใกล้ชิด โดยเมื่อกำหนดให้ $S_0(t)$ เป็นฟังก์ชันการคงอยู่ของตัวแปรสุ่ม $\exp(\mu + \sigma w)$ แล้ว ดังนั้น ตัวแบบลอการิทึมเวลาเชิงเส้นจะเท่ากับ AFT model ที่ $\theta = -\gamma$

ยกตัวอย่าง กรณี AFT model และตัวแบบลอการิทึมเวลาเชิงเส้น เมื่อมีการแจกแจงเป็นแบบต่างๆ คือ

1) การแจกแจงแบบไวบูลล์ (The Weibull Distribution)

การแจกแจงแบบไวบูลล์เป็นการแจกแจงที่มีฟังก์ชันภาวะภัยเป็นได้ทั้งค่าคงที่ เพิ่มขึ้น หรือลดลง และเป็นตัวแบบการถดถอยอิงพารามิเตอร์ (Parametric regression model) เพียงตัวแบบเดียว ที่แสดงได้ทั้ง Cox PH model และ AFT model โดยฟังก์ชันต่างๆ สำหรับการแจกแจงไวบูลล์จะมีลักษณะเป็น

ฟังก์ชันการคงอยู่

$$S_T(t) = \exp(-\lambda t^{\alpha})$$

ฟังก์ชันภาวะภัย

$$h_T(t) = \alpha\lambda t^{\alpha-1}$$

เมื่อเปลี่ยนเวลาให้อยู่ในรูปลอการิทึม จะได้ฟังก์ชันการคงอยู่ของหนึ่งตัวแปรสำหรับ Y = ln (T) คือ

$$S_Y(y) = \exp(-\lambda e^{\alpha y})$$

กำหนดพารามิเตอร์ใหม่โดยให้ $\lambda = \exp(-\mu / \sigma)$ และ $\sigma = 1/\alpha$ ดังนั้น Y จะเป็นไปตามตัวแบบลอการิทึมเชิงเส้นที่แสดงได้ว่า

$$Y = \ln(T) = \mu + \sigma W$$

เมื่อ W มีการแจงแจงเป็นแบบค่าสุดขีด (The extreme value distribution) ที่มีฟังก์ชันความน่าจะเป็นหนาแน่น

$$f_w(w) = \exp(w - e^{w})$$

ฟังก์ชันการคงอยู่

$$S_W(w) = \exp(-e^{w})$$

ดังนั้น ฟังก์ชันความน่าจะเป็นหนาแน่น และฟังก์ชันการคงอยู่สำหรับ Y จะเป็น

$$f_y(y) = (1/\sigma)\exp\{[(y-\mu)/\sigma] - e^{[(y-\mu)/\sigma]}\}$$

$$S_Y(y) = \exp(-e^{[(y-\mu)/\sigma]})$$

เมื่อ $\alpha = 1$ หรือ $\sigma = 1$ การแจกแจงแบบไวบูลล์จะลดรูปลงเป็นการแจกแจงแบบเลขชี้กำลัง

เมื่อทราบฟังก์ชันต่างๆ แล้ว จึงนำมาสร้างฟังก์ชันความควรจะเป็นเพื่อประมาณค่าพารามิเตอร์ μ และ σ จากค่าสังเกตของเมทริกซ์สารสนเทศ (Observed Information Matrix) ตามคุณสมบัติคงที่ของตัวประมาณความควรจะเป็นสูงสุด (Maximum Likelihood Estimator : MLE) จึงสามารถหาตัวประมาณ MLE ของ λ และ α โดย

$$\hat{\lambda} = \exp(-\hat{\mu}/\hat{\sigma})$$

และ $$\hat{\alpha} = 1/\hat{\sigma}$$

ค่าความแปรปรวนและความแปรปรวนร่วม

$$\hat{Var}(\hat{\lambda}) = \exp(-2\hat{\mu}/\hat{\sigma})[\hat{Var}(\hat{\mu})/\hat{\sigma}^2 + \hat{\mu}^2\, Var(\hat{\sigma})/\hat{\sigma}^4 - 2\hat{\mu}Cov(\hat{\mu},\hat{\sigma})/\hat{\sigma}^3]$$

$$\hat{Var}(\hat{\alpha}) = \hat{Var}(\hat{\sigma})/\hat{\sigma}^4$$

$$\hat{Cov}(\hat{\lambda},\hat{\alpha}) = \exp(-\hat{\mu}/\hat{\sigma})[Cov(\hat{\mu},\hat{\sigma})/\hat{\sigma}^3 - \hat{\mu}Var(\hat{\sigma})/\hat{\sigma}^4]$$

ตัวแปรอิสระที่เข้าไปในการแจกแจงแบบไวบูลล์ จะได้ตัวแบบลอการิทึมเวลาเชิงเส้น คือ

$$Y = \mu + \gamma Z + \sigma W$$

เมื่อ W มีการแจกแจงเป็นแบบค่าสุดขีด ตัวแบบนี้จะนำไปสู่ Cox PH model สำหรับ T ที่มีภาวะภัยพื้นฐานเป็นการแจกแจงแบบไวบูลล์ นั่นคือ

$$h(t|Z) = (\alpha\lambda t^{\alpha-1})\exp(\beta' Z)$$

โดย $\alpha = 1/\sigma$, $\lambda = \exp(-\mu/\sigma)$ และ $\beta_j = -\gamma_j / \sigma$ $\quad j = 1,2,\ldots,p$

AFT model ของตัวแบบการถดถอยไวบูลล์ จะได้ฟังก์ชันภาวะภัยสำหรับหน่วยสังเกตที่มีเวกเตอร์ตัวแปรอิสระ Z คือ

$$h(t|Z) = \exp(\theta' Z)h_0[t\,\exp(\theta' Z)]$$

เมื่อ $h_0(t)$ คือ $\alpha\lambda t^{\alpha-1}$ และ $\exp(\theta' Z)$ คือ ตัวประกอบเร่ง

สำหรับตัวแบบการถดถอยไวบูลล์ จะประมาณโดยมีพื้นฐานจากตัวแบบลอการิทึมเวลาเชิงเส้นตามคุณสมบัติไม่แปรเปลี่ยน (invariance property) ของตัวประมาณ MLE

2) การแจกแจงแบบลอกลอจิสติค (The Log-Logistic Distribution)

ตามลักษณะการแจกแจงแบบลอกลอจิสติคที่มีฟังก์ชันภาวะภัยเป็นรูปตัวยูคว่ำ ฟังก์ชันการคงอยู่และฟังก์ชันภาวะภัยสะสมสำหรับ T เมื่อ T เป็นไปตามการแจกแจงแบบลอกลอจิสติค คือ

$$S_T(t) = \frac{1}{1+\lambda t^{\alpha}} \tag{2.5.2}$$

และ

$$H_T(t) = \ln(1+\lambda t^{\alpha})$$

แปลงรูปแบบเวลาโดยการใส่ลอการิทึม ฟังก์ชันการคงอยู่สำหรับ Y = ln (T) จะเท่ากับ

$$S_Y(y) = \frac{1}{1+\lambda e^{\alpha y}} \tag{2.5.3}$$

ตัวแบบลอการิทึมเชิงเส้นที่ไม่มีตัวแปรอิสระจะเท่ากับ

$$Y = \ln(T) = \mu + \sigma W$$

เมื่อ W มีการแจกแจงแบบลอจิสติคมาตรฐาน (The standard logistic distribution) ที่มีฟังก์ชันความน่าจะเป็นหนาแน่นคือ

$$f_w(w) = e^{w} /(1+e^{w})^2$$

ฟังก์ชันการคงอยู่

$$S_w(w) = 1/(1+e^{w})$$

ดังนั้น ฟังก์ชันความน่าจะเป็นหนาแน่น และฟังก์ชันการคงอยู่สำหรับ Y จะเป็นไปตาม

$$f_y(y) = (1/\sigma)\exp[(y-\mu)/\sigma]/[1+\exp[(y-\mu)/\sigma]^2$$

และ

$$S_Y(y) = 1/[1+e^{[(y-u)/\sigma]}] \tag{2.5.4}$$

ค่าพารามิเตอร์ของการแจกแจงลอกลอจิสติคสำหรับตัวแปรสุ่ม T ในสมการ (2.5.2) และสำหรับการแจกแจงของตัวแปร Y ที่เปลี่ยนอยู่ในรูปลอการิทึม ฟังก์ชันการคงอยู่สำหรับ Y = ln(T) จะเป็นไปตามฟังก์ชันของพารามิเตอร์ลอการิทึมเชิงเส้นใน (2.5.4)

นอกจากนี้ค่า α และ λ จะคำนวณได้จาก $\alpha = 1/\sigma$ และ $\lambda = \exp(-\mu/\sigma)$ เช่นเดียวกับตัวแบบการถดถอยไวบูลล์ และสามารถประมาณ μ , σ, λ , α กับเมทริกซ์ความแปรปรวนร่วมได้เช่นเดียวกัน

รูปแบบที่ใช้ในการสร้างตัวแบบอันเกิดจากผลของตัวแปรอิสระตามเวลาการคงอยู่กับการแจกแจงลอกลอจิสติคมีอยู่ด้วยกัน 3 ตัวแบบ ตัวแบบแรก จะเป็นตัวแบบลอการิทึมเวลาเชิงเส้นกับ $Y = \ln(T) = \mu + \gamma Z + \sigma W$ เมื่อ W มีการแจกแจงเป็นลอจิสติคพื้นฐาน ตัวแบบที่สองได้มาจากการแทนที่ λ ใน (2.5.3) โดย $\exp(\beta' Z)$ ฟังก์ชันการคงอยู่แบบมีเงื่อนไขสำหรับค่าสังเกตจะให้โดย

$$S_T(t|Z) = \frac{1}{1 + \lambda \exp(\beta' Z) t^{\alpha}}$$

ค่าพารามิเตอร์จะมีความสัมพันธ์ คือ

$$\lambda = \exp(-\mu / \sigma)$$

$$\alpha = 1 / \sigma$$

และ

$$\beta = -\gamma / \sigma$$

บนพื้นฐานตัวประมาณ MLE สำหรับ μ ,γ , σ และเมทริกซ์ความแปรปรวนร่วม ซึ่งค่าประมาณสำหรับ λ , β , α และเมทริกซ์ความแปรปรวนร่วมจะได้มาจากวิธีเดลต้า (Delta Method)

ตัวแบบที่สาม ตัวแบบการถดถอยลอกลอจิสติคจะเป็นเหมือนกับ AFT model ที่เป็นไปตามฟังก์ชันการคงอยู่แบบลอกลอจิสติค (log-logistic survival function) ตัวแบบลอกลอจิสติคนี้เป็นตัวแบบอิงพารามิเตอร์เพียงตัวแบบเดียวเท่านั้น ที่เป็นทั้ง Proportional Odds Model และ AFT model (Klien et al., 1997)

3) การแจกแจงพารามิเตอร์แบบอื่นๆ (Other Parametric Distribution)

ตัวแบบแรกที่จะพิจารณา คือ ตัวแบบลอก-นอร์มอล ในที่นี้จะกำหนดให้ตัวแปรอิสระคือ $Z = (Z_1, ..., Z_p)'$ ซึ่งจะได้ลอการิทึมของค่าสังเกตเป็นไปตามตัวแบบการถดถอยปกติทั่วไป นั่นคือ

$$Y = \ln(T) = \mu + \gamma Z + \sigma W$$

เมื่อ W มีการแจกแจงเป็นแบบปกติมาตรฐาน (The standard normal distribution) และมีรูปแบบทั่วไปของฟังก์ชันภาวะภัยเหมือนกับการแจกแจงแบบลอกลอจิสติค

สำหรับการแจกแจงแบบลอก-นอร์มอล ฟังก์ชันการคงอยู่ของค่าสังเกต T จะให้โดย

$$S(t) = 1 - \Phi\{[\ln(t) - (\mu + \gamma Z)] / \sigma\}$$

เมื่อ $\Phi\{\ \}$ เป็นฟังก์ชันการแจกแจงสะสมปกติมาตรฐาน (The standard normal cumulative distribution function)

ตัวแบบต่อมา คือ ตัวแบบแกมมาทั่วไป (The generalized gamma model) ตัวแบบนี้มีประโยชน์มากในการเลือกตัวแบบที่ใช้พารามิเตอร์ต่างๆ เพราะตัวแบบนี้จะรวมตัวแบบไวบูลล์ ตัวแบบเลขชี้กำลัง และตัวแบบลอก-นอร์มอลในตัวเอง สำหรับตัวแบบนี้ Y = ln (T) จะเป็นไปตามตัวแบบการถดถอยปกติทั่วไปกับ W ที่มีฟังก์ชันความน่าจะเป็นหนาแน่นเป็น

$$f(w) = \frac{|\theta|[\exp(\theta w)/\theta^2]^{(1/\theta^2)}\exp[-\exp(\theta w)/\theta^2]}{\Gamma(1/\theta^2)} \quad ,-\infty < w < \infty$$

กรณี $\theta = 1$ ตัวแบบนี้จะลดรูปลงเป็นตัวแบบการถดถอยไวบูลล์ สำหรับ $\theta = 0$ ตัวแบบนี้จะลดรูปลงเป็นตัวแบบการถดถอยลอก-นอร์มอล และเมื่อ $\theta = 1$ กับ $\sigma = 1$ จะลดรูปลงเป็นตัวแบบการถดถอยเลขชี้กำลัง (Klien et al., 1997)

การประมาณค่าพารามิเตอร์ด้วยวิธีความควรจะเป็นสูงสุด

วิธีความควรจะเป็นสูงสุดเป็นวิธีที่นิยมกันอย่างกว้างขวางในการประมาณค่าพารามิเตอร์ เนื่องจากวิธีความควรจะเป็นสูงสุด จะให้ตัวประมาณของข้อมูลตัวอย่างมีคุณสมบัติความน่าจะเป็นเข้าใกล้ค่าที่แท้จริงเหมือนกับขนาดตัวอย่างใหญ่ คือ จะได้ตัวประมาณไม่เอนเอียง และมีประสิทธิภาพ เมื่อเทียบกับการประมาณด้วยวิธีอื่นๆ (สุชาดา กีระนันท์, 2538: 77-82; Hosmer et al., 1998)

หลักในการประมาณด้วยวิธีความควรจะเป็นสูงสุด

สมมติค่าสังเกตมีจำนวนทั้งหมด n ราย (i= 1,2,....,n) สำหรับ i แต่ละรายนั้น ข้อมูลจะประกอบไปด้วย t_i δ_i และ Z_i เมื่อ t_i เป็นตัวแปรแสดงถึงค่าสังเกต δ_i แสดงถึงความสมบูรณ์ของค่าสังเกต ซึ่งจะมีค่าเท่ากับ 1 เมื่อ t_i เป็นค่าสังเกตที่สมบูรณ์ หรือมีค่าเท่ากับ 0 เมื่อค่าสังเกตไม่สมบูรณ์ $Z_i = [1\ Z_{i1}\ Z_{i2} \quad Z_{ik}]$ เป็นเวกเตอร์ค่าของตัวแปรอิสระ (independent value) โดยค่า 1 คือ ส่วนตัด (intercept)

ยกตัวอย่าง การศึกษาข้อมูลที่ไม่สมบูรณ์ทางขวา และมีการตัดออกทางซ้าย (right censor and left truncation) สามารถแสดงได้โดยการจับคู่ของตัวแปรสุ่ม (T, δ) ค่าของ T จะถูกกำหนดตาม $T = \min(t, C_r)$ สำหรับข้อมูลที่มีการถูกตัดออกทางซ้าย คือ จะมีเพียงข้อมูลที่ $t \geq Y$ เท่านั้นที่ถูกสังเกต เมื่อ Y เป็นเวลาที่ค่าสังเกตถูกตัดออก (Elandt-Johnson and Johnson, 1980; Klien et al., 1997: 67)

ฟังก์ชันความน่าจะเป็นหนาแน่นสำหรับข้อมูลที่ไม่สมบูรณ์ คือ

สำหรับ $\delta = 0$

$$P(T,\delta) = \left[\frac{f(t)}{S(y)}\right]^{\delta}\left[\frac{S(c_r)}{S(y)}\right]^{1-\delta}$$

เมื่อค่าสังเกตมีจำนวน n ราย จะได้ฟังก์ชันความควรจะเป็น

$$L(\beta) = \prod_{j=1}^{n}\left[\frac{f(t)}{S(y)}\right]^{\delta_j}\left[\frac{S(c_r)}{S(y)}\right]^{1-\delta_j}$$

การประมาณค่าพารามิเตอร์ในฟังก์ชันความควรจะเป็นข้างต้น จะใช้วิธีนิวตัน-ราฟสันในการคำนวณ โดยให้ $U(\beta)$ เป็นเวกเตอร์ (px1) ของอนุพันธ์อันดับที่ 1 ของลอการิทึม $L(\beta)$ เทียบกับ β และให้ $I(\beta)$ เป็นเมทริกซ์ขนาด (pxp) ของลบอนุพันธ์อันดับที่ 2 ของลอการิทึม $L(\beta)$ เทียบกับ β นั่นคือ

$$U(\beta) = \frac{\partial LL(\beta)}{\partial\beta}$$

$$I(\beta) = -\frac{\partial^2 LL(\beta)}{\partial\beta\partial\beta'}$$

ขั้นตอนการประมาณพารามิเตอร์ด้วยวิธีนิวตัน-ราฟสัน ของเวกเตอร์พารามิเตอร์ β_{j+1} จะเป็นไปตามอนุกรม

$$\beta_{j+1} = \beta_j + I^{-1}(\beta_j)U(\beta_j) \qquad \text{สำหรับ } j = 1,2,\ldots$$

เมื่อ I^{-1} คือ ตัวผกผัน (inverse) ของ I โดยให้ค่าเริ่มต้น β_0 มีค่าเท่ากับศูนย์ และแทนค่า β_0 ทางขวามือในสมการเพื่อหาค่า β_1 เมื่อได้ค่า β_1 แล้ว ก็จะทำการแทนค่าทางขวามือในสมการเพื่อหาค่า β_2 วนทำซ้ำเช่นนี้ไปเรื่อยๆ จนกระทั่งการเปลี่ยนแปลงมากที่สุดในการประมาณค่าพารามิเตอร์จากขั้นหนึ่งไปยังขั้นที่ต่อเนื่องกัน มีค่าน้อยกว่า 0.0001 (Allison, 1995: 83)

2.6 การเลือกตัวแบบเชิงพหุของการวิเคราะห์ข้อมูลการคงอยู่

2.6.1 การเลือกตัวแปรอิสระเข้าตัวแบบเชิงพหุ

เนื่องจากการสร้างตัวแบบเชิงพหุในการวิเคราะห์การคงอยู่ เป็นการศึกษาถึงความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระ ในการพยากรณ์หรือประมาณตัวแปรที่สนใจ โดยที่ทราบค่าของตัวแปรอิสระ การประมาณที่ได้จะมีความถูกต้องหรือไม่ จึงขึ้นอยู่กับตัวแปรอิสระที่เลือกเข้าตัวแบบ

ก) การเลือกตัวแปรอิสระที่เหมาะสมสำหรับ Cox PH model ด้วยวิธีการถดถอยแบบขั้นบันได (stepwise regression) (Hosmer et al., 1998: 180-183)

กระบวนการในการคัดเลือกตัวแปรอิสระเข้า Cox PH model ด้วยวิธีการถดถอยแบบขั้นบันไดเต็มรูปแบบ จะประกอบไปด้วยวิธีการเลือกตัวแปรแบบเลือกเข้า (forward selection) ตามด้วยวิธีการกำจัดตัวแปรแบบถอยหลัง (backward elimination) รวมกัน นั่นคือ

ขั้นที่ A สมมติว่ามีตัวแปรอิสระทั้งหมดอยู่ p ตัว ให้สัญลักษณ์เป็น Z_j โดยที่ $j = 1,2\ldots,p$ ขั้นที่ A จะทดสอบอัตราส่วนความควรจะเป็นบางส่วน และค่า P-value สำหรับระดับนัยสำคัญของแต่ละตัวแปรอิสระ เพื่อเปรียบเทียบลอการิทึมความควรจะเป็นบางส่วนของตัวแบบที่มี Z_j กับลอการิทึมความควรจะเป็นบางส่วนของตัวแบบในขั้น A ที่ไม่มีตัวแปรอิสระใดในตัวแบบ

ตัวสถิติทดสอบ คือ

$$G^{(A)}(j) = -2[L^{(A)}(j) - L(0)] \qquad j = 1,2\ldots, p$$

เมื่อ L(0) เท่ากับลอการิทึมความควรจะเป็นบางส่วนของตัวแบบ A และ $L^{(A)}(j)$ เท่ากับลอการิทึมความควรจะเป็นบางส่วนของตัวแบบที่มีตัวแปรอิสระ Z_j ระดับนัยสำคัญของการทดสอบคือ

$$p^{(A)}(j) = \Pr[\chi^2(1) \geq G^{(A)}(j)]$$

ตัวอักษรในวงเล็บที่ยกกำลังเป็นการบอกถึงลำดับขั้น และสัญลักษณ์ j เป็นการบอกถึงลำดับของตัวแปรอิสระ ตัวแปรอิสระที่เข้าไปในตัวแบบที่ขั้น A นี้ จะเป็นตัวแปรอิสระที่มีนัยสำคัญมากที่สุด และให้สัญลักษณ์เป็น Z_{e_1} เมื่อ

$$p^{(A)}(e_1) = \min_j [p^{(A)}(j)]$$

สำหรับตัวแปร Z_{e_1} ที่ถูกเลือกเข้าไปในตัวแบบ ค่า P-value ของตัวแปรอิสระตัวนี้ จะต้องมีค่าน้อยกว่าเกณฑ์นัยสำคัญที่กำหนดไว้ก่อน (ให้สัญลักษณ์เป็น p_E) ถ้าตัวแปรที่เลือกเข้าไปมีนัยสำคัญ $(p^{(A)}(e_1) < p_E)$ ก็จะทำต่อในขั้น B และหากค่า P-value ที่ได้มีค่านอกเหนือจากนี้ จะหยุดคำนวณ

ขั้น B ขั้นตอนนี้จะเริ่มต้นด้วยตัวแปร Z_{e_1} ในตัวแบบ ดังนั้น Cox PH model ใหม่จะมีตัวแปรอิสระเหลืออยู่ p-1 ตัว ผลลัพธ์นี้จะถูกใช้ในการคำนวณด้วยวิธีอัตราส่วนความควรจะเป็นบางส่วนของตัวแบบที่มีตัวแปรอิสระ 2 ตัวที่เหมาะสมกับตัวแบบ กับตัวแบบที่มีตัวแปรอิสระแค่ 1 ตัว คือ Z_{e_1}

$$G^{(B)}(j) = -2[L^{(B)}(j) - L(Z_{e_1})] \qquad j = 1,2..., p \text{ และ } j \neq e_1$$
(2.6.1)

ค่า P-value สำหรับการทดสอบนัยสำคัญของการเพิ่ม Z_j เข้าไปในตัวแบบที่มี Z_{e_1}

$$p^{(B)}(j) = \Pr[\chi^2(1) \geq G^{(B)}(j)] \tag{2.6.2}$$

ตัวแบบที่ถูกคัดเลือกเข้าไปในขั้น C คือ Z_{e_2} เมื่อ

$$p^{(B)}(e_2) = \min_{j \neq e_1}[p^{(B)}(j)]$$

ถ้าการเลือกตัวแปร Z_{e_2} มีนัยสำคัญ $(p^{(B)}(e_2) < p_E)$ ก็จะทำในขั้น C ต่อไป ถ้าเป็นค่าอื่นนอกเหนือจากนี้จะหยุดคำนวณ

ขั้น C ขั้นตอนนี้จะเริ่มต้นจากตัวแปรเริ่มต้น Z_{e_1} และ Z_{e_2} โดยเริ่มจากหลักการของวิธีการกำจัดตัวแปรแบบถอยหลังในการตรวจสอบความสัมพันธ์ของ Z_{e_1} นั่นคือ ตรวจสอบว่า Z_{e_1} ยังคงมีนัยสำคัญอยู่หรือไม่ หลังจากที่ได้นำตัวแปรอิสระ Z_{e_2} เข้าไปในตัวแบบ ขั้นตอนต่อไปจะทำซ้ำสมการที่ (2.6.1) และ (2.6.2) ด้วยกฎการสลับกันของ 2 ตัวแปร ในการคำนวณนี้ จะดูจากเกณฑ์ความแตกต่างของนัยสำคัญสำหรับการตรวจสอบ ซึ่งให้สัญลักษณ์เป็น p_R โดยจะเลือกค่าที่ $p_R > p_E$ ซึ่งเป็นไปได้ที่จะมีการเข้าและออกของตัวแปรตัวเดิมในขั้นที่ต่อเนื่องกัน ในกรณีนี้สมมติว่าตัวแปรที่เข้าไปในขั้น B ยังคงมีนัยสำคัญ

ดังนั้น ตัวแปรอิสระ p-2 ตัว ก็จะเป็น Cox PH model ใหม่ (รวมถึง Z_{e_1} และ Z_{e_2} ด้วย) ผลลัพธ์ที่ได้จะนำไปคำนวณอัตราส่วนความควรจะเป็นบางส่วน และค่า P-value ของตัวแบบ สำหรับการเพิ่มตัวแปรอิสระตัวใหม่ที่เข้าไปในตัวแบบ

$$G^{(C)}(j) = -2[L^{(C)}(j) - L(Z_{e_1}, Z_{e_2})]$$
$$j = 1,2..., p \text{ และ } j \neq e_1, e_2$$

และ

$$p^{(C)}(j) = \Pr[\chi^2(1) \geq G^{(C)}(j)]$$

ตัวแบบ Z_{e_3} ที่ถูกคัดเลือกเข้ามาในขั้น C จะเป็นตัวแปรที่มีค่า P-value น้อยที่สุด นั่นคือ

$$p^{(C)}(e_3) = \min_{j \neq e_1, e_2}[p^{(C)}(j)]$$

ถ้า $p^{(C)}(e_3) < p_E$ จะทำการคำนวณต่อไปยังขั้นตอน D สำหรับกรณีอื่น จะหยุดคำนวณ

ขั้น D ขั้นตอนนี้จะทำซ้ำกับขั้นตอน C ในการพิจารณาคัดตัวแปรอิสระออกจากตัวแบบ และเมื่อกรณีตัวแปรทุกตัวที่เข้าไปในตัวแบบขั้นก่อนหน้านี้ยังคงมีนัยสำคัญ กระบวนการคัดเลือกก็จะทำซ้ำขั้นตอนที่ A ถึง C จนกระทั่งขั้นตอนสุดท้าย ขั้นที่ Z

ขั้น Z เป็นขั้นตอนสุดท้าย โดยผลลัพธ์ที่เกิดขึ้นจะอยู่ภายใต้ 2 ลักษณะ คือ (1) ตัวแปรทุกตัวยังคงอยู่ในตัวแบบ ไม่มีตัวแปรตัวใดถูกคัดเลือกออก หรือ (2) ไม่มีตัวแปรใดอยู่ในตัวแบบ คือ $p^{(Z)}(j) > p_E$ ที่จุดนี้ จะไม่มีตัวแปรที่ถูกเลือกเข้าและตัวแปรในตัวแบบถูกคัดเลือกออก

ข) เกณฑ์การคัดเลือกตัวแบบที่ดีที่สุดโดย Akaike 's Information Criterion

Akaike ได้เสนอเกณฑ์ในการคัดเลือกตัวแบบ ซึ่งเป็นเกณฑ์ที่พิจารณาจากการประมาณความคลาดเคลื่อน รวมเข้ากับค่าสังเกตสารสนเทศ (Observed Information) ของค่าสังเกต และใช้แนวคิดจากค่าต่ำสุดของสารสนเทศ Kulback-Leibler เพื่อนำมาใช้ในการปรับค่าประมาณของการพยากรณ์ให้มีความแม่นยำมากขึ้น โดยมีข้อกำหนดเบื้องต้นคือ ตัวประมาณได้มาจากค่าความควรจะเป็นสูงสุด (พจนา แววสวัสดิ์, 2544)

เกณฑ์การคัดเลือกตัวแบบโดยสารสนเทศของ Akaike คือ

$$AIC_k = -2\log(ML_k) + 2p_k$$

เมื่อ ML_k คือ ฟังก์ชันความควรจะเป็นสูงสุดของตัวแบบที่ k

p_k คือ จำนวนพารามิเตอร์ของตัวแบบที่ k

โดยตัวแบบที่ให้ค่า AIC ต่ำสุด จะเป็นตัวแบบที่ให้ค่าพยากรณ์แม่นยำและถูกต้อง

2.6.2 การตรวจสอบตัวแบบเชิงพหุสำหรับการวิเคราะห์ข้อมูลการคงอยู่

ค่าตกค้าง (residuals) ที่ใช้ในการตรวจสอบตัวแบบ เพื่อพิจารณาดูความเหมาะสมระหว่างตัวแบบกับข้อมูลที่นำมาวิเคราะห์ มีอยู่หลายประเภท คือ

ก) Cox-Snell residuals

ในการวิเคราะห์ข้อมูลการคงอยู่ทั่วไป Cox-Snell residuals ที่เสนอโดย Cox and Snell จะเป็นค่าตกค้างที่นิยมใช้กันอย่างกว้างขวาง โดยกำหนดว่าสำหรับแต่ละหน่วยสังเกตที่ j เมื่อ j = 1,2,…,n ค่าตกค้างประเภทนี้จะมีสูตรการคำนวณเป็น (Collet, 1994: 150)

$$r_j = \hat{H}_0(t_j)\exp(\hat{\beta}' Z_j)$$

เมื่อ $\hat{H}_0(t_j)$ คือ ค่าประมาณฟังก์ชันภาวะภัยพื้นฐานสะสม (cumulative baseline hazard function) ที่เวลา t_j (เมื่อ t_j คือ ค่าที่สังเกตได้จากหน่วยสังเกต) Cox-Snell residuals คือ ค่าของ $\hat{H}_j(t_j) = -\log \hat{S}_j(t_j)$ เมื่อ $\hat{H}_j(t_j)$ และ $\hat{S}_j(t_j)$ เป็นค่าประมาณที่ได้ของฟังก์ชันภาวะภัยสะสม และฟังก์ชันการคงอยู่สะสมของหน่วยสังเกตลำดับที่ j ณ เวลา t_j

ค่าตกค้างนี้สามารถคำนวณได้จากผลลัพธ์ทั่วไปในสถิติคณิตศาสตร์ สำหรับการแจกแจงของฟังก์ชันตัวแปรสุ่ม ในที่นี้ ถ้า T เป็นตัวแปรสุ่มที่สัมพันธ์กับเวลาการคงอยู่ของหน่วยสังเกต และ S(t) คือ ฟังก์ชันการคงอยู่ที่สมนัยกัน ดังนั้น ตัวแปรสุ่ม Y = - log S(t) จะมีการแจกแจงแบบเลขชี้กำลังที่มีค่าเฉลี่ยเท่ากับหนึ่ง (เมื่อไม่พิจารณารูปแบบของ S(t))

ข) Martingale residuals

Fleming and Harrington ได้เสนอค่าตกค้างประเภทนี้ โดยได้กำหนดว่าสำหรับหน่วยสังเกตที่ j ในตัวอย่าง จะมีเวกเตอร์ $Z_j(t)$ ของตัวแปรอิสระที่ไม่ขึ้นอยู่กับเวลา (time-independent covariate) ให้ $N_j(t)$ มีค่าเท่ากับ 1 ที่เวลา t ถ้าหน่วยสังเกตประสบกับเหตุการณ์ที่สนใจ และเป็น 0 ถ้าหน่วยสังเกตยังไม่ประสบกับเหตุการณ์ที่สนใจ $n_j(t)$ เป็นค่าที่แสดงว่าหน่วยสังเกตที่ j นี้ อยู่ภายใต้การศึกษาก่อนจะถึงเวลา t และ b เป็นเวกเตอร์ของพารามิเตอร์การถดถอย (Collet, 1994: 153)

$$\hat{M}_j = N_j(\infty) - \int_0^\infty n_j(t)\exp[b'Z_j(t)]d\hat{H}_0(t) \qquad j = 1,2,\ldots, n$$

เมื่อข้อมูลมีค่าไม่สมบูรณ์ทางขวา และตัวแปรอิสระทุกตัวถูกกำหนดที่เวลาเริ่มต้นของการศึกษา ดังนั้น Martingale residuals จะลดรูปลงเหลือ

$$\hat{M}_j = \delta_j - \hat{H}_0(T_j)\exp\left(\sum_{k=1}^{p} Z_{jk}b_k\right) = \delta_j - r_j \qquad j = 1,2,\ldots, n$$

Martingale residuals จะมีค่าได้ตั้งแต่ 1 ถึง -∞ โดยค่าตกค้างจะมีค่าเป็นลบเมื่อ $\delta_i = 0$ สำหรับค่าสังเกตไม่สมบูรณ์

ดังสมการข้างต้น สามารถพิจารณา Martingale residuals เหมือนกับความแตกต่างระหว่างจำนวนค่าสังเกตของการเสียชีวิตสำหรับหน่วยสังเกตที่ j ในช่วง $(0,t_j)$ และค่าประมาณของจำนวนค่าคาดหวังบนตัวแบบที่เหมาะสม หมายเหตุว่า จำนวนค่าสังเกตของการเสียชีวิตเป็น 1 ถ้าค่าสังเกต t_j เป็นค่าสังเกตสมบูรณ์ และมีค่าเท่ากับ 0 ถ้าเป็นค่าสังเกตไม่สมบูรณ์ ในเทอมของ r_j คือ ค่าประมาณของฟังก์ชันภาวะภัยสะสม สำหรับหน่วยสังเกตลำดับที่ j ตลอดช่วง $(0,t_j)$

ค) Deviance residuals

ค่าตกค้างประเภทนี้ถูกคิดขึ้นโดย Therneau, Grambsch and Fleming ซึ่งมีลักษณะการแจกแจงเป็นแบบสมมาตร กำหนดโดย (Collet, 1994: 153-154)

$$D_j = sign(\hat{M}_j)\{-2[\hat{M}_j + \delta_j \ln(\delta_j - \hat{M}_j)]\}^{1/2} \qquad (2.6.3)$$

เมื่อ $\hat{M}_j$ คือ Martingale residuals สำหรับหน่วยสังเกตที่ j sign เป็นเครื่องหมายที่แสดงถึงค่าสัมประสิทธิ์ โดยมีค่าเป็น +1 ถ้า $\hat{M}_j$ เป็นค่าบวก และมีค่าเท่ากับ –1 ถ้า $\hat{M}_j$ มีค่าเป็นลบ ดังนั้น sign($\hat{M}_j$) ทำให้ทราบว่า Deviance residuals จะมีเครื่องหมายเช่นเดียวกับ Martingale residuals

ค่าตกค้างประเภทนี้เกิดมาจาก Deviance Statistics ซึ่งเป็นตัวสถิติที่ใช้ในการประเมินความเหมาะสมของตัวแบบปัจจุบัน เทียบกับความเบี่ยงเบนจากตัวแบบที่เหมาะสมที่สุดจากข้อมูล ตัวแบบท้ายสุดจะเรียกว่า ตัวแบบสมบูรณ์ (full model) และเป็นตัวแบบที่พารามิเตอร์ β เป็นเหมือนความแตกต่างสำหรับแต่ละหน่วยสังเกต ตัวสถิตินี้ให้โดย

$$D = -2\{\log\hat{L}_C - \log\hat{L}_f\}$$

เมื่อ $\hat{L}_C$ คือ ความควรจะเป็นบางส่วนสูงสุด (Maximized Partial Likelihood) ภายใต้ตัวแบบปัจจุบัน และ $\hat{L}_f$ เป็นความควรจะเป็นบางส่วนสูงสุดภายใต้ตัวแบบสมบูรณ์ โดยตัวแบบที่ดีที่สุด จะเป็นตัวแบบที่มีค่าเบี่ยงเบนน้อยที่สุด

Deviance residuals อาจพิจารณาได้เหมือนกับกรณี Martingale residuals ที่ถูกแปลงรูป (transform) เพื่อให้ได้ค่าสมมาตรรอบจุดศูนย์ เมื่อตัวแบบที่พิจารณาเป็นตัวแบบที่เหมาะสม ดังที่ได้ทราบในตอนต้นแล้วว่า Martingale residuals จะมีค่าอยู่ในช่วง (-∞,1) สำหรับ Martingale residuals ที่มีค่าติดลบมาก รวมทั้งในส่วนของเครื่องหมายปีกกาดังสมการที่ (2.6.3) ก็ถูกจำกัดอยู่กับ Martingale residuals รากที่สอง(square root) ของจำนวนนี้ จึงได้ค่าตกค้างที่มีค่าเข้าใกล้ศูนย์ ดังนั้น Martingale residuals ในช่วงระยะ (-∞,0) จึงลดลงเข้าสู่ศูนย์

การพิจารณา Martingale residuals ในช่วง (0,1) ส่วนของ $\delta_j \ln(\delta_j - \hat{M}_j)$ ในสมการ (2.6.3) จะมีเฉพาะค่าที่ไม่เป็นศูนย์เท่านั้นสำหรับค่าสังเกตสมบูรณ์ รวมทั้งจะได้ $\ln(1-\hat{M}_j)$ มีค่าติดลบมาก เช่นเดียวกับกรณีถ้า $\hat{M}_j$ มีค่าเข้าใกล้หนึ่ง หรือ 1 -$\hat{M}_j$ ที่มีค่าเข้าใกล้ศูนย์

ง) Score residuals

ให้ b เป็นค่าประมาณพารามิเตอร์ของสัมประสิทธิ์ β และ $\hat{H}_0(t)$ เป็นค่าประมาณ $h_0(t)$ ดังเช่นที่ได้นิยามไว้ใน Martingale residuals $N_j(t)$ แสดงถึงหน่วยสังเกตที่ j ที่ประสบกับเหตุการณ์ที่สนใจ และ $n_j(t)$ เป็นค่าที่แสดงว่าหน่วยสังเกตที่ j นี้ อยู่ภายใต้การศึกษาก่อนจะถึงเวลา t สำหรับตัวแปรอิสระที่ k เมื่อ k = 1,2,...p ให้

$$\bar{Z}_k(t) = \frac{\sum_{j=1}^{n} n_j(t) Z_{jk}(t) \exp[b'Z_j(t)]}{\sum_{j=1}^{n} n_j(t) \exp[Z_j(t)]}$$

เมื่อ $\hat{M}_j(t)$ เป็น Martingale residuals ที่เวลา t สำหรับหน่วยสังเกตที่ j

$$\hat{M}_j = N_j(t) - \int_0^t n_j(t) \exp[b'Z_j(t)] d\hat{H}_0(t) \qquad j = 1,2,..., n$$

Score residuals สำหรับตัวแปรอิสระที่ k และหน่วยสังเกตที่เวลา t จะถูกกำหนดโดย

$$S_{jk}(t) = \int_0^t \{Z_{jk}(s) - \bar{Z}_k(t)\} d\hat{M}_j(t)$$

กรณีที่ตัวแปรอิสระทุกตัวถูกกำหนด ณ เวลาเริ่มต้น Score residuals ที่เวลาสังเกต จะเป็น

$$S_{jk}(t) = \delta_j n_j(t)[Z_{jk} - \bar{Z}_k(T_j)] - \sum_{t_h \le t} [Z_{jk} - \bar{Z}_k(t_h)] n_j(t_h) \exp\{b'Z_j\}[\hat{H}_0(t_h) - \hat{H}_0(t_{h-1})] \tag{2.6.4}$$

สำหรับแต่ละหน่วยสังเกต n จะกำหนดให้ Score process สำหรับตัวแปรอิสระที่ k เป็นเช่นเดียวกับ

$$U_k(t) = \sum_{j=1}^{n} S_{jk}(t)$$

โดยกระบวนการนี้ เป็นการหาอนุพันธ์บางส่วนอันดับที่หนึ่งของฟังก์ชันความควรจะเป็นบางส่วนสำหรับ Cox PH model ที่เหมาะสม ดังนั้น กระบวนการ $W_k(t) = U_k(t)$ x ค่าเบี่ยงเบนมาตรฐานของ b_k จะลู่เข้าสู่ Brownian bridge (โดยมีข้อแม้ว่า $Cov(b_k, b_k') = 0$ สำหรับ $k = k'$)

2.7 งานวิจัยที่เกี่ยวข้อง

การศึกษาด้านการขาดอายุของกรมธรรม์ประกันชีวิต ส่วนใหญ่งานวิจัยมักจะเน้นการวิเคราะห์ไปในเชิงคุณภาพ ในเรื่องการให้ความสำคัญเกี่ยวกับบทบาทและประสิทธิภาพของตัวแทนประกันชีวิต ในการป้องกันการสิ้นผลบังคับลงก่อนกำหนดของกรมธรรม์ประกันชีวิต แต่สำหรับการศึกษาที่ใช้วิธีการทางด้านสถิติเข้ามาช่วยในการวิเคราะห์ข้อมูลเกี่ยวกับการสิ้นผลบังคับลงของกรมธรรม์ที่ผ่านมายังมีอยู่น้อยมาก

Willet (1980) ได้เน้นถึงความสำคัญของการมีสัมพันธไมตรีที่ดี การเพิ่มโอกาสในการสร้างความนิยม ความมั่นใจ และความไว้วางใจต่อผู้เอาประกันภัย (cited in Lian et al., 1998: 102) ทางด้าน Santos (1992) เชื่อว่า ตัวแทนประกันชีวิตสามารถป้องกันการสิ้นผลบังคับลงของกรมธรรม์ประกันชีวิตเป็นอย่างดี ด้วยการยังคงทัศนคติของการให้บริการ เข้าใจในความต้องการของผู้เอาประกันภัย และให้ความกระจ่างในรายละเอียดเกี่ยวกับเงื่อนไขของกรมธรรม์ พร้อมทั้งแนะนำวิธีการแก้ไขปัญหา รวมถึงการรักษาสัญญาที่ให้ไว้ (cited in Lian et al., 1998: 103) สอดคล้องเช่นเดียวกับผลการศึกษาของ บริบูรณ์ ดิฐกมล (2525) ที่ทำการสัมภาษณ์กลุ่มผู้เอาประกันภัยที่มีการขาดอายุ หรือต่ออายุกรมธรรม์ในปีพ.ศ. 2524 จำนวน 196 คน ในเขตกรุงเทพมหานคร นนทบุรี ปทุมธานี และสมุทรปราการ พบว่า การตัดสินใจเลือกแบบประกันชีวิตของผู้เอาประกันภัยส่วนใหญ่ ร้อยละ 58.16 เลือกตามที่ตัวแทนประกันชีวิตเสนอมา อีกร้อยละ 41.84 เลือกตามความพอใจของตัวเอง ซึ่งโดยส่วนหนึ่งของผู้ที่เลือกตามความพอใจของตัวเองนี้ ได้เลือกแบบประกันชีวิตตามที่ตัวแทนประกันชีวิตเสนอมา โดยมีอยู่เพียงไม่กี่แบบเท่านั้นที่นำมาเสนอ

ในขณะที่ผู้วิจัยบางท่าน ศึกษาเกี่ยวกับมาตรการการจัดการการขาดอายุลงของกรมธรรม์ประกันชีวิต ไม่ว่าจะเป็น Goodwin (1993) ที่แนะนำให้ผู้รับประกันภัยสร้างจิตสำนึกแก่ตัวแทนประกันชีวิต ปลูกฝังจริยธรรม ส่งเสริมผู้เอาประกันภัยได้เลือกทำกรมธรรม์ประกันชีวิตที่ตนเองมีความสนใจมากที่สุด แนะนำให้มีบทลงโทษเพื่อบังคับตัวแทนประกันชีวิต สำหรับกรมธรรม์ที่สิ้นผลบังคับลงก่อนกำหนด (cited in Lian et al., 1998: 103) เช่นเดียวกับ สุทธิ รจิตรังสรรค์ ในฐานะนายกสมาคมประกันชีวิตไทย ที่ได้มีการตั้งมาตรฐานการกำกับดูแลบริษัทที่เป็นสมาชิกของสมาคมประกันชีวิตไทยตามนโยบายของกรมการประกันภัย โดยเฉพาะการป้องกันปัญหาตัวแทนประกันชีวิตโยกย้ายบริษัทในลักษณะที่ผิดธรรมชาติ ซึ่งอาจส่งผลกระทบต่อบริษัทสมาชิก และผู้เอาประกันภัยที่ต้องเสียประโยชน์จากการโยกย้ายกรมธรรม์ โดยอาจดำเนินการด้วยการระงับไม่ต่อใบอนุญาตให้กับตัวแทนประกันชีวิต โดยกรมการประกันภัยเป็นผู้พิจารณา (บรรยงค์ สุวรรณผ่อง, 2545: 26)

Lauzenheiser and Barks (1991) และ Howard (1997) มุ่งเน้นความสนใจไปในด้านการออกรูปแบบกรมธรรม์ เพื่อที่จะให้ผู้ถือกรมธรรม์ดำรงรักษากรมธรรม์ของตนเองไว้ให้นานที่สุด โดย Lauzenheiser and Barks สนับสนุนการออกแบบกรมธรรม์ประกันชีวิต ที่เตรียมลักษณะความคุ้ม

ครองผลประโยชน์ครบวงจร และสร้างความเชื่อมั่นแก่ผู้ถือกรมธรรม์ในการรักษากรมธรรม์ว่า จะได้รับผลประโยชน์เป็นที่แน่นอนเมื่อมีเหตุการณ์เกิดขึ้น ในส่วน Howard แนะนำว่า การขาดอายุของกรมธรรม์ประกันชีวิตสามารถพัฒนาดีขึ้นได้จากการเข้าเยี่ยมเยียนของตัวแทนประกันชีวิต การให้บริการ และส่งข่าวสารเกี่ยวกับข้อมูลอย่างสม่ำเสมอ รวมทั้งสามารถเปลี่ยนแปลงกรมธรรม์ได้ตลอดเวลาตามความต้องการของผู้เอาประกันภัย (cited in Lian et al., 1998: 103) ต่างจาก Sweeney (1996) Hall (1990) และ Berlin and Janette (1990) ที่ได้ให้ความสำคัญด้านการตลาดและสื่อโฆษณา โดยได้แสดงความคิดเห็นว่า การโฆษณาที่ไม่ถูกต้องกับลักษณะงานขายในทางปฏิบัติ จะส่งผลในทางตรงกันข้ามกับอัตราการคงอยู่ และยอดขายของกรมธรรม์ใหม่ (cited in Lian et al., 1998: 103)

สำหรับในประเทศไทย จากการศึกษาของ บริบูรณ์ คิฐกมล (2525) นอกเหนือจากที่ได้กล่าวมาข้างต้น บริบูรณ์ คิฐกมล ได้เพิ่มความสนใจเกี่ยวกับการศึกษาหาสาเหตุของการขาดอายุกรมธรรม์ประกันชีวิต มากกว่าจะให้ความสนใจในอัตราการขาดอายุของกรมธรรม์ โดยได้วิเคราะห์ข้อมูลทั่วไปเกี่ยวกับผู้เอาประกันภัย พบว่า กรมธรรม์ที่ขาดอายุส่วนใหญ่นั้น อายุของผู้เอาประกันภัยจะอยู่ในช่วง 25-44 ปี สถานภาพสมรสแล้ว มีระดับการศึกษาส่วนใหญ่ต่ำกว่าอนุปริญญาหรือเทียบเท่า ผู้เอาประกันภัยประกอบอาชีพลูกจ้างเอกชนมากที่สุด รองลงมาคือประกอบธุรกิจส่วนตัว สำหรับอาชีพอื่นๆ มีเพียงเล็กน้อย รายได้ต่อเดือนของผู้เอาประกันภัยน้อยกว่า 6,000 บาท ส่วนใหญ่ไม่มีเงินออม ในด้านข้อมูลเกี่ยวกับกรมธรรม์ประกันชีวิตที่มีการสิ้นผลบังคับลงก่อนกำหนด จำนวนเงินเอาประกันภัยอยู่ระหว่าง 50,001 บาท ถึง 150,000 บาท เบี้ยประกันภัยต่อปีอยู่ระหว่าง 2,001 บาท ถึง 4,000 บาท โดยจะชำระเบี้ยประกันภัยเป็นรายปีมากที่สุด นอกจากนี้ เมื่อ บริบูรณ์ คิฐกมล ทำการสัมภาษณ์ถึงสาเหตุจูงใจในการตัดสินใจทำประกันชีวิต พบว่า มีเพียง 2 สาเหตุคือ ผู้เอาประกันภัยทราบถึงประโยชน์ของการทำประกันชีวิต สาเหตุต่อมาคือ เกรงใจในตัวแทนประกันชีวิต สำหรับเหตุผลของการขาดส่งเบี้ยประกันภัย คือ ฐานะการเงินที่เปลี่ยนแปลงไป ทำให้ผู้เอาประกันภัยไม่สามารถที่จะชำระเบี้ยประกันภัยต่อไปได้ แบบประกันชีวิตที่เลือกไว้ไม่เหมาะสม หรือเป็นเพราะโดยส่วนตัวแล้ว ไม่เห็นประโยชน์ของการทำประกันชีวิต แต่ที่ตัดสินใจเลือกทำเพราะเห็นแก่ตัวแทนประกันชีวิต หรือเห็นแก่คนที่แนะนำตัวแทนประกันชีวิตมา

สำหรับการศึกษาที่ใช้วิธีการทางด้านสถิติเข้ามาช่วยในการวิเคราะห์ข้อมูลเกี่ยวกับการสิ้นผลบังคับลงก่อนกำหนดของกรมธรรม์ประกันชีวิต ได้มีผู้เชี่ยวชาญหลายท่านเสนอวิธีวิเคราะห์แบบต่างๆ อาทิเช่น Lian et al. (1993) ได้ประยุกต์ใช้เกี่ยวกับการวิเคราะห์การถดถอยเชิงพหุ (Multivariate Regression Analysis) ในการศึกษาถึงปัจจัยที่มีผลต่อการสิ้นผลบังคับลงก่อนกำหนดของกรมธรรม์ประกันชีวิต ผลการศึกษาพบว่า ตัวแปรในเรื่องของ อายุที่เริ่มทำประกันชีวิต เพศผู้เอาประกันภัย งวดการชำระเบี้ยประกันภัย แบบกรมธรรม์ประกันชีวิต และวิธีการชำระเบี้ยประกันภัย มีนัยสำคัญต่อระยะเวลาการมีผลบังคับอยู่ของกรมธรรม์ประกันชีวิต (cited in Lian et al., 1998: 104) คล้ายกับ

ผลการศึกษาของ Renshaw and Haberman (1986) ที่ได้ทำการพิจารณาการสิ้นผลบังคับลงก่อนกำหนดของกรมธรรม์ประกันชีวิตจากรูปแบบกรมธรรม์ประกันชีวิตแบบต่างๆ ในรอบปีค.ศ. 1976 เพื่อนำมาพิจารณาเหตุผลที่ผู้ถือกรมธรรม์ในประเทศสหรัฐอเมริกาไม่ต่ออายุกรมธรรม์ Renshaw et al. ได้นำตัวแบบเชิงเส้นทั่วไปมาทดสอบปัจจัยต่างๆ (เช่น แบบกรมธรรม์ประกันชีวิต สาขาของบริษัทประกันชีวิต งวดการชำระเบี้ยประกันภัย และอื่นๆ) ที่คาดว่าจะมีผลกระทบต่ออัตราการสิ้นผลบังคับลงของกรมธรรม์ประกันชีวิต ซึ่งผลการศึกษาพบว่า ปัจจัยที่ส่งผลกระทบสำคัญต่อระยะเวลาการคงอยู่ของกรมธรรม์คือ ระยะเวลาเอาประกันภัย และแบบกรมธรรม์ประกันชีวิต (cited in Lian et al., 1998: 103-104)

ในขณะที่ Lian et al. (1998) ได้นำตัวแบบเชิงเส้นทั่วไปมาวิเคราะห์กรมธรรม์ประกันชีวิตที่สิ้นผลบังคับก่อนกำหนดในรอบปีค.ศ. 1990 เนื่องจากการไม่ชำระเบี้ยประกันภัยของผู้ถือกรมธรรม์ในสมาคมประกันชีวิตของประเทศสิงคโปร์ จำนวน 48,243 กรมธรรม์ โดยปัจจัยที่นำมาพิจารณาประกอบไปด้วย ระยะเวลาที่กรมธรรม์สิ้นผลบังคับลง (ปี) อายุที่เริ่มทำประกันชีวิต เพศผู้เอาประกันภัย แบบกรมธรรม์ประกันชีวิต วิธีการชำระเบี้ยประกันภัย งวดการชำระเบี้ยประกันภัย สภาพของกรมธรรม์ (เช่น การขาดอายุ การเวนคืน การเปลี่ยนเป็นกรมธรรม์ใช้เงินสำเร็จ หรือการเปลี่ยนเป็นกรมธรรม์ขยายเวลา) จำนวนเงินเอาประกันภัย สถานะของการบริการ (มีตัวแทน/ไม่มีตัวแทน) ขนาดของบริษัทรับประกันภัย และระยะเวลาเอาประกันภัย ซึ่งผลการศึกษาพบว่า ปัจจัยในเรื่องของระยะเวลาเอาประกันภัย เป็นเพียงปัจจัยเดียวที่ไม่มีนัยสำคัญต่อระยะเวลาการคงอยู่ของกรมธรรม์ประกันชีวิตที่สิ้นผลบังคับลงก่อนกำหนด

ในเวลาต่อมา Lian et al. (1998) จึงได้เพิ่มขั้นตอนในการศึกษา และปรับเปลี่ยนวิธีการศึกษาใหม่ โดยใช้การวิเคราะห์การคงอยู่เข้ามาศึกษาแทน และคำนวณฟังก์ชันการคงอยู่ ค่าเฉลี่ยจำนวนปีการคงอยู่ของกรมธรรม์โดยวิธีตารางชีพ (Life-Table Method) รวมทั้งประมาณค่าพารามิเตอร์ของแต่ละปัจจัยที่มีนัยสำคัญต่อการสิ้นผลบังคับลงก่อนกำหนดของกรมธรรม์ประกันชีวิต ด้วยตัวแบบที่มีการแจกแจงแบบไวบูลล์ ผลการวิจัยพบว่า โดยส่วนใหญ่ที่กรมธรรม์ประกันชีวิตจะสิ้นผลบังคับลงก่อนกำหนดนั้นจะอยู่ในช่วง 2 ปีแรก และมีค่าเฉลี่ยอยู่ประมาณ 3.569 ปี โดยปัจจัยที่มีผลต่อระยะเวลาการสิ้นผลบังคับลงของกรมธรรม์ประกันชีวิตประเภทสามัญ ไม่ค่อยมีความแตกต่างจากการศึกษาด้วยวิธีตัวแบบเชิงเส้น (เพียงแต่การศึกษาในครั้งนี้ Lian et al. ได้ตัดตัวแปรในเรื่อง สภาพของกรมธรรม์ ออกจากการศึกษา) โดยปัจจัยที่ว่านี้คือ อายุเริ่มทำประกันชีวิต เพศผู้เอาประกันภัย แบบกรมธรรม์ประกันชีวิต งวดการชำระเบี้ยประกันภัย วิธีการชำระเบี้ยประกันภัย จำนวนเงินเอาประกันภัย สถานะการให้บริการ นอกจากนี้ เพื่อความเข้าใจเกี่ยวกับพฤติกรรมความต้องการของผู้เอาประกันภัย ในการเป็นแนวทางให้กับบริษัทประกันชีวิตเข้าใจถึงปัจจัยที่มีอิทธิพลต่อพฤติกรรมของผู้บริโภค Wu, Lian and Koh (2001) จึงได้นำวิธีแผนผังต้นไม้ (Decision trees) เข้ามาช่วยในการวิเคราะห์ครั้งนี้ด้วย