บทที่ 2

ตรวจเอกสาร

2.1 ตัวแบบการถดถอยเชิงเส้น (Linear Regression Model)

ตัวแบบการถดถอยเชิงเส้นที่มีตัวแปรตาม $Y_1, Y_2, ..., Y_n$ เป็นตัวแปรสุ่มต่อเนื่องที่เป็นอิสระต่อกัน และ $X = [1, \underset{\sim}{x}_1, \underset{\sim}{x}_2, ..., \underset{\sim}{x}_n]^T$ เป็นเมตริกซ์ของตัวแปรอิสระที่มีขนาด $n \times (p+1)$ โดยที่ตัวแปรอิสระเป็นตัวแปรเชิงปริมาณ หรือเชิงคุณภาพก็ได้ ถ้า Y_i และ $\underset{\sim}{x}_i$, $i = 1, 2, ..., n$ มีความสัมพันธ์เชิงเส้นต่อกันแล้วตัวแบบที่ใช้แทนความสัมพันธ์ดังกล่าวคือ ตัวแบบการถดถอยเชิงเส้น

$$\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon} \tag{2.1}$$

$\underset{\sim}{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$ คือเวคเตอร์ของค่าคลาดเคลื่อนสุ่มที่เป็นอิสระต่อกัน และ $\varepsilon_i \sim N(0, \sigma^2)$ และ $\underset{\sim}{Y} = [Y_1, Y_2, \ldots, Y_n]^T$ เป็นเวคเตอร์ของตัวแปรสุ่มที่ $Y_i \sim N(\mu_i, \sigma^2)$ ดังนั้นค่าเฉลี่ยของ Y_i คือ $E(Y_i) = \mu_i = \underset{\sim}{x}_i^T\underset{\sim}{\beta}$ โดยที่ $\underset{\sim}{x}_i = [1, x_{1i}, ..., x_{pi}]^T$ เป็นเวคเตอร์ของตัวแปรอิสระ และ $\underset{\sim}{\beta} = [\beta_0, \beta_1, ..., \beta_p]^T$ เป็นเวคเตอร์ของพารามิเตอร์ ดังนั้นฟังก์ชันการถดถอยเชิงเส้นของค่าเฉลี่ยของตัวแปรสุ่ม Y_i คือ

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} \tag{2.2}$$

ถ้าประมาณค่าพารามิเตอร์ $\beta_0, \beta_1, ..., \beta_p$ โดยวิธีกำลังสองน้อยสุด (Least square method) แล้วค่าประมาณของ $\beta_0, \beta_1, ..., \beta_p$ เป็นค่าที่ทำให้ผลบวกกำลังสองของค่าคลาดเคลื่อนมีค่าน้อยสุด นั่นคือจะต้องหาค่าของ $\beta_0, \beta_1, ..., \beta_p$ ที่ทำให้ $\sum_{i=1}^{n}\varepsilon_i^2 = \sum_{i=1}^{n}\left[Y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip})\right]^2$ มีค่าน้อยสุด

2.2 ตัวแบบเชิงเส้นวางนัยทั่วไป (Generalized Linear Models)

ตัวแบบเชิงเส้นวางนัยทั่วไป (Generalized linear models: glms) ซึ่งเสนอโดย Nelder and Wedderburn (1972) เป็นตัวแบบเชิงสถิติที่เป็นพื้นฐานของการจำลองตัวแบบเชิงสถิติ ที่ประกอบด้วย 3 องค์ประกอบได้แก่

1) Random component ได้แก่ตัวแปรสุ่ม Y ที่มีการแจกแจงความน่าจะเป็น เป็นสมาชิกของ Exponential family distribution โดยมีฟังก์ชันความหนาแน่นน่าจะเป็น (Probability density function: p.d.f) คือ

$$f(y;\theta,\phi) \quad = \quad \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

ที่ทราบรูปแบบของ $b(\theta)$, $c(y,\phi)$ และ $a(\phi)$ โดยที่ $a(\phi)$ คือฟังก์ชันของ Dispersion parameter Log – likelihood function ของ θ_i เมื่อกำหนด $y_1, y_2,...,y_n$ คือ

$$l \quad = \quad \sum_{i=1}^{n}\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, a(\phi))\right\} \tag{2.3}$$

l มีคุณสมบัติดังนี้คือ $E\left[\frac{\partial l}{\partial \theta_i}\right] = 0$ และ $E\left[\frac{\partial^2 l}{\partial \theta_i^2}\right] + Var\left[\frac{\partial l}{\partial \theta_i}\right] = 0$ โดยใช้คุณสมบัติดังกล่าวสามารถพิสูจน์ได้ว่า $E(Y_i) = b'(\theta_i) = \mu_i$ และ $Var(Y_i) = a(\phi)b''(\theta_i) = a(\phi)V(\mu_i)$ เมื่อ $b'(\theta_i)$ และ $b''(\theta_i)$ คืออนุพันธ์อันดับที่ 1 และ 2 ของ $b(\theta_i)$ เทียบกับ θ_i ตามลำดับ และ $V(\mu_i)$ คือ Variance function ของตัวแบบ

2) Systematic component เป็นฟังก์ชันเชิงเส้นของตัวแปรอิสระซึ่งเรียกว่า Linear predictors นิยามด้วย $\eta_i = \underset{\sim}{x}_i^T\underset{\sim}{\beta}$ โดยที่ $\underset{\sim}{x}_i = \left[1, x_{1i},..., x_{pi}\right]^T$; $i = 1, 2, \ldots, n$

3) Link function คือฟังก์ชัน $g(\mu_i)$ ที่ใช้เชื่อมระหว่าง μ_i และ Linear predictors $\eta_i = \underset{\sim}{x}_i^T\underset{\sim}{\beta}$ นั่นคือ

$$g(\mu_i) = \eta_i = \underset{\sim}{x}_i^T\underset{\sim}{\beta} = \beta_0 + \beta_1 x_{11} + ... + \beta_p x_{ip} \tag{2.4}$$

ฟังก์ชัน $g(\mu_i)$ ขึ้นอยู่กับ θ_i

ตัวอย่างของการแจกแจงที่เป็นสมาชิกของ glms ได้แก่ การแจกแจงปรกติ การแจกแจงปัวซง และการแจกแจงทวินาม ทั้งนี้เนื่องจากฟังก์ชันความน่าจะเป็นของสามการแจกแจงนี้เป็นสมาชิกของ Exponential family distribution ซึ่งสามารถตรวจสอบได้ ดังนี้

- ถ้า Y มีการแจกแจงปรกติที่มีค่าเฉลี่ย μ และความแปรปรวน σ^2 หรือแทนด้วย $Y \sim N(\mu, \sigma^2)$ ฟังก์ชันความหนาแน่นน่าจะเป็นของ Y ในรูป Exponential family distribution คือ

$$f\left(y;\mu,\sigma^2\right) = \exp\left\{\frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} + \left(\frac{-y^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\right)\right\} \quad \text{ที่มี } \theta = \mu,$$

$b(\theta) = \frac{1}{2}\mu^2$, $a(\phi) = \sigma^2$ และ $c(y,\phi) = \frac{-y^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)$ ดังนั้น $E(Y) = \mu$, $V(\mu) = 1$ และ Link function คือ Identity link ที่ $g(\mu) = \mu$ ซึ่งเป็นที่มาของตัวแบบการถดถอยเชิงเส้นที่แสดงในสมการ (2.2)

- ถ้า Y มีการแจกแจงปัวซง ที่มีค่าเฉลี่ย μ หรือแทนด้วย $Y \sim Pois(\mu)$ ฟังก์ชันมวลน่าจะเป็นของ Y ในรูป Exponential family distribution คือ $f\left(y;\mu\right) = \exp\left\{\frac{y\ln\mu - \mu}{1} - \ln y!\right\}$ ที่มี $\theta = \ln(\mu)$, $b(\theta) = \mu$, $a(\phi) = 1$ และ $c(y,\phi) = \ln y!$ ดังนั้น $E(Y) = \mu$, $V(\mu) = \mu$ และ Link function คือ log link ที่ $g(\mu) = \ln\mu$ ซึ่งเป็นที่มาของตัวแบบการถดถอยปัวซง หรือ Log-linear model

- ถ้า Y มีการแจกแจงทวินาม ที่เกิดจากการกระทำซ้ำ ๆ กัน m ครั้ง และ π แทนความน่าจะเป็นของการเกิดความสำเร็จในการกระทำแต่ละครั้ง เขียนแทนด้วย $Y \sim Bin(m,\pi)$ ฟังก์ชันมวลน่าจะเป็นของ Y ในรูป Exponential family distribution คือ

$$f\left(y;m,\pi\right) = \exp\left\{y\ln\left(\frac{\pi}{1-\pi}\right) + m\ln\left(1-\pi\right) + \ln\binom{m}{y}\right\} \quad \text{ที่มี}$$

$\theta = \ln\left(\frac{\pi}{1-\pi}\right)$, $b(\theta) = m\ln(1-\pi)$, $a(\phi) = 1$ และ $c(y,\phi) = \ln\binom{m}{y}$ ดังนั้น $E(Y) = m\pi$, $V(\pi) = \pi(1-\pi)$ และ Link function คือ logit link ที่ $g(\pi) = \ln\left(\frac{\pi}{1-\pi}\right)$ ซึ่งเป็นที่มาของตัวแบบการถดถอยทวินาม หรือ Logistic regression model

ถ้าประมาณค่าพารามิเตอร์ $\beta_0,\beta_1,...,\beta_p$ ในสมการ (2.4) โดยวิธีความควรจะเป็นสูงสุดแล้วจะพิจารณาหาค่า β_j ที่ทำให้ Log – likelihood function (2.3) มีค่าสูงสุด โดยพิจารณาจากสมการที่ถูกกำหนดโดยให้อนุพันธ์อันดับหนึ่งของ 1 ใน (2.3) เทียบกับ β_j มีค่าเท่ากับศูนย์

$$\frac{\partial l}{\partial \beta_j} = \frac{\partial l}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} = 0 \;;\; j=0,1,...,p \qquad (2.5)$$

แต่สมการ (2.5) ไม่เป็นสมการเชิงเส้นของ β_j ดังนั้นการแก้สมการ (2.5) เพียงครั้งเดียวไม่สามารถหาค่าของ β_j ที่ทำให้ 1 มีค่าสูงสุด จึงต้องใช้วิธีการวนซ้ำ โดยวิธีการวนซ้ำที่ใช้ในวิทยานิพนธ์ฉบับนี้คือ วิธีนิวตัน – ราฟสัน (Newton – Raphson method) สมการสำหรับวิธีวนซ้ำนิวตัน – ราฟสันในการประมาณค่าความควรจะเป็นสูงสุดของ $\underset{\sim}{\beta}$ ในสมการ (2.4) คือ

$$\beta^{(m+1)} = \beta^{(m)} + \left[I^{(m)}\right]^{-1}\underset{\sim}{s}^{(m)} \qquad (2.6)$$

เมื่อ $\underset{\sim}{s}^{(m)} = \left[\begin{array}{c} \frac{\partial l}{\partial \beta_0} \\ \frac{\partial l}{\partial \beta_1} \\ \vdots \\ \frac{\partial l}{\partial \beta_p} \end{array}\right]_{\underset{\sim}{\beta}=\underset{\sim}{\beta}^{(m)}}$ และ $I^{(m)} = \left[\begin{array}{ccc} -\frac{\partial^2 l}{\partial \beta_0^2} & -\frac{\partial^2 l}{\partial \beta_0 \partial \beta_1} & \cdots & -\frac{\partial^2 l}{\partial \beta_0 \partial \beta_p} \\ -\frac{\partial^2 l}{\partial \beta_1 \partial \beta_0} & -\frac{\partial^2 l}{\partial \beta_1^2} & \cdots & -\frac{\partial^2 l}{\partial \beta_1 \partial \beta_p} \\ \vdots & \vdots & & \vdots \\ -\frac{\partial^2 l}{\partial \beta_p \partial \beta_0} & -\frac{\partial^2 l}{\partial \beta_p \partial \beta_1} & \cdots & -\frac{\partial^2 l}{\partial \beta_p^2} \end{array}\right]_{\underset{\sim}{\beta}=\beta^{(m)}}$

$\underset{\sim}{s}$ คือ Score vector และ I คือ Observed information matrix การแก้สมการ (2.6) จะถูกวนซ้ำไปเรื่อย ๆ จนกระทั่ง $\underset{\sim}{\beta}$ จะลู่เข้าสู่ค่าใดค่าหนึ่ง แล้วค่าของ $\underset{\sim}{\beta}$ ในกระบวนการสุดท้าย คือค่าประมาณพารามิเตอร์ความควรจะเป็นสูงสุด $\left(\underset{\sim}{\hat{\beta}}\right)$ ที่ต้องการ และ I^{-1} ในกระบวนการสุดท้ายคือ Asymptotic variance-covariance matrix ของ $\underset{\sim}{\hat{\beta}}$ (Nelder and Wedderburn, 1972)

2.3 ตัวแบบการถดถอยปัวซง (Poisson regression model)

การวิเคราะห์การถดถอยปัวซงเป็นการศึกษาความสัมพันธ์ของตัวแปรตาม Y_i ที่มีการแจกแจงปังซง และตัวแปรอิสระ $\underset{\sim}{x}_i = \left[1, x_{1i},...,x_{pi}\right]^T$ โดยความสัมพันธ์ถูกเชื่อมด้วย Log link function ผ่านค่าเฉลี่ย μ_i หรือ Log – linear model หรือ ตัวแบบการถดถอยปัวซง

$$\eta_i = \ln \mu_i = \underset{\sim}{x}_i^T \underset{\sim}{\beta} \tag{2.7}$$

ที่ให้ $\mu_i = \exp(\underset{\sim}{x}_i^T \underset{\sim}{\beta})$ Log – likelihood function ของการแจกแจงปัวซงสำหรับค่า $y_1, y_2, \ldots, y_n$ คือ

$$l(\underset{\sim}{\mu}, \underset{\sim}{y}) \quad = \quad l(\underset{\sim}{\mu}) = \quad \sum_{i=1}^{n} \{ y_i \ln(\mu_i) - \mu_i - \ln(y_i !) \} \tag{2.8}$$

ข้อสมมติเบื้องต้นของตัวแปรสุ่มปัวซงคือ

1) จำนวนความสำเร็จที่เกิดขึ้นใน 1 หน่วยเวลา หรือในขอบเขตหนึ่งต้องเป็นอิสระกับจำนวนความสำเร็จที่เกิดขึ้นใน 1 หน่วยเวลา หรือในขอบเขตอื่น ๆ

2) ค่าเฉลี่ยและความแปรปรวนต้องเท่ากัน

ข้อสมมติเบื้องต้นข้อที่ 2) มีความสำคัญที่ต้องตรวจสอบ เมื่อใช้ตัวแบบการถดถอยปัวซง การตรวจสอบเบื้องต้นอาจทำได้โดยพิจารณาจาก Deviance residual โดยที่ Deviance function สำหรับการถดถอยปัวซง คือ

$$\begin{aligned} D(\underset{\sim}{y}; \underset{\sim}{\hat{\mu}}) \quad &= \quad -2 \times \left\{ l(\underset{\sim}{\hat{\mu}}, \underset{\sim}{y}) - l(\underset{\sim}{y}, \underset{\sim}{y}) \right\} \\ &= \quad 2 \times \sum_{i=1}^{n} \left\{ y_i \ln\left(\frac{y_i}{\hat{\mu}_i} \right) - (y_i - \hat{\mu}_i) \right\} \end{aligned} \tag{2.9}$$

หรือพิจารณาจาก Pearson chi-square statistic

$$X^2 \quad = \quad \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} \tag{2.10}$$

ถ้าการถดถอยปัวซงเหมาะสมกับข้อมูลแล้ว $D(y; \hat{\mu})$ และ X^2 มีการแจกแจงไคกำลังสอง (Chi-square) ที่มี $df = n - p - 1$ ดังนั้นถ้าตัวแบบการถดถอยปัวซงที่มีพารามิเตอร์ $p+1$ ตัว เหมาะสมกับข้อมูล แสดงว่าค่าเฉลี่ยเท่ากับค่าความแปรปรวน แล้ว $D(y; \hat{\mu}) \approx n - p - 1$ หรือ $X^2 \approx n - p - 1$ ถ้า $D(y; \hat{\mu})$ หรือ X^2 มีค่ามากกว่า $n - p - 1$ แสดงว่าค่าความแปรปรวนมากกว่าค่าเฉลี่ย หรือน่าจะเกิด Overdispersion และข้อมูลที่มีลักษณะดังกล่าวจะถูกเรียกว่า Overdispersed Poisson counts ถ้า $D(y; \hat{\mu})$ หรือ X^2 มีค่าน้อยกว่า $n - p - 1$ แสดงว่าค่าความแปรปรวนน้อยกว่าค่าเฉลี่ย หรือน่าจะเกิด Underdispersion และข้อมูลที่มีลักษณะดังกล่าวจะถูกเรียกว่า Underdispersed Poisson counts แต่ในทางปฏิบัติ Underdispersed Poisson counts เกิดขึ้นน้อยมาก (McCullagh and Nelder, 1989)

2.4 Overdispersion

Overdispersion เป็นปรากฏการณ์ที่เกิดขึ้นเนื่องจากคาดหวังว่าข้อมูลแบบนับที่ได้เกิดขึ้นภายใต้การแจกแจงปัวซง แต่ในสภาพการณ์ที่เป็นจริงข้อมูลนั้นมีความแปรปรวนมากกว่าค่าเฉลี่ยซึ่งไม่เป็นไปตามข้อสมมติของการแจกแจงปัวซง Hinde and Demètrio (1998) ได้อธิบายถึงสาเหตุที่เป็นไปได้ที่ทำให้เกิด Overdispersion และผลของการที่ไม่นำ Overdispersion มาพิจารณาในการวิเคราะห์ข้อมูล ดังนี้

สาเหตุที่อาจทำให้เกิด Overdispersion

1) การจัดตัวอย่างแบบเกาะกลุ่ม (Cluster sampling) จะแบ่งประชากรออกเป็นกลุ่มย่อย ๆ โดยที่ในแต่ละกลุ่มย่อยจะมีคุณลักษณะคล้ายกัน และภายในกลุ่มเดียวกันจะมีคุณลักษณะที่แตกต่างกัน ถ้าการแบ่งกลุ่มไม่เหมาะสม อาจทำให้เกิดความผันแปรระหว่างกลุ่มซึ่งทำให้เกิดความแปรปรวนมาก

2) จำนวนนับที่เกิดขึ้นใน 1 หน่วยเวลา หรือในขอบเขตหนึ่งไม่เป็นอิสระกับจำนวนของเหตุการณ์หรือลักษณะที่สนใจที่เกิดขึ้นใน 1 หน่วยเวลาหรือในขอบเขตอื่น ทำให้เกิดความแปรปรวนระหว่างตัวแปรตาม

3) การรวมกลุ่มระหว่างระดับของตัวแปร ทำให้เกิดการแจกแจงแบบผสม (Compound distributions)

ผลของการที่ไม่นำ Overdispersion มาพิจารณาในการวิเคราะห์ข้อมูล

1) สัมประสิทธิ์ของการถดถอยปัวซงมีค่าความคลาดเคลื่อนมาตรฐาน (Standard error) ต่ำกว่าความเป็นจริง ทำให้ตัดสินใจเลือกตัวแปรอิสระเข้าตัวแบบไม่ถูกต้อง

2) ส่งผลให้การเปลี่ยนของค่า Deviance ที่เกี่ยวข้องกับการนำตัวแปรอิสระเข้าหรือออกจากตัวแบบมีค่ามากเกินไป อาจทำให้เลือกตัวแบบที่มีความซับซ้อนมากเกินไป

วิธีการวิเคราะห์ข้อมูลแบบนับที่เกิด Overdispersion มีหลายวิธีแต่ละวิธีจะมีที่มาแตกต่างกันขึ้นอยู่กับสาเหตุที่ทำให้เกิด Overdispersion Poortema (1999) ได้รวบรวมวิธีการวิเคราะห์ข้อมูลแบบนับที่เกิด Overdispersion ใน Binomial counts, Poisson counts และ Multinomial counts

2.5 การแจกแจงทวินามแบบลบ (Negative binomial model)

Lawless (1987) และ Hinde and Demètrio (1998) ได้แสดงให้เห็นว่า การแจกแจงทวินามแบบลบ สำหรับ Overdispersed Poisson count ถูกพัฒนามาจากแนวคิดของเบส์ (Bayes' concept) โดยสมมติให้ ตัวแปรสุ่ม Y_i มีการแจกแจงปัวซง ที่ $f(y_i;\theta_i) = f(y_i \mid \theta_i)$ เมื่อ θ_i เป็นค่าของตัวแปรสุ่มที่มีการแจกแจงแบบแกมมา มีพารามิเตอร์ ϕ และ λ_i และมีฟังก์ชันความหนาแน่นน่าจะเป็น $g(\theta_i;\phi,\lambda_i) = \frac{1}{\Gamma\left(\phi^{-1}\right)\lambda_i^{\phi^{-1}}}\theta_i^{\phi^{-1}-1}e^{\frac{-\theta_i}{\lambda_i}}$, $\theta_i, \lambda_i, \phi > 0$ ดังนั้น joint p.m.f. ของ Y_i และ θ_i คือ

$$f(y_i,\theta_i) = \frac{e^{-\theta_i}\theta_i^{y_i}}{y_i!}\cdot\frac{1}{\Gamma(\phi^{-1})\lambda_i^{\phi^{-1}}}\theta_i^{\phi^{-1}-1}e^{\frac{-\theta_i}{\lambda_i}}$$

และ marginal p.m.f ของ Y_i คือ

$$f(y_i;\phi,\lambda_i) = \frac{\lambda_i^{y_i}\Gamma(y_i+\phi^{-1})}{y_i!\Gamma(\phi^{-1})(1+\lambda_i)^{y_i+\phi^{-1}}}$$

ที่มี $E(Y_i) = E\left[E(Y_i \mid \theta_i)\right] = \frac{\lambda_i}{\phi} = \mu_i$ และ

$Var(Y_i) = E\left[Var(Y_i \mid \theta_i)\right] + Var\left[E(Y_i \mid \theta_i)\right] = \mu_i(1+\phi\mu_i)$ ดังนั้น Y_i มีการแจกแจงทวินามแบบลบ ที่มี p.m.f

$$f(y_i;\mu_i,\phi) = \frac{\Gamma(y_i+\phi^{-1})\mu_i^{y_i}\phi^{y_i}}{y_i!\Gamma(\phi^{-1})(1+\phi\mu_i)^{y_i+\phi^{-1}}} \quad ; \quad y_i = 0,1,2,\ldots\ ; \phi > 0 \qquad (2.11)$$

Lawless (1987) เรียก ϕ ว่า Overdispersion parameter ถ้า $\phi \to 0$ แล้วการแจกแจง NB ก็จะลู่เข้าสู่การแจกแจงปัวซง

Gurmu and Trivedi (1992) ได้แสดงให้เห็นว่า p.m.f ของการแจกแจง NB มีหลายรูปแบบตามความสัมพันธ์ระหว่างความแปรปรวนและค่าเฉลี่ย เขาได้เสนอรูปแบบทั่วไปของการแจกแจง NB โดยการให้ ϕ เป็นฟังก์ชันของ μ_i ตัวอย่างเช่น $\phi = \alpha\mu_i^{c-1}$ โดยที่ α เป็นค่าคงที่ c เป็นสเกลาร์ (scalar) ดังนี้

$$f(y_i;\mu_i,\alpha) = \begin{cases} \dfrac{\Gamma(y_i+\alpha^{-1}\mu_i^{1-c})}{y_i!\Gamma(\alpha^{-1}\mu_i^{1-c})}\dfrac{\mu_i^{cy_i}\alpha^{y_i}}{(1+\alpha\mu_i^{c})^{y_i+\alpha^{-1}\mu_i^{1-c}}} & , \quad y_i = 0,1,2,\ldots\ ;\alpha>0 \\ 0 & , \quad \text{otherwise} \end{cases} \tag{2.12}$$

โดยที่ $E(Y_i)=\mu_i$ และ $Var(Y_i)=\mu_i(1+\alpha\mu_i^{c})$ ถ้า $\alpha\to 0$ การแจกแจง NB จะลดรูปเป็นการแจกแจงปัวซง c เป็นดัชนีที่แสดงความสัมพันธ์ระหว่างความแปรปรวน และค่าเฉลี่ยของการแจกแจงทวินามแบบลบ การแจกแจง NB ที่นิยมใช้กันอย่างแพร่หลายมี 2 การแจกแจงคือ การแจงแจงทวินามแบบลบที่มีความแปรปรวนเป็นฟังก์ชันเชิงเส้นของค่าเฉลี่ย (Linear mean – variance Negative Binomial) หรือเรียกว่า การแจกแจง NB1 ซึ่งได้จากการกำหนดให้ $c=0$ และการแจงแจงทวินามแบบลบที่มีความแปรปรวนเป็นฟังก์ชันกำลังสองของค่าเฉลี่ย (Quadratic mean – variance Negative Binomial) หรือเรียกว่าการแจกแจง NB2 ซึ่งได้จากการกำหนดให้ $c=1$

1. การแจกแจง NB1

เมื่อ $c=0$ ตัวแปรสุ่ม Y_i มีการแจกแจงทวินามแบบลบที่มีค่าเฉลี่ย μ_i และ Overdispersion α ซึ่งมักเขียนย่อ ๆ เป็น $Y_i \sim NB1(\mu_i,\alpha)$ แล้ว p.m.f ของ Y_i คือ

$$f(y_i;\mu_i,\alpha) = \begin{cases} \dfrac{\Gamma(y_i+\alpha^{-1}\mu_i)}{y_i!\Gamma(\alpha^{-1}\mu_i)}\dfrac{\alpha^{y_i}}{(1+\alpha)^{y_i+\alpha^{-1}\mu_i}} & , \quad y_i = 0,1,2,\ldots\ ;\alpha>0 \\ 0 & , \quad \text{otherwise} \end{cases} \tag{2.13}$$

ที่มี $E(Y_i)=\mu_i$ และ $Var(Y_i)=\mu_i(1+\alpha)$

2. การแจกแจง NB2

เมื่อ $c=1$ ตัวแปรสุ่ม Y_i มีการแจกแจงทวินามแบบลบที่มีค่าเฉลี่ย μ_i และ Overdispersion α ซึ่งมักเขียนย่อ ๆ เป็น $Y_i \sim NB2(\mu_i,\alpha)$ แล้ว p.m.f ของ Y_i คือ

$$f(y_i;\mu_i,\alpha) = \begin{cases} \dfrac{\Gamma(y_i+\alpha^{-1})}{y_i!\Gamma(\alpha^{-1})}\dfrac{\mu_i^{y_i}\alpha^{y_i}}{(1+\alpha\mu_i)^{y_i+\alpha^{-1}}} & , \quad y_i = 0,1,2,\ldots\ ;\alpha>0 \\ 0 & , \quad \text{otherwise} \end{cases} \tag{2.14}$$

ที่มี $E(Y_i)=\mu_i$ และ $Var(Y_i)=\mu_i(1+\alpha\mu_i)$

ทั้ง NB1 และ NB2 ตัวแบบสำหรับค่าเฉลี่ย คือ $\ln(\mu_i)=\underset{\sim}{x}_i^{T}\underset{\sim}{\beta}$

2.6 การเลือกตัวแบบ (Model Selection)

การเลือกตัวแบบจำแนกออกเป็น 2 กรณี คือ การเลือกระหว่างตัวแบบถดถอยปัวซง และตัวแบบถดถอย NB กับการเลือกตัวแบบที่เหมาะสมที่สุดโดยพิจารณาตัวแปรอิสระเข้าในตัวแบบหรือออกจากตัวแบบ

2.6.1 การเลือกระหว่างตัวแบบถดถอยปัวซงและตัวแบบถดถอย NB

สถิติที่ใช้ทดสอบความเหมาะสมระหว่างตัวแบบถดถอยปัวซง และตัวแบบ NB ซึ่งจะทดสอบผ่านค่า Overdispersion Parameter (α) โดยสมมติฐานของการทดสอบคือ

$$\begin{aligned} H_0 &: \alpha = 0 \\ H_1 &: \alpha > 0 \end{aligned} \tag{2.15}$$

สถิติที่ใช้ทดสอบที่นิยมใช้มี 3 แบบทดสอบ ดังนี้

1) Likelihood ratio test

Likelihood ratio test ที่ใช้ทดสอบ (2.15) นิยามด้วย

$$LR = -2\{l(\hat{\mu}) - l_{NB2}(\hat{\mu}, \hat{\alpha})\}$$

เมื่อ $l(\hat{\mu})$ คือ Log – likelihood function ของตัวแบบการถดถอยปัวซง และ $l_{NB2}(\hat{\mu}, \hat{\alpha})$ คือ Log – likelihood function ของตัวแบบการถดถอย NB2 ภายใต้ข้อสมมติว่า H_0 จริง LR มีการแจกแจงไคกำลังสองที่มี Degree of freedom เท่ากับ 1 (Lawless, 1987)

2) Wald test

Wald test ที่ใช้ทดสอบ (2.15) นิยามด้วย $W = \dfrac{\hat{\alpha}^2}{Var(\hat{\alpha})}$

ภายใต้ข้อสมมติว่า H_0 จริง W มีการแจกแจงไคกำลังสองที่มี Degree of freedom เท่ากับ 1 (Lawless, 1987)

3) Score tests

Score test ที่ใช้ทดสอบ (2.15) คืออนุพันธ์อันดับที่ 1 ของ l_{NB2} เทียบกับ α เมื่อ $\alpha = 0$ $\left(\left. \dfrac{\partial l_{NB2}}{\partial \alpha} \right|_{\alpha=0} \right)$ ซึ่งก็คือการสมมติให้ H_0 จริง ดังนั้น Score test จึงเป็นแบบทดสอบที่มีข้อดี

กว่า Likelihood ratio test และ Wald test ตรงที่ค่าสถิติที่คำนวณจากค่า $\hat{\underset{\sim}{\beta}}$ หรือ $\hat{\underset{\sim}{\mu}}$ จากการถดถอยปัวซงเท่านั้น โดยไม่จำเป็นต้อง fit ตัวแบบ NB2

Dean (1992) ได้เสนอ Score test ของตัวแบบ NB2 สำหรับการทดสอบสมมติฐาน (2.15) ดังนี้

Score test statistic $$P_B = \frac{\sum_{i=1}^{n}\{(Y_i - \hat{\mu}_i)^2 - Y_i\}}{\{2\sum_{i=1}^{n}\hat{\mu}_i^2\}^{\frac{1}{2}}}$$

Adjusted statistic $$P'_B = \frac{\sum_{i=1}^{n}\{(Y_i - \hat{\mu}_i)^2 - Y_i + \hat{h}_{ii}\hat{\mu}_i\}}{\{2\sum_{i=1}^{n}\hat{\mu}_i^2\}^{\frac{1}{2}}}$$

โดยที่ $\hat{h}_{ii}$ คือ Leverage ซึ่งเป็นสมาชิกบนแนวทแยงมุมหลักของ $\hat{W}^{\frac{1}{2}}X(X^T\hat{W}X)^{-1}X^T\hat{W}^{\frac{1}{2}}$ ของการถดถอยปัวซง และ $\hat{W}$ คือ Weight matrix ที่มี $\hat{\mu}_i$ เป็นสมาชิกบนเส้นทแยงมุมหลัก
ค่า Adjusted statistic ใน Score test ของ Dean (1992) มีข้อดีดังนี้

1) ช่วยปรับความเอนเอียงของค่าประมาณของค่าเฉลี่ย P_B
2) เข้าสู่การแจกแจงปรกติที่มีค่าเฉลี่ย 0 ความแปรปรวน 1 ได้เร็วกว่า P_B

ต่อมา Wang-Shu Lu (1997) ได้เสนอ Score test ของตัวแบบ NB2 สำหรับการทดสอบสมมติฐาน (2.15) ดังนี้

Score test statistic $$S_2 = \frac{\sum_{i=1}^{n}\left\{(Y_i - \hat{\mu}_i)^2 - Y_i\right\}}{\{2\sum_{i=1}^{n}\hat{\mu}_i^2\}^{\frac{1}{2}}}$$

Adjusted statistic $$S'_2 = \frac{\sum_{i=1}^{n}\{(Y_i - \hat{\mu}_i)^2 - cY_i\}}{\{2\sum_{i=1}^{n}\hat{\mu}_i^2\}^{\frac{1}{2}}}$$

โดยที่ $c = \frac{(n-p)}{n}$, n คือ ขนาดของตัวอย่าง และ p คือจำนวนตัวแปรอิสระในตัวแบบการถดถอยปัวซง ค่า Adjusted statistic ใน Score test ของ Wang-Shu Lu (1997) มีข้อดีดังนี้

1) ช่วยปรับความเอนเอียงของค่าประมาณของค่าเฉลี่ย S_2
2) เข้าสู่การแจกแจงปรกติที่มีค่าเฉลี่ย 0 ความแปรปรวน 1 ได้เร็วกว่า S_2

ค่า Adjusted statistic S_2' ของ Wang-Shu Lu (1997) จะคล้ายคลึงกับค่า Adjusted statistic P_B' ของ Dean (1992) แต่ P_B' จะขึ้นอยู่กับ Leverage ของการถดถอยปัวซง ส่วน S_2' จะขึ้นอยู่กับค่า c นอกจากนี้ Wang-Shu Lu (1997) ยังยืนยันว่า ถ้าข้อมูลมีการกระจายไม่มากแล้วค่า S_2' จะมีค่าใกล้เคียงกับ P_B'

2.6.2 การเลือกตัวแบบ

การพิจารณาตัวแบบที่เหมาะสม ตามแนวคิดเชิงสถิติคือเลือกตัวแบบที่มีจำนวนตัวแปรอิสระน้อยที่สุด แต่สามารถอธิบายการแปรผันของตัวแปรตามได้ดีพอ ๆ กับตัวแบบที่มีจำนวนตัวแปรอิสระมาก ๆ กฎเกณฑ์ที่นิยมใช้ในการเลือกตัวแบบที่เหมาะสมที่สุดในการจำลองตัวแบบเชิงสถิติคือ Akaike information criterion (AIC) (Akaike, 1973) โดยที่กฎเกณฑ์นี้ได้ปรับสถิติ Likelihood ratio test ด้วยจำนวนตัวประมาณค่าพารามิเตอร์ในตัวแบบ ดังนี้

$$\text{AIC} = -2\times l + 2\,(\text{จำนวนตัวประมาณค่าพารามิเตอร์ในตัวแบบ}) \qquad (2.16)$$

ตัวแบบที่ทำให้ค่า AIC น้อยที่สุด จะเป็นตัวแบบที่เหมาะสมกว่าตัวแบบที่มีค่า AIC มาก

2.7 Robust standard error

Royall (1986) ได้กล่าวว่าการประมาณค่าพารามิเตอร์ด้วยวิธีความควรจะเป็นสูงสุดจะให้ตัวประมาณค่าที่มีคุณสมบัติ consistency และ asymptotic normal distribution โดยคุณสมบัตินี้จะ Robust ถึงแม้การแจกแจงของตัวแปรสุ่มจะไม่ถูกต้องสมบูรณ์ร้อยเปอร์เซ็นต์ แต่ตัวประมาณค่าคลาดเคลื่อนจะไม่ Robust นั่นคือถ้าใช้การแจกแจงไม่ถูกต้องจะทำให้ค่าประมาณของ Asymptotic standard error ของตัวประมาณค่าไม่ถูกต้องด้วย ดังนั้นเขาจึงเสนอตัวประมาณค่าของ standard error ที่ Robust และเรียกมันว่า Robust standard error ในกรณีที่พารามิเตอร์ของการแจกแจงมีเพียงหนึ่งตัว คือ θ และถ้าให้ $\hat{\theta}$ เป็นตัวประมาณค่าความควรจะเป็นสูงสุดของ θ แล้ว Robust variance ของ $\hat{\theta}$ คือ

$$\hat{\Lambda} = {n\sum\left\{\dot{l}_i(\hat{\theta})\right\}^2} \Big/ {\left\{I(\hat{\theta})\right\}^2}$$

เมื่อ $\dot{l}(\hat{\theta})$ คือ $\left.\frac{\partial l}{\partial \theta}\right|_{\theta=\hat{\theta}}$ และ $I(\hat{\theta})$ คือ Observed information matrix ในกรณีที่ $\hat{\underset{\sim}{\theta}}$ เป็นเวคเตอร์ของตัวประมาณค่าความควรจะเป็นสูงสุดของ $\underset{\sim}{\theta}$ และ Robust variance-covariance matrix ของ $\hat{\underset{\sim}{\theta}}$ คือ

$$\hat{\Lambda} = \{I(\hat{\theta})\}^{-1}\sum\{\dot{l}_i(\hat{\underset{\sim}{\theta}})\}\{\dot{l}_i(\hat{\underset{\sim}{\theta}})\}^T\{I(\hat{\theta})\}^{-1} \quad (2.17)$$

โดยที่ $\dot{l}$ คือเวคเตอร์ของอนุพันธ์อันดับหนึ่งของ Log-likelihood function เทียบกับพารามิเตอร์ θ และแทนค่า θ ด้วย $\hat{\theta}$ และ I คือเมตริกซ์ของอนุพันธ์อันดับสองของ Log-likelihood function เทียบกับพารามิเตอร์ θ Robust standard error ของ $\hat{\underset{\sim}{\theta}}$ คือรากที่สองของสมาชิกบนเส้นทแยงมุมหลักของ $\hat{\Lambda}$ ถ้าการแจกแจงที่ใช้ในการวิเคราะห์ข้อมูลถูกต้องแล้ว จะให้ Robust standard error มีค่าเท่ากับ Asymptotic standard error ที่ได้จากกระบวนการย้อนซ้ำ (2.6) ด้วยคุณสมบัติที่ดีของ Robust standard error ทำให้เราสนใจที่จะใช้มันเป็นกฎเกณฑ์หนึ่งในการตรวจสอบการใช้ mean–variance NB กับ Overdispersed Poisson counts

2.8 การวินิจฉัยตัวแบบ (Model Diagnostics)

การวินิจฉัยตัวแบบโดยใช้ half normal plot with simulated envelope สำหรับ normal regression model ซึ่งเสนอโดย Atkinson (1985) เป็นการตรวจสอบการแจกแจงของเศษตกค้าง (Residual) โดยการนำค่าสัมบูรณ์ของ Pearson residual จากตัวแบบที่ได้ มาวาดกราฟกับ Half-normal scores $\phi^{-1}\left(\frac{i+n}{n}\right)$ โดยที่ ϕ คือฟังก์ชันการแจกแจงสะสมของการแจกแจงปรกติมาตรฐาน ต่อมา Demètrio and Hinde (1997) ได้ปรับเป็น $\phi^{-1}\left(\frac{i+n-0.125}{2n+0.5}\right)$ ซึ่งให้ค่าที่ใกล้เคียงกับค่าคาดหวังของสถิติลำดับมากขึ้น ถ้าค่าของ Pearson residual ทุกค่าอยู่ใน envelope แสดงว่า Pearson residual นั้นมีการแจกแจงปรกติ ต่อมา Williams (1987) ได้เสนอการตรวจสอบความถูกต้องของตัวแบบเชิงสถิติโดยใช้ Half normal plot with simulate envelope โดยใช้ค่าสัมบูรณ์ของ Standardized deviance residual ของตัวแบบ และเขายังกล่าวว่า Half normal plot with simulated envelope สามารถใช้ตรวจสอบได้ทั้งข้อสมมติเบื้องต้นของตัวแบบ และ Overdispersion (ทั้งค่าน้อยและค่ามากของ Standardized deviance residual ของตัวแบบ อยู่นอก upper envelope) ได้ด้วย ตรวจสอบความถูกต้องของตัวแบบโดย Half normal plot with simulated envelope มีขั้นตอนการสร้างดังนี้

1) Fit ตัวแบบ และคำนวณค่าสัมบูรณ์ของ Standardized Pearson residual แทนด้วย $r^{\bullet}$ เรียงลำดับจากน้อยไปมาก สมมติแทนด้วย $d_{(i)}$

2) จำลองตัวแปรตาม 19 ชุด ภายใต้ข้อสมมติว่าตัวแบบในข้อ 1) เหมาะสม

3) Fit ตัวแบบโดยใช้ตัวแปรตามจากข้อ 2) ใช้ตัวแปรอิสระชุดเดียวกับตัวแบบที่ Fit ในข้อ 1) คำนวณค่าสัมบูรณ์ของ $r^{\bullet}$ เรียงลำดับจากน้อยไปมากแล้วแทนด้วย $d^{*}_{k(i)}$ โดยที่ $k = 1, 2, ..., 19$; $i = 1, 2, ..., n$

4) คำนวณค่าต่ำสุด, ค่าสูงสุด และค่าเฉลี่ย ของ $d^{*}_{k(i)}$

5) พล็อตค่า $d_{(i)}$ ค่าต่ำสุด, ค่าสูงสุด และค่าเฉลี่ยของ $d^{*}_{k(i)}$ กับค่า Half–normal scores