บทที่ 3

การวิเคราะห์การถดถอยทวินามแบบลบที่มีความแปรปรวนเป็นฟังก์ชันกำลังสองของค่าเฉลี่ย

ในบทนี้เป็นการศึกษาเกี่ยวกับการถดถอยทวินามแบบลบที่มีความแปรปรวนเป็นฟังก์ชันกำลังสองของค่าเฉลี่ย หรือเรียก การถดถอย NB2 ตัวแบบการถดถอย NB2 เป็นอีกทางเลือกหนึ่งในการแก้ปัญหาการวิเคราะห์ Overdispersed Poisson counts โดยวิทยานิพนธ์ฉบับนี้จะกล่าวถึงวิธีการประมาณค่าพารามิเตอร์ของตัวแบบการถดถอย NB2 และ Overdispersion การเลือกตัวแบบที่ดีที่สุด การตรวจสอบ Overdispersion ในตัวแบบการถดถอยปัวซง ศึกษาแบบทดสอบ Score test ที่ใช้เปรียบเทียบระหว่างการถดถอยแบบปัวซงกับการถดถอย NB2 เสนอ Robust standard error ของตัวประมาณค่าความควรจะเป็นสูงสุด พร้อมทำการจำลองข้อมูล (Simulation study) เพื่อศึกษาว่าสามารถใช้ Robust standard error เป็นเกณฑ์ในการตรวจสอบว่าตัวแบบการถดถอย NB2 เหมาะสมกับ Overdispersed Poisson counts หรือไม่ และใช้ในการวินิจฉัยตัวแบบ

3.1 การประมาณค่าความควรจะเป็นสูงสุด และการเลือกตัวแบบที่ดีที่สุด สำหรับตัวแบบ NB2

จาก (2.14) เราได้ว่า p.m.f ของตัวแปรสุ่ม $Y_i \sim NB2(\mu_i, \alpha)$ คือ

$$f(y_i;\mu_i,\alpha) = \begin{cases} \dfrac{\Gamma(y_i+\alpha^{-1})}{y_i!\Gamma(\alpha^{-1})}\dfrac{\mu_i^{y_i}\alpha^{y_i}}{(1+\alpha\mu_i)^{y_i+\alpha^{-1}}} & , \quad y_i = 0,1,2,\ldots\ ;\alpha>0 \\ 0 & , \quad \text{otherwise} \end{cases} \tag{3.1}$$

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

3.1.1 การประมาณค่าความควรจะเป็นสูงสุด

การประมาณค่าพารามิเตอร์โดยวิธีความควรจะเป็นสูงสุดของตัวแบบการถดถอย NB2 ที่มีตัวทำนายเชิงเส้น $\eta_i = \ln(\mu_i) = \underset{\sim}{x}_i^T\underset{\sim}{\beta} = \beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}$ จะเกี่ยวข้องกับการประมาณค่า $\underset{\sim}{\beta}$ และ Overdispersion parameter α ถ้าให้ตัวอย่างสุ่ม $Y_1, Y_2, \ldots, Y_n$ มีค่าเป็น $y_1, y_2, \ldots, y_n$ แล้ว Log-likelihood function ของ (3.1) คือ

$$l_{NB2} = \ln L(\underset{\sim}{\mu}, \alpha; \underset{\sim}{y}) = \ln \prod_{i=1}^{n} f(y_i; \mu_i, \alpha) = \sum_{i=1}^{n} \ln f(y_i; \mu_i, \alpha)$$

$$= \sum_{i=1}^{n} \left\{ y_i \ln\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right) - \frac{1}{\alpha}\ln(1+\alpha\mu_i) + \ln\Gamma(y_i + \alpha^{-1}) - \ln\Gamma(\alpha^{-1}) - \ln y_i ! \right\} \quad (3.2)$$

ค่าประมาณความควรจะเป็นสูงสุด $\underset{\sim}{\beta}$ และ α ซึ่งแทนด้วย $\hat{\underset{\sim}{\beta}}$ และ $\hat{\alpha}$ ตามลำดับ โดยปกติแล้วทำได้โดยการหาอนุพันธ์ย่อย (Partial derivative) ของ l_{NB2} เทียบกับ $\underset{\sim}{\beta}$ และ α และกำหนดให้เท่ากับ 0 แล้วทำการแก้สมการดังกล่าว แต่เนื่องจากสมการที่ได้ไม่เป็นสมการเชิงเส้น ดังนั้นการแก้สมการเพียงครั้งเดียวไม่สามารถทำให้ค่า $\hat{\underset{\sim}{\beta}}$ และ $\hat{\alpha}$ ที่ทำให้ค่า l_{NB2} มีค่าสูงสุดได้ จึงต้องใช้วิธีแก้สมการไม่เชิงเส้น (Nonlinear equation) แบบมีการวนซ้ำ เช่น วิธี นิวตัน – ราฟสัน (Newton– Raphson method) ซึ่งใช้ในวิทยานิพนธ์นี้

วิธีวนซ้ำนิวตัน – ราฟสัน (Newton – Raphson Iterative Method)

วิธีวนซ้ำนิวตัน – ราฟสัน เป็นวิธีการแก้สมการไม่เชิงเส้นแบบมีการวนซ้ำโดยการนำค่าประมาณที่ได้จากกระบวนการก่อนหน้ามาหาผลเฉลยในกระบวนการปัจจุบัน ดังนั้น ค่าประมาณของ $\underset{\sim}{\beta}$, α ในกระบวนการที่ $m+1$ ซึ่งแทนด้วย $\hat{\underset{\sim}{\beta}}^{(m+1)}$, $\hat{\alpha}^{(m+1)}$ จะได้จากการใช้ค่าประมาณของ $\underset{\sim}{\beta}$, α ในกระบวนการที่ m ซึ่งแทนด้วย $\hat{\underset{\sim}{\beta}}^{(m)}$, $\hat{\alpha}^{(m)}$ ในสมการนิวตัน – ราฟสัน ซึ่งมีเขียนได้ ดังนี้

$$\begin{bmatrix} \hat{\underset{\sim}{\beta}}^{(m+1)} \\ \hat{\alpha}^{(m+1)} \end{bmatrix} = \begin{bmatrix} \hat{\underset{\sim}{\beta}}^{(m)} \\ \hat{\alpha}^{(m)} \end{bmatrix} + \left[I^{(m)} \right]^{-1} \hat{\underset{\sim}{s}}^{(m)} \quad (3.3)$$

เมื่อ $\underset{\sim}{s} = \underset{\sim}{s}(\underset{\sim}{\beta}, \alpha)$ คือ Score vector ที่

$$\underset{\sim}{s}\left(\underset{\sim}{\beta}, \alpha\right) = \underset{\sim}{s} = \left[\frac{\partial l_{NB2}}{\partial \underset{\sim}{\beta}}, \frac{\partial l_{NB2}}{\partial \alpha} \right]^T = \left[\frac{\partial l_{NB2}}{\partial \beta_0} \; \frac{\partial l_{NB2}}{\partial \beta_1} \; \cdots \; \frac{\partial l_{NB2}}{\partial \beta_p} \; \frac{\partial l_{NB2}}{\partial \alpha} \right]^T_{(p+2)\times 1}$$

และ $I = I(\underset{\sim}{\beta}, \alpha)$ คือ Observed information matrix ขนาด $(p+2)\times(p+2)$ ที่ $I = \begin{bmatrix} I_{\underset{\sim}{\beta}\underset{\sim}{\beta}} & I_{\underset{\sim}{\beta}\alpha} \\ I_{\alpha\underset{\sim}{\beta}} & I_{\alpha\alpha} \end{bmatrix}$

เป็น Partitioned matrix ที่มี $I_{\underset{\sim}{\beta}\underset{\sim}{\beta}}$ เป็นเมตริกซ์สมมาตรที่มีมิติ $(p+1)\times(p+1)$ ที่มี $-\frac{\partial^2 l_{NB2}}{\partial \underset{\sim}{\beta} \partial \underset{\sim}{\beta}^T}$ เป็น

สมาชิก $I_{\alpha\alpha} = -\frac{\partial^2 l_{NB2}}{\partial \alpha^2}$ เป็นสเกลาร์ และ $I_{\underset{\sim}{\beta}\alpha} = I_{\alpha\underset{\sim}{\beta}}$ เป็นเมตริกซ์ที่มีมิติ $(p+1)\times 1$ ที่มี $-\frac{\partial^2 l_{NB2}}{\partial \underset{\sim}{\beta}\partial \alpha}$ เป็นสมาชิก ดังนั้นรูปเต็มของ I คือ

$$I = \begin{bmatrix} -\frac{\partial^2 l_{NB2}}{\partial \beta_0^2} & -\frac{\partial^2 l_{NB2}}{\partial \beta_0 \partial \beta_1} & \cdots & -\frac{\partial^2 l_{NB2}}{\partial \beta_0 \partial \beta_p} & -\frac{\partial^2 l_{NB2}}{\partial \beta_0 \partial \alpha} \\ -\frac{\partial^2 l_{NB2}}{\partial \beta_1 \partial \beta_0} & -\frac{\partial^2 l_{NB2}}{\partial \beta_1^2} & \cdots & -\frac{\partial^2 l_{NB2}}{\partial \beta_1 \partial \beta_p} & -\frac{\partial^2 l_{NB2}}{\partial \beta_1 \partial \alpha} \\ \vdots & \vdots & & \vdots & \vdots \\ -\frac{\partial^2 l_{NB2}}{\partial \beta_p \partial \beta_0} & -\frac{\partial^2 l_{NB2}}{\partial \beta_p \partial \beta_1} & \cdots & -\frac{\partial^2 l_{NB2}}{\partial \beta_p^2} & -\frac{\partial^2 l_{NB2}}{\partial \beta_p \partial \alpha} \\ -\frac{\partial^2 l_{NB2}}{\partial \alpha \partial \beta_0} & -\frac{\partial^2 l_{NB2}}{\partial \alpha \partial \beta_1} & \cdots & -\frac{\partial^2 l_{NB2}}{\partial \alpha \partial \beta_p} & -\frac{\partial^2 l_{NB2}}{\partial \alpha^2} \end{bmatrix}$$

แทนค่า $\underset{\sim}{\beta}$ ด้วย $\hat{\underset{\sim}{\beta}}^{(m)}$ และ α ด้วย $\hat{\alpha}^{(m)}$ ในสมการที่ (3.3) และทำกระบวนการวนซ้ำไปเรื่อย ๆ จนกว่าได้ค่าประมาณลู่เข้าสู่ค่าใดค่าหนึ่ง กล่าวคือจนกว่าค่าประมาณพารามิเตอร์ในครั้งที่ $m+1$ แตกต่างจากครั้งที่ m น้อยมากหรือเกือบเท่ากับศูนย์ โดยมีเกณฑ์การหยุดกระบวนการวนซ้ำ เช่น

$$|\alpha^{(m+1)} - \alpha^{(m)}| < \varepsilon \text{ และ } |l_{NB2}{}^{(m+1)} - l_{NB2}{}^{(m)}| < \varepsilon$$

ซึ่งในที่นี้ใช้ $|\alpha^{(m+1)} - \alpha^{(m)}| < \varepsilon$ ค่า $\hat{\underset{\sim}{\beta}}$ และ $\hat{\alpha}$ ที่ได้จากกระบวนการสุดท้าย คือค่าประมาณความควรจะเป็นสูงสุดของ $\underset{\sim}{\beta}$ และ α ที่ต้องการ

จุดอ่อนของวิธีวนซ้ำ นิวตัน – ราฟสัน คือถ้ากำหนดค่าเริ่มต้นไม่เหมาะสมแล้วจะทำให้การวนซ้ำไม่มีการลู่เข้า Hinde and Demètrio (1998) ได้เสนอค่าเริ่มต้นสำหรับการประมาณค่าของตัวแบบถดถอย NB2 สำหรับ $\underset{\sim}{\beta}$ และ α ซึ่งแทนด้วย $\hat{\underset{\sim}{\beta}}^{(0)}$ และ $\hat{\alpha}^{(0)}$ ดังนี้

1) ใช้ $\hat{\underset{\sim}{\beta}}^{(0)}$ จากการถดถอยปัวซง แล้วคำนวณ $\hat{\mu}_i^{(0)} = \exp(\underset{\sim}{x}_i^T \hat{\underset{\sim}{\beta}}^{(0)})$

2) ให้ $\hat{\alpha}^{(0)} = \dfrac{\sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} - (n-p-1)}{\sum_{i=1}^{n} \hat{\mu}_i (1 - q_{ii}\hat{\mu}_i)}$ ซึ่งเป็นการปรับ Pearson chi-square

โดยที่ q_{ii} เป็นสมาชิกบนเส้นทแยงมุมหลักของ Covariance matrix ของ Linear predictor $\text{Cov}(x_i^T \hat{\underset{\sim}{\beta}}) = X(X^T \hat{W} X)^{-1} X^T$ โดยที่ $\hat{W}$ คือ Weight matrix ที่มี $\hat{\mu}_i$ เป็นสมาชิกบนเส้นทแยงมุมหลัก

เมื่อแทนค่า $\mu_i = \exp(\underset{\sim}{x}_i^T \underset{\sim}{\beta})$ แล้ว Log-likelihood function (3.2) เขียนได้ดังนี้

$$l_{NB2} = \sum_{i=1}^{n} \Big\{ y_i \ln\alpha + y_i \underset{\sim}{x}_i^T \underset{\sim}{\beta} - y_i \ln(1+\alpha\exp(\underset{\sim}{x}_i^T \underset{\sim}{\beta})) - \alpha^{-1}\ln(1+\alpha\exp(\underset{\sim}{x}_i^T \underset{\sim}{\beta})) + \ln\Gamma(y_i+\alpha^{-1}) - \ln\Gamma(\alpha^{-1}) - \ln y_i ! \Big\}$$

และอนุพันธ์ย่อยอันดับที่หนึ่ง และสอง สำหรับ $\underset{\sim}{\beta}$ และ I มีดังนี้

$$\frac{\partial l_{NB2}}{\partial \beta_j} = \sum_{i=1}^{n} \left\{ y_i x_{ij} - \frac{y_i x_{ij}\alpha\exp(\underset{\sim}{x}_i^T \underset{\sim}{\beta})}{1+\alpha\exp(\underset{\sim}{x}_i^T \underset{\sim}{\beta})} - \frac{1}{\alpha}\cdot\frac{x_{ij}\alpha\exp(\underset{\sim}{x}_i^T \underset{\sim}{\beta})}{1+\alpha\exp(\underset{\sim}{x}_i^T \underset{\sim}{\beta})} \right\} \quad , \quad j = 0,1,2,...,p$$

$$= \sum_{i=1}^{n} \left\{ \frac{y_i x_{ij} + y_i x_{ij}\alpha\exp(\underset{\sim}{x}_i^T \underset{\sim}{\beta}) - y_i x_{ij}\alpha\exp(\underset{\sim}{x}_i^T \underset{\sim}{\beta}) - \alpha^{-1}\alpha x_{ij}\exp(\underset{\sim}{x}_i^T \underset{\sim}{\beta})}{1+\alpha\exp(\underset{\sim}{x}_i^T \underset{\sim}{\beta})} \right\}$$

$$= \sum_{i=1}^{n} \left\{ \frac{y_i x_{ij} - x_{ij}\exp(\underset{\sim}{x}_i^T \underset{\sim}{\beta})}{1+\alpha\exp(\underset{\sim}{x}_i^T \underset{\sim}{\beta})} \right\}$$

$$= \sum_{i=1}^{n} \left\{ \frac{y_i - \exp(\underset{\sim}{x}_i^T \underset{\sim}{\beta})}{1+\alpha\exp(\underset{\sim}{x}_i^T \underset{\sim}{\beta})} \right\} x_{ij}$$

$$= \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{1+\alpha\mu_i} \quad , \quad j = 0,1,2,...,p$$

$$\frac{\partial l_{NB2}}{\partial \alpha} = \sum_{i=1}^{n} \left\{ \frac{y_i}{\alpha} - \frac{y_i\exp(\underset{\sim}{x}_i^T \underset{\sim}{\beta})}{1+\alpha\exp(\underset{\sim}{x}_i^T \underset{\sim}{\beta})} - \left(\frac{\alpha^{-1}\exp(\underset{\sim}{x}_i^T \underset{\sim}{\beta})}{1+\alpha\exp(\underset{\sim}{x}_i^T \underset{\sim}{\beta})} - \frac{\ln(1+\alpha\exp(\underset{\sim}{x}_i^T \underset{\sim}{\beta}))}{\alpha^2} \right) + \frac{\partial\ln\Gamma(y_i+\alpha^{-1})}{\partial\alpha} - \frac{\partial\ln\Gamma(\alpha^{-1})}{\partial\alpha} \right\}$$

$$= \sum_{i=1}^{n} \left\{ \frac{y_i + y_i\alpha\exp(\underset{\sim}{x}_i^T \underset{\sim}{\beta}) - y_i\alpha\exp(\underset{\sim}{x}_i^T \underset{\sim}{\beta}) - \exp(\underset{\sim}{x}_i^T \underset{\sim}{\beta})}{\alpha\left(1+\alpha\exp(\underset{\sim}{x}_i^T \underset{\sim}{\beta})\right)} + \frac{\ln(1+\alpha\exp(\underset{\sim}{x}_i^T \underset{\sim}{\beta}))}{\alpha^2} + \frac{\partial\ln\Gamma(y_i+\alpha^{-1})}{\partial\alpha} - \frac{\partial\ln\Gamma(\alpha^{-1})}{\partial\alpha} \right\}$$

$$= \sum_{i=1}^{n} \left\{ \frac{y_i - \exp(\underset{\sim}{x}_i^T \underset{\sim}{\beta})}{\alpha\left(1+\alpha\exp(\underset{\sim}{x}_i^T \underset{\sim}{\beta})\right)} + \alpha^{-2}\ln(1+\alpha\exp(\underset{\sim}{x}_i^T \underset{\sim}{\beta})) + \mathrm{ddg}\left(y_i, \alpha^{-1}\right) \right\}$$

$$= -\alpha^{-2}\sum_{i=1}^{n}\left\{\frac{\exp(\underset{\sim}{x}_i^T\underset{\sim}{\beta})-y_i}{\alpha^{-1}+\exp(\underset{\sim}{x}_i^T\underset{\sim}{\beta})}-\ln(1+\alpha\exp(\underset{\sim}{x}_i^T\underset{\sim}{\beta}))-ddg\left(y_i,\alpha^{-1}\right)\right\}$$

$$= -\alpha^{-2}\sum_{i=1}^{n}\left\{\frac{\mu_i-y_i}{\alpha^{-1}+\mu_i}-\ln(1+\alpha\mu_i)-ddg\left(y_i,\alpha^{-1}\right)\right\}$$

โดยที่ $ddg\left(y_i,\alpha^{-1}\right) = \frac{\partial \ln\Gamma(y_i+\alpha^{-1})}{\partial\alpha}-\frac{\partial\ln\Gamma(\alpha^{-1})}{\partial\alpha}$ หรือ Digamma function

$$\frac{\partial^2 l_{NB2}}{\partial\beta_j\partial\beta_k} = \sum_{i=1}^{n}\left\{\frac{\partial}{\partial\beta_k}\left(\frac{y_i x_{ij}-x_{ij}\exp(\underset{\sim}{x}_i^T\underset{\sim}{\beta})}{1+\alpha\exp(\underset{\sim}{x}_i^T\underset{\sim}{\beta})}\right)\right\}$$

$$= \sum_{i=1}^{n}\left\{-\left(1+\alpha x_{ij}x_{ik}\left(\exp(\underset{\sim}{x}_i^T\underset{\sim}{\beta})\right)^2\right)-\frac{\left(y_i x_{ij}-x_{ij}\exp(\underset{\sim}{x}_i^T\underset{\sim}{\beta})\right)\alpha x_{ik}\exp(\underset{\sim}{x}_i^T\underset{\sim}{\beta})}{\left(1+\alpha\exp(\underset{\sim}{x}_i^T\underset{\sim}{\beta})\right)^2}\right\}$$

$$= \sum_{i=1}^{n}\left\{\frac{-x_{ij}x_{ik}\exp(\underset{\sim}{x}_i^T\underset{\sim}{\beta})-\alpha x_{ij}x_{ik}\left(\alpha\exp(\underset{\sim}{x}_i^T\underset{\sim}{\beta})\right)^2-\alpha y_i x_{ij}x_{ik}\exp(\underset{\sim}{x}_i^T\underset{\sim}{\beta})+\alpha x_{ik}x_{ij}\exp^2(\underset{\sim}{x}_i^T\underset{\sim}{\beta})}{\left(1+\alpha\exp(\underset{\sim}{x}_i^T\underset{\sim}{\beta})\right)^2}\right\}$$

$$= -\sum_{i=1}^{n}\left\{\frac{(1+\alpha y_i)x_{ij}x_{ik}\exp(\underset{\sim}{x}_i^T\underset{\sim}{\beta})}{\left(1+\alpha\exp(\underset{\sim}{x}_i^T\underset{\sim}{\beta})\right)^2}\right\}$$

$$= -\sum_{i=1}^{n}\left\{\frac{(1+\alpha y_i)}{(1+\alpha\mu_i)^2}\mu_i x_{ij}x_{ik}\right\}$$

$$\frac{\partial^2 l_{NB2}}{\partial\underset{\sim}{\beta}_j\partial\alpha} = \sum_{i=1}^{n}\frac{\partial}{\partial\alpha}\left(\frac{y_i x_{ij}-\exp\left(\underset{\sim}{x}_i^T\underset{\sim}{\beta}\right)x_{ij}}{1+\alpha\exp\left(\underset{\sim}{x}_i^T\beta\right)}\right)$$

$$= \sum_{i=1}^{n}\left\{\left(y_i x_{ij}-\exp\left(\underset{\sim}{x}_i^T\underset{\sim}{\beta}\right)x_{ij}\right)\left(-\frac{\exp\left(\underset{\sim}{x}_i^T\beta\right)}{\left(1+\alpha\exp\left(\underset{\sim}{x}_i^T\beta\right)\right)^2}\right)\right\}$$

$$= -\sum_{i=1}^{n}\left\{-\frac{\left(y_i-\exp\left(\underset{\sim}{x}_i^T\underset{\sim}{\beta}\right)\right)\exp\left(\underset{\sim}{x}_i^T\beta\right)x_{ij}}{\left(1+\alpha\exp\left(\underset{\sim}{x}_i^T\beta\right)\right)^2}\right\}$$

$$= -\sum_{i=1}^{n}\left\{\frac{\mu_i(y_i-\mu_i)}{(1+\alpha\mu)^2}x_{ij}\right\}$$

$$\frac{\partial^2 l_{NB2}}{\partial\alpha^2} = \sum_{i=1}^{n}\frac{\partial}{\partial\alpha}\left\{\frac{-\alpha^{-2}(\mu_i - y_i)}{\alpha^{-1}+\mu_i} + \alpha^{-2}\ln(1+\alpha\mu_i) + \alpha^{-2}\mathrm{ddg}\left(y_i,\alpha^{-1}\right)\right\}$$

$$= \sum_{i=1}^{n}\left\{-\frac{\partial}{\partial\alpha}\left((\mu_i - y_i)\frac{\alpha^{-2}}{\alpha^{-1}+\mu_i}\right) + \frac{\partial}{\partial\alpha}\left(\alpha^{-2}\ln\left(1+\alpha\mu_i\right)\right) - \frac{\partial}{\partial\alpha}\alpha^{-2}\mathrm{ddg}\left(y_i,\alpha^{-1}\right)\right\}$$

$$= \sum_{i=1}^{n}\left\{-(\mu_i - y_i)\left(\frac{\alpha^{-1}+\mu_i\left(-2\alpha^{-3}\right)+\alpha^{-4}}{\left(\alpha^{-1}+\mu_i\right)^2}\right) - 2\alpha^{-3}\ln\left(1+\alpha\mu_i\right) + \alpha^{-2}\frac{\mu_i}{1+\alpha\mu_i}\right.$$
$$\left. + \frac{\partial}{\partial\alpha}\alpha^{-2}\mathrm{ddg}\left(y_i,\alpha^{-1}\right)\right\}$$

$$= \sum_{i=1}^{n}\left\{\frac{-\left(-2\alpha^{-3}\right)(\mu_i - y_i)\left(\alpha^{-1}+\mu_i\right)}{\left(\alpha^{-1}+\mu_i\right)^2} - \frac{\alpha^{-4}(\mu_i - y_i)}{\left(\alpha^{-1}+\mu_i\right)^2} - 2\alpha^{-3}\ln\left(1+\alpha\mu_i\right) + \frac{\alpha^{-2}\mu_i}{1+\alpha\mu_i}\right.$$
$$\left. - \frac{\partial}{\partial\alpha}\alpha^{-2}\mathrm{ddg}\left(y_i,\alpha^{-1}\right)\right\}$$

$$= \sum_{i=1}^{n}\left\{\frac{2\alpha^{-3}(\mu_i - y_i)}{\alpha^{-1}+\mu_i} + \frac{\alpha^{-4}(y_i - \mu_i)}{\left(\alpha^{-1}+\mu_i\right)^2} - 2\alpha^{-3}\ln\left(1+\alpha\mu_i\right) + \frac{\alpha^{-4}\alpha\mu_i}{\alpha^{-1}+\mu_i}\right.$$
$$\left. - 2\alpha^{-3}\mathrm{ddg}\left(y_i,\alpha^{-1}\right) - \alpha^{-4}\mathrm{dtg}\left(y_i,\alpha^{-1}\right)\right\}$$

$$\therefore \quad \frac{\partial^2 l_{NB2}}{\partial\alpha^2} = \sum_{i=1}^{n}\left\{2\alpha^{-3}\left[\frac{(\mu_i - y_i)}{\alpha^{-1}+\mu_i} - \ln\left(1+\alpha\mu_i\right) - \mathrm{ddg}\left(y_i,\alpha^{-1}\right)\right]\right.$$
$$\left. + \alpha^{-4}\left[\frac{y_i - \mu_i}{\left(\alpha^{-1}+\mu_i\right)^2} + \frac{\alpha\mu_i}{\alpha^{-1}+\mu_i} - \mathrm{dtg}\left(y_i,\alpha^{-1}\right)\right]\right\}$$

โดยที่ $\mathrm{dtg}\left(y_i,\alpha^{-1}\right) = \frac{\partial^2 \ln\Gamma(y_i+\alpha^{-1})}{\partial\alpha^2} - \frac{\partial^2 \ln\Gamma(\alpha^{-1})}{\partial\alpha^2}$ หรือ Trigamma function

3.1.2 การเลือกตัวแบบที่เหมาะสมที่สุด

เนื่องจากตัวแบบที่มีพารามิเตอร์มากค่าประมาณที่ได้มักจะใกล้เคียงกับค่าจริง และเหมาะสมกับข้อมูลได้มากกว่าตัวแบบที่มีจำนวนพารามิเตอร์น้อยกว่า แต่การพิจารณาตัวแบบที่เหมาะสมตามแนวคิดเชิงสถิติคือ เลือกตัวแบบที่มีจำนวนตัวแปรอิสระน้อยที่สุดแต่สามารถอธิบายการแปรผันของตัวแปรตามได้ดีพอ ๆ กับตัวแบบที่มีจำนวนตัวแปรอิสระมาก ๆ กฎเกณฑ์ที่นิยมใช้

ในการเลือกตัวแบบที่เหมาะสมที่สุดในการจำลองตัวแบบเชิงสถิติ คือ Akaike information criterion (AIC) (Akaike, 1973) โดยที่กฎเกณฑ์นี้ได้ปรับสถิติ Likelihood ($-2\times l$) ด้วยจำนวนตัวประมาณค่าพารามิเตอร์ในตัวแบบ สำหรับตัวแบบ NB2 ที่มีตัวแปรอิสระ p ตัว

$$\text{AIC} = -2\times l_{NB2} + 2(\text{จำนวนตัวประมาณค่าพารามิเตอร์ในตัวแบบ}) \qquad (3.4)$$

ตัวแบบที่ทำให้ค่า AIC น้อยที่สุดจะเป็นตัวแบบที่ถูกเลือก

3.2 การตรวจสอบ Overdispersion ใน Poisson counts

ถึงแม้ว่าการใช้อัตราส่วนระหว่าง Residual deviance และ df ของ Residual deviance เป็นเกณฑ์ในการพิจารณาว่าเกิด Overdispersion ใน Poisson counts หรือไม่ แต่เกณฑ์ดังกล่าวเป็นแค่เกณฑ์พื้นฐานซึ่งมีโอกาสที่จะเกิดความผิดพลาด ดังนั้นจึงได้มีการเสนอสถิติที่ใช้ทดสอบเพื่อเลือกระหว่างตัวแบบการถดถอยปัวซง และตัวแบบ NB2 โดยพิจารณาจากค่า Overdispersion Parameter (α) ขึ้น โดยสมมติฐานของการทดสอบ คือ

$$\begin{aligned} H_0 &: \ \alpha = 0 \\ H_1 &: \ \alpha > 0 \end{aligned} \qquad (3.5)$$

สถิติที่ใช้ทดสอบที่นิยมใช้มี 3 ชนิด คือ

1) Likelihood ratio test

Likelihood ratio test ที่ใช้ทดสอบ (3.5) คือ

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

เมื่อ $l(\hat{\mu})$ คือ Log – likelihood function ขอตัวแบบการถดถอยปัวซง 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 (3.5) คือ $\text{W} = \dfrac{\hat{\alpha}^2}{\text{Var}(\hat{\alpha})}$

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

3) Score tests

Score test ที่ใช้ทดสอบ (3.5) ได้จากการหาอนุพันธ์อันดับที่ 1 ของ l_{NB2} เทียบกับ α แล้วแทนค่า $\alpha = 0$ $\left(\frac{\partial l_{NB2}}{\partial \alpha}\Big|_{\alpha=0} \right)$ ซึ่งก็คือการสมมติให้ H_0 จริง ดังนั้น Score test จึงเป็นแบบทดสอบที่มีข้อดีกว่า Likelihood ratio test และ Wald test ตรงที่ค่าสถิติที่ใช้ทดสอบคำนวณจากค่า $\hat{\underset{\sim}{\beta}}$ หรือ $\hat{\underset{\sim}{\mu}}$ จากการถดถอยปัวซงเท่านั้น โดยไม่จำเป็นต้อง fit ตัวแบบ NB2

Dean (1992) ได้เสนอ Score test ของตัวแบบ NB2 สำหรับการทดสอบสมมติฐาน (3.5) 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 โดยที่ $\hat{W}$ คือ Weight matrix ที่มี $\hat{\mu}_i$ เป็นสมาชิกบนเส้นทแยงมุมหลัก P_B และ $P_B^{'}$ มีการแจกแจงปรกติมาตรฐาน Dean (1992) ยืนยันว่า $P_B^{'}$ ข้อดีดังนี้

1) ช่วยปรับความเอนเอียงของค่าประมาณของค่าเฉลี่ย P_B ซึ่งมีค่าเท่ากับ 0

2) เข้าสู่การแจกแจงปรกติที่มีค่าเฉลี่ย 0 ความแปรปรวน 1 ได้เร็วกว่า P_B

ต่อมา Wang – Shu Lu (1997) ได้เสนอ Score test ของตัวแบบ NB2 สำหรับการทดสอบสมมติฐาน (3.5) 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, p คือจำนวนตัวแปรอิสระในตัวแบบการถดถอยปัวซง ทำนองเดียวกัน S_2 และ $S_2^{'}$ มีการแจกแจงปรกติมาตรฐาน และ Wang – Shu Lu (1997) ยืนยันว่า $S_2^{'}$ ข้อดี ดังนี้

1) ช่วยปรับความเอนเอียงของค่าประมาณของค่าเฉลี่ย S_2 ซึ่งมีค่าเท่ากับ 0

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 Wang – Shu Lu (1997) อ้างว่า ถ้าข้อมูลมีการกระจายไม่มากแล้ว ค่า $S_2^{'}$ จะมีค่าใกล้เคียงกับ P_B'

3.3 Simulation study สำหรับ Score test

เพื่อศึกษาคุณสมบัติของแบบทดสอบ Score tests ของ Dean (1992) และของ Wang – Shu Lu (1997) เราดำเนินการโดยใช้ Simulation study ซึ่งมีขั้นตอน ดังนี้

1. ศึกษาการแจกแจงของ Score test ที่เสนอโดย Dean (1992) และโดย Wang – Shu Lu (1997) โดยพิจารณาจากค่าประมาณของความน่าจะเป็นของการเกิด Type I error หรือการเกิดความผิดพลาดอันเนื่องมาจากปฏิเสธ $H_0 : \alpha = 0$ เมื่อ H_0 จริง โดยการจำลองข้อมูลภายใต้การแจกแจงปัวซง โดยมีตัวแบบที่ศึกษา (Working models) และขนาดตัวอย่าง ดังแสดงในตาราง 3.1 และ 3.2 จำนวน 5,000 ชุด แต่ละชุด fit ตัวแบบปัวซง แล้วคำนวณค่าความน่าจะเป็นของการเกิด Type I error เพื่อศึกษา Score test ของ Dean (1992) คือ $P_D = \frac{\sum_{R=1}^{5,000} [s_D \geq Z_\delta]}{R}$ และของ Wang-Shu Lu (1997) คือ $P_W = \frac{\sum_{R=1}^{5,000} [s_w \geq Z_\delta]}{R}$ ซึ่งค่า P_D จะถูกจำแนกเป็น P_{D1} และ P_{D2} สำหรับ P_B และ P_B' ตามลำดับ และ P_W จะถูกจำแนกเป็น P_{W1} และ P_{W2} สำหรับ S_2 และ S_2' เพื่อเปรียบเทียบกับค่าระดับนัยสำคัญที่ศึกษา $\delta =$ 0.10, 0.05, 0.01 ดังแสดงในตาราง 3.3

ตารางที่ 3.1 Working models สำหรับการจำลองข้อมูลเพื่อศึกษา Score test โดยที่ค่าเฉลี่ยมีการกระจายน้อย

ตัวแบบ	ค่าต่ำสุดของ μ	ค่าสูงสุดของ μ	ค่าเฉลี่ยของ μ
$\ln(\mu) = 1.05$	-	-	2.858
$\ln(\mu) = 1.05 - 0.45x_1$	1.162	7.029	3.470
$\ln(\mu) = 1.05 - 0.45x_1 + 0.25x_2$	1.492	9.025	4.039
$\ln(\mu) = 0.65 - 0.55x_1 + 0.50x_2 + 0.15x_1x_2$	1.419	7.029	3.160

x_1 คือตัวแปรเชิงปริมาณ มีค่า -2, -1, 0, 1, 2 และ x_2 คือตัวแปรเชิงคุณภาพ มีค่า 1, 0, 0, 1, 1 โดยแต่ละค่าของ x_1 และ x_2 มีจำนวนซ้ำ r = 5, 10, 15, 20 สำหรับ n = 25, 50, 75, 100 ตามลำดับ

ตารางที่ 3.2 Working models สำหรับการจำลองข้อมูลเพื่อศึกษา Score test โดยที่ค่าเฉลี่ยมีการกระจายมาก

ตัวแบบ	ค่าต่ำสุดของ μ	ค่าสูงสุดของ μ	ค่าเฉลี่ยของ μ
$\ln(\mu) = 2.5 + 0.5x_1$	4.482	33.115	15.451
$\ln(\mu) = 2.5 - 0.15x_1 + 0.25x_2$	0.779	314.190	77.048
$\ln(\mu) = 3.25 - 0.65x_1 + 0.75x_2 + 0.25x_1x_2$	24.533	121.510	51.567

x_1 คือตัวแปรเชิงปริมาณ มีค่า -2, -1, 0, 1, 2 และ x_2 คือตัวแปรเชิงคุณภาพ มีค่า 1, 0, 0, 1, 1 โดยแต่ละค่าของ x_1 และ x_2 มีจำนวนซ้ำ r = 5, 10, 15, 20 สำหรับ n = 25, 50, 75, 100 ตามลำดับ

จากตารางที่ 3.3 เมื่อพิจารณาระหว่างค่า Score test statistic และค่า Adjusted statistic พบว่าการประมาณค่าความน่าจะเป็นของการเกิด Type I error ที่ขนาดตัวอย่างเท่ากัน Adjusted statistic จะมีค่าใกล้เคียงกับระดับนัยสำคัญ (δ) มากกว่าค่า Score test statistic ทั้งของ Dean (1992) และ Wang – Shu Lu (1997) นั่นคือ ค่า Adjusted statistic จะลู่เข้าสู่การแจกแจงปรกติได้เร็วกว่าค่า Score test statistic เมื่อพิจารณาขนาดตัวอย่างที่แตกต่างกันพบว่า ถ้าขนาดตัวอย่างมาก ค่า Adjusted statistic และค่า Score test statistic จะมีค่าใกล้เคียงกับค่าระดับนัยสำคัญ (δ) มากกว่า

ขนาดตัวอย่างน้อย และเมื่อพิจารณาระหว่างตัวแบบที่ค่าเฉลี่ยมีการกระจายน้อย และตัวแบบที่ค่าเฉลี่ยมีการกระจายมาก พบว่าถ้าตัวแบบที่ค่าเฉลี่ยมีการกระจายน้อยจะทำให้ค่า Adjusted statistic และค่า Score test statistic ของ Dean (1992) และ Wang – Shu Lu (1997) มีค่าใกล้เคียงกัน

ตารางที่ 3.3 ค่าประมาณความน่าจะเป็นของการเกิด Type I error เปรียบเทียบกับค่าระดับนัยสำคัญที่ศึกษา $\delta =$ 0.10, 0.05, 0.01

	Score test statistic $P_{D1}{}^{+}$			**Adjusted statistic** $P_{D2}{}^{+}$			**Score test statistic** $P_{W1}{}^{*}$			**Adjusted statistic** $P_{W2}{}^{*}$		
δ	**0.10**	**0.05**	**0.01**	**0.10**	**0.05**	**0.01**	**0.10**	**0.05**	**0.01**	**0.10**	**0.05**	**0.01**
$\ln(\mu) = 1.05$ ค่าเฉลี่ยคงที่												
n=25	0.076	0.045	0.016	0.093	0.056	0.018	0.082	0.051	0.013	0.099	0.064	0.016
n=50	0.089	0.047	0.013	0.103	0.056	0.017	0.088	0.049	0.017	0.100	0.057	0.020
n=75	0.090	0.049	0.013	0.100	0.058	0.016	0.094	0.051	0.017	0.106	0.061	0.019
n=100	0.088	0.049	0.014	0.100	0.057	0.015	0.096	0.054	0.013	0.106	0.060	0.016
$\ln(\mu) = 1.05 - 0.45x_1$ ค่าเฉลี่ยมีการกระจายน้อย												
n=25	0.057	0.030	0.009	0.087	0.049	0.016	0.061	0.031	0.012	0.087	0.050	0.015
n=50	0.074	0.044	0.010	0.100	0.058	0.014	0.080	0.043	0.017	0.101	0.054	0.021
n=75	0.077	0.044	0.014	0.102	0.059	0.017	0.078	0.044	0.014	0.098	0.056	0.019
n=100	0.084	0.049	0.012	0.104	0.061	0.016	0.084	0.039	0.015	0.102	0.051	0.018
$\ln(\mu) = 2.5 + 0.5x_1$ ค่าเฉลี่ยมีการกระจายมาก												
n=25	0.064	0.033	0.011	0.093	0.051	0.018	0.069	0.037	0.010	0.095	0.055	0.018
n=50	0.074	0.039	0.013	0.098	0.057	0.019	0.073	0.046	0.012	0.095	0.058	0.016
n=75	0.084	0.039	0.013	0.103	0.049	0.015	0.081	0.045	0.012	0.098	0.054	0.016
n=100	0.079	0.045	0.013	0.100	0.056	0.017	0.079	0.043	0.013	0.098	0.052	0.015

$^{+}$ Dean (1992), * Wang – Shu Lu (1997)

ตารางที่ 3.3 (ต่อ)

	Score test statistic P_{D1}^{+}			**Adjusted statistic** P_{D2}^{+}			**Score test statistic** P_{W1}^{*}			**Adjusted statistic** P_{W2}^{*}		
δ	**0.10**	**0.05**	**0.01**	**0.10**	**0.05**	**0.01**	**0.10**	**0.05**	**0.01**	**0.10**	**0.05**	**0.01**
$\ln(\mu) = 1.05 - 0.45x_1 + 0.25x_2$ ค่าเฉลี่ยมีการกระจายน้อย												
n=25	0.045	0.026	0.007	0.085	0.049	0.014	0.046	0.025	0.008	0.082	0.046	0.015
n=50	0.057	0.031	0.009	0.095	0.055	0.017	0.059	0.033	0.012	0.090	0.049	0.018
n=75	0.070	0.038	0.012	0.106	0.058	0.017	0.069	0.037	0.011	0.095	0.051	0.017
n=100	0.075	0.037	0.014	0.102	0.053	0.019	0.068	0.040	0.010	0.089	0.054	0.015
$\ln(\mu) = 2.5 - 0.15x_1 + 0.25x_2$ ค่าเฉลี่ยมีการกระจายมาก												
n=25	0.046	0.027	0.008	0.088	0.054	0.018	0.048	0.021	0.007	0.093	0.048	0.015
n=50	0.059	0.031	0.009	0.099	0.054	0.016	0.054	0.032	0.007	0.089	0.053	0.014
n=75	0.066	0.036	0.008	0.102	0.055	0.014	0.061	0.034	0.012	0.096	0.048	0.017
n=100	0.078	0.037	0.012	0.109	0.056	0.016	0.066	0.039	0.012	0.088	0.055	0.016
$\ln(\mu) = 0.65 - 0.55x_1 + 0.50x_2 + 0.15x_1x_2$ ค่าเฉลี่ยมีการกระจายน้อย												
n=25	0.038	0.025	0.006	0.091	0.055	0.015	0.037	0.022	0.007	0.077	0.042	0.014
n=50	0.053	0.031	0.010	0.093	0.054	0.016	0.054	0.034	0.008	0.090	0.054	0.017
n=75	0.068	0.033	0.011	0.102	0.056	0.019	0.063	0.035	0.010	0.091	0.054	0.015
n=100	0.063	0.038	0.012	0.094	0.059	0.017	0.063	0.035	0.008	0.096	0.057	0.015
$\ln(\mu) = 3.25 - 0.65x_1 + 0.75x_2 + 0.25x_1x_2$ ค่าเฉลี่ยมีการกระจายมาก												
n=25	0.042	0.023	0.008	0.092	0.052	0.019	0.039	0.024	0.007	0.076	0.045	0.016
n=50	0.060	0.027	0.011	0.101	0.054	0.020	0.062	0.031	0.010	0.097	0.053	0.019
n=75	0.071	0.034	0.010	0.108	0.059	0.019	0.065	0.031	0.010	0.094	0.049	0.018
n=100	0.069	0.032	0.008	0.103	0.055	0.015	0.067	0.037	0.010	0.097	0.054	0.016

$^{+}$ Dean (1992), * Wang – Shu Lu (1997)

2. ศึกษากำลังของการทดสอบ (Power of the test) ของ Score test ที่เสนอโดย Dean (1992) และโดย Wang – Shu Lu (1997) โดยพิจารณาจากค่าประมาณของค่าความน่าจะเป็นของการเกิด Type II error หรือการเกิดความผิดพลาดอันเนื่องมาจากยอมรับ $H_0 : \alpha = 0$ เมื่อ H_0 ไม่จริง โดยการจำลองข้อมูลภายใต้การแจกแจง NB2 จำนวน 5,000 ชุด ที่มี Working models สำหรับค่าเฉลี่ยดังแสดงในตาราง 3.1 และ 3.2 และกำหนดให้ Overdispersion parameter (α) มีค่าเท่ากับ 1.5 แต่ละชุด fit ตัวแบบปัวซง แล้วคำนวณค่าความน่าจะเป็นของการเกิด Type II error แทนด้วย $Q_D = \frac{\sum_{R=1}^{5,000}[s_D < Z_\delta]}{R}$ สำหรับ Score test ของ Dean (1992) และ $Q_W = \frac{\sum_{R=1}^{5,000}[s_w < Z_\delta]}{R}$ สำหรับ Score test ของ Wang – Shu Lu (1997) ซึ่งค่า Q_D จะถูกจำแนกเป็น Q_{D1} และ Q_{D2} สำหรับ P_B และ P'_B ตามลำดับ และค่า Q_W จะถูกจำแนกเป็น Q_{W1} และ Q_{W2} สำหรับ S_2 และ S'_2 เพื่อเปรียบเทียบกับค่าระดับนัยสำคัญที่ศึกษา $\delta =$ 0.10, 0.05, 0.01 แล้วหาค่า Power of the test $1-Q$ ดังแสดงในตาราง 3.4

ตารางที่ 3.4 Power of the test ของ Score test ที่ระดับนัยสำคัญที่ศึกษา $\delta =$ 0.10, 0.05, 0.01 กำหนด $\alpha = 1.5$

	Score test statistic $1-Q_{D1}{}^{+}$			**Adjusted statistic** $1-Q_{D2}{}^{+}$			**Score test statistic** $1-Q_{W1}{}^{*}$			**Adjusted statistic** $1-Q_{W2}{}^{*}$		
δ	**0.10**	**0.05**	**0.01**	**0.10**	**0.05**	**0.01**	**0.10**	**0.05**	**0.01**	**0.10**	**0.05**	**0.01**
$\ln(\mu_i) = 1.05$ ค่าเฉลี่ยคงที่												
n=25	0.661	0.569	0.402	0.692	0.603	0.431	0.657	0.553	0.418	0.691	0.586	0.447
n=50	0.919	0.870	0.752	0.929	0.886	0.769	0.918	0.873	0.746	0.928	0.889	0.763
n=75	0.963	0.939	0.855	0.966	0.945	0.867	0.965	0.938	0.855	0.970	0.947	0.866
n=100	0.988	0.977	0.933	0.990	0.979	0.941	0.988	0.977	0.937	0.989	0.980	0.943

ตารางที่ 3.4 (ต่อ)

	Score test statistic $1-Q_{D1}{}^{+}$			**Adjusted statistic** $1-Q_{D2}{}^{+}$			**Score test statistic** $1-Q_{W1}{}^{*}$			**Adjusted statistic** $1-Q_{W2}{}^{*}$		
δ	**0.10**	**0.05**	**0.01**	**0.10**	**0.05**	**0.01**	**0.10**	**0.05**	**0.01**	**0.10**	**0.05**	**0.01**
$\ln(\mu_i) = 1.05 - 0.45x_{1i}$ ค่าเฉลี่ยมีการกระจายน้อย												
n=25	0.792	0.720	0.589	0.838	0.770	0.644	0.804	0.720	0.591	0.843	0.766	0.639
n=50	0.942	0.908	0.824	0.955	0.926	0.850	0.937	0.895	0.822	0.951	0.917	0.849
n=75	0.977	0.963	0.914	0.981	0.971	0.926	0.973	0.960	0.915	0.979	0.967	0.926
n=100	0.993	0.987	0.959	0.995	0.991	0.966	0.992	0.984	0.961	0.994	0.986	0.966
$\ln(\mu_i) = 2.5 + 0.5x_{1i}$ ค่าเฉลี่ยมีการกระจายมาก												
n=25	0.997	0.998	0.992	0.998	0.998	0.995	0.998	0.997	0.993	0.998	0.998	0.995
n=50	1	1	1	1	1	1	1	1	1	1	1	1
n=75	1	1	1	1	1	1	1	1	1	1	1	1
n=100	1	1	1	1	1	1	1	1	1	1	1	1
$\ln(\mu) = 1.05 - 0.45x_1 + 0.25x_2$ ค่าเฉลี่ยมีการกระจายน้อย												
n=25	0.722	0.653	0.543	0.802	0.731	0.621	0.728	0.662	0.534	0.785	0.725	0.598
n=50	0.943	0.914	0.836	0.963	0.939	0.873	0.941	0.914	0.837	0.958	0.934	0.867
n=75	0.992	0.984	0.966	0.994	0.989	0.974	0.991	0.986	0.966	0.994	0.989	0.973
n=100	0.999	0.996	0.992	0.999	0.998	0.994	0.999	0.998	0.991	0.999	0.999	0.993
$\ln(\mu) = 2.5 - 0.15x_1 + 0.25x_2$ ค่าเฉลี่ยมีการกระจายมาก												
n=25	0.998	0.995	0.989	0.999	0.997	0.992	0.997	0.994	0.987	0.999	0.996	0.993
n=50	1	1	1	1	1	1	1	1	1	1	1	1
n=75	1	1	1	1	1	1	1	1	1	1	1	1
n=100	1	1	1	1	1	1	1	1	1	1	1	1

$^{+}$ Dean (1992), * Wang – Shu Lu (1997)

ตารางที่ 3.4 (ต่อ)

	Score test statistic $1-Q_{D1}$ [+]			**Adjusted statistic** $1-Q_{D2}$ [+]			**Score test statistic** $1-Q_{W1}$ [*]			**Adjusted statistic** $1-Q_{W2}$ [*]		
δ	**0.10**	**0.05**	**0.01**	**0.10**	**0.05**	**0.01**	**0.10**	**0.05**	**0.01**	**0.10**	**0.05**	**0.01**
$\ln(\mu) = 0.65 - 0.55x_1 + 0.50x_2 + 0.15x_1x_2$ ค่าเฉลี่ยมีการกระจายน้อย												
n=25	0.596	0.511	0.371	0.705	0.620	0.465	0.593	0.507	0.369	0.694	0.609	0.454
n=50	0.902	0.858	0.754	0.933	0.900	0.813	0.898	0.848	0.766	0.931	0.886	0.808
n=75	0.978	0.961	0.918	0.986	0.974	0.940	0.979	0.959	0.924	0.984	0.976	0.944
n=100	0.985	0.973	0.945	0.989	0.980	0.960	0.990	0.969	0.939	0.993	0.979	0.956
$\ln(\mu) = 3.25 - 0.65x_1 + 0.75x_2 + 0.25x_1x_2$ ค่าเฉลี่ยมีการกระจายมาก												
n=25	0.998	0.995	0.989	0.999	0.997	0.992	0.997	0.994	0.987	0.999	0.996	0.993
n=50	1	1	1	1	1	1	1	1	1	1	1	1
n=75	1	1	1	1	1	1	1	1	1	1	1	1
n=100	1	1	1	1	1	1	1	1	1	1	1	1

[+] Dean (1992), [*] Wang – Shu Lu (1997)

จากตารางที่ 3.4 เมื่อพิจารณาระหว่างค่า Score test statistic และค่า Adjusted statistic พบว่าที่ขนาดตัวอย่างเท่ากัน ค่า Adjusted statistic มีค่า Power of the test ใกล้เคียง 1 มากกว่าค่า Score test statistic ทั้งของ Dean (1992) และของ Wang – Shu Lu (1997) และเมื่อพิจารณาขนาดตัวอย่างที่แตกต่างกันพบว่า ถ้าขนาดตัวอย่างมาก ค่า Adjusted statistic และค่า Score test statistic มีค่า Power of the test ใกล้เคียง 1 มากกว่าขนาดตัวอย่างน้อย

3.4 Robust standard error

สูตรในการคำนวณ Robust variance สำหรับ $\hat{\underset{\sim}{\beta}}$ และ $\hat{\alpha}$ ของ NB2 สามารถหาได้ง่ายเนื่องจากเรามี $\dot{l}_{NB2i}$ ซึ่งเป็นสมาชิกของ Score vector $\underset{\sim}{s}$ และ Observed information matrix $I(\hat{\underset{\sim}{\beta}}, \hat{\alpha})$ จากกระบวนการวนซ้ำนิวตัน – ราฟสัน ใน (3.3) ดังนั้นค่าประมาณ Robust variance ของตัวแบบ NB2 แทนด้วย $\hat{\Lambda}_{\hat{\underset{\sim}{\beta}},\hat{\alpha}}$ คือ

$$\hat{\Lambda}_{\hat{\underset{\sim}{\beta}},\hat{\alpha}} = \{I(\hat{\underset{\sim}{\beta}},\hat{\alpha})\}^{-1}\sum\{\dot{I}_{NB2i}(\hat{\underset{\sim}{\beta}},\hat{\alpha})\}\{\dot{I}_{NB2i}(\hat{\underset{\sim}{\beta}},\hat{\alpha})\}^{T}\{I(\hat{\underset{\sim}{\beta}},\hat{\alpha})\}^{-1} \quad (3.6)$$

และค่าประมาณ Robust standard error ของ $\hat{\underset{\sim}{\beta}}$, $\hat{\alpha}$ คือรากที่สองของสมาชิกบนเส้นทแยงมุมหลักของ $\hat{\Lambda}_{\hat{\underset{\sim}{\beta}},\hat{\alpha}}$ ถ้าตัวแบบถดถอย NB2 เป็นตัวแบบที่ถูกต้องแล้ว Asymptotic standard error ของ $\hat{\underset{\sim}{\beta}}$ และ $\hat{\alpha}$ ที่ได้จากกระบวนการวนซ้ำ (3.3) กับค่าประมาณ Robust standard error ของ $\hat{\underset{\sim}{\beta}}$ และ $\hat{\alpha}$ จะเท่ากันโดยประมาณ ด้วยคุณสมบัตินี้ทำให้เราสนใจที่จะใช้เป็นกฎเกณฑ์หนึ่งในการตรวจสอบการใช้ mean – variance NB กับ Overdispersed Poisson counts วิธีการตรวจสอบที่นิยมใช้ก็คือ Simulation study

ชุดคำสั่งชื่อ nb2.fit.robust ที่เขียนขึ้นเพื่อคำนวณสมการนิวตัน – ราฟสัน (3.3) รวมทั้งคำนวณ $\hat{\Lambda}_{\hat{\underset{\sim}{\beta}},\hat{\alpha}}$ และค่าประมาณ Robust standard error ของ $\hat{\underset{\sim}{\beta}}$, $\hat{\alpha}$ ในโปรแกรม R ได้แสดงในภาคผนวก ข หัวข้อ 1.

3.5 Simulation study สำหรับ Robust standard error

ในการใช้ Simulation study เพื่อศึกษาเปรียบเทียบระหว่างค่าประมาณ Robust standard error และ Asymptotic standard error ของ $\hat{\underset{\sim}{\beta}}$ และ $\hat{\alpha}$ ที่ได้จากการเลือกใช้ตัวแบบ Mean – variance NB กับ Overdispersed Poisson counts เราได้ ดำเนินการ ดังนี้

1) จำลองข้อมูลภายใต้การแจกแจง NB2 จำนวน R = 5,000 ชุด โดยใช้ Working models ดังแสดงในตารางที่ 3.5 และแต่ละตัวแบบ ใช้ขนาดตัวอย่าง n = 25, 50 และ 100 ข้อมูลที่จำลองได้ในแต่ละชุดทำการ fit NB2 โดยมีตัวแปรอิสระเช่นเดียวกับ Working models และคำนวณค่าเฉลี่ยของ $\hat{\beta}$, $\hat{\alpha}$, Asymptotic standard error, Robust standard error และคำนวณ Unbiased standard error จากสูตร $\sqrt{\dfrac{\sum_{j=1}^{5000}\left(\hat{\beta}_j-\bar{\hat{\beta}}_j\right)^2}{5000-1}}$ แล้วนำเสนอค่าที่คำนวณได้ในตารางที่ 3.6 – 3.9

2) จำลองข้อมูลภายใต้การแจกแจง NB1 จำนวน R = 5,000 ชุด โดยใช้ Working models เดียวกันกับ 1) แต่ละชุดทำการ fit NB2 และดำเนินการเช่นเดียวกันกับ 1) ดังแสดงในตารางที่ 3.6 – 3.9

ตารางที่ 3.5 Working models สำหรับการจำลองข้อมูลเพื่อศึกษา Robust variance

ตัวแบบ	α	ค่าต่ำสุดของ μ	ค่าสูงสุดของ μ	ค่าเฉลี่ยของ μ
$\ln(\mu) = 1.95$	1.5	-	-	7.029
$\ln(\mu) = 2.5 + 0.5x_1$	1.5	4.482	33.115	15.451
$\ln(\mu) = 2.5 - 0.15x_1 + 0.25x_2$	1.5	0.779	314.190	77.048
$\ln(\mu) = 3.25 - 0.65x_1 + 0.75x_2 + 0.25x_1x_2$	1.5	24.533	121.510	51.567

x_1 คือตัวแปรเชิงปริมาณ มีค่า -2, -1, 0, 1, 2 และ x_2 คือตัวแปรเชิงคุณภาพ มีค่า 1, 0, 0, 1, 1 โดยแต่ละค่าของ x_1 และ x_2 มีจำนวนซ้ำ r = 5, 10, 20 สำหรับ n = 25, 50, 100 ตามลำดับ

ตารางที่ 3.6 Parameter estimates (P.est), Asymptotic standard error (A.SE), Robust standard error (R.SE) และ Unbiased standard error estimates (U.SE) เปรียบเทียบระหว่างข้อมูลที่มีการแจกแจง NB2 fit NB2 กับข้อมูลที่มีการแจกแจง NB1 fit NB2 ภายใต้ตัวแบบ $\ln(\mu) = 1.95$

n	$\hat{\beta}, \hat{\alpha}$	ข้อมูลที่มีการแจกแจง NB2 Fit NB2				ข้อมูลที่มีการแจกแจง NB1 Fit NB2			
		P.est	A.SE	R.SE	U.SE	P. est	A.SE	R.SE	U.SE
25	β_0	1.92268	0.24781	0.23946	0.25933	1.94292	0.11654	0.11572	0.11779
	α	1.43840	0.47407	0.45949	0.54874	0.20308	0.10127	0.09604	0.10248
50	β_0	1.93262	0.17822	0.17556	0.18287	1.94681	0.08336	0.08306	0.08300
	α	1.46969	0.33787	0.33084	0.39121	0.20837	0.07234	0.07023	0.07338
100	β_0	1.94185	0.12716	0.12608	0.12628	1.94846	0.05916	0.05906	0.05888
	α	1.48642	0.24021	0.23775	0.27336	0.20937	0.05118	0.05033	0.05138

ตารางที่ 3.7 Parameter estimates (P.est), Asymptotic standard error (A.SE), Robust standard error (R.SE) และ Unbiased standard error estimates (U.SE) เปรียบเทียบระหว่างข้อมูลที่มีการแจกแจง NB2 fit NB2 กับข้อมูลที่มีการแจกแจง NB1 fit NB2 ภายใต้ตัวแบบ $\ln(\mu) = 2.5 + 0.5x_1$

n	$\hat{\underset{\sim}{\beta}}, \hat{\alpha}$	**ข้อมูลที่มีการแจกแจง NB2 Fit NB2**				**ข้อมูลที่มีการแจกแจง NB1 Fit NB2**			
		P.est	**A.SE**	**R.SE**	**U.SE**	**P. est**	**A.SE**	**R.SE**	**U.SE**
25	β_0	2.44558	0.23888	0.23772	0.50189	2.49023	0.08766	0.09647	0.12236
	β_1	0.46967	0.17477	0.17366	1.09196	0.50582	0.06207	0.06565	0.09807
	α	1.35727	0.41749	0.40559	0.50429	0.09133	0.05547	0.07467	0.05506
50	β_0	2.46888	0.17294	0.17308	0.30774	2.49603	0.06255	0.06914	0.07112
	β_1	0.51238	0.12459	0.12213	0.44607	0.50164	0.04429	0.04714	0.04842
	α	1.41596	0.30354	0.29751	0.38390	0.09457	0.03793	0.04054	0.04090
100	β_0	2.48333	0.12461	0.12312	0.12866	2.49881	0.04458	0.04957	0.04990
	β_1	0.50049	0.08916	0.08635	0.09108	0.50028	0.03160	0.03391	0.03443
	α	1.46121	0.21953	0.21712	0.26541	0.09734	0.02723	0.02933	0.02923

ตารางที่ 3.8 Parameter estimates (P.est), Asymptotic standard error (A.SE), Robust standard error (R.SE) และ Unbiased standard error estimates (U.SE) เปรียบเทียบระหว่างข้อมูลที่มีการแจกแจง NB2 fit NB2 กับข้อมูลที่มีการแจกแจง NB1 fit NB2 ภายใต้ตัวแบบ $\ln(\mu) = 2.5 - 0.15x_1 + 0.25x_2$

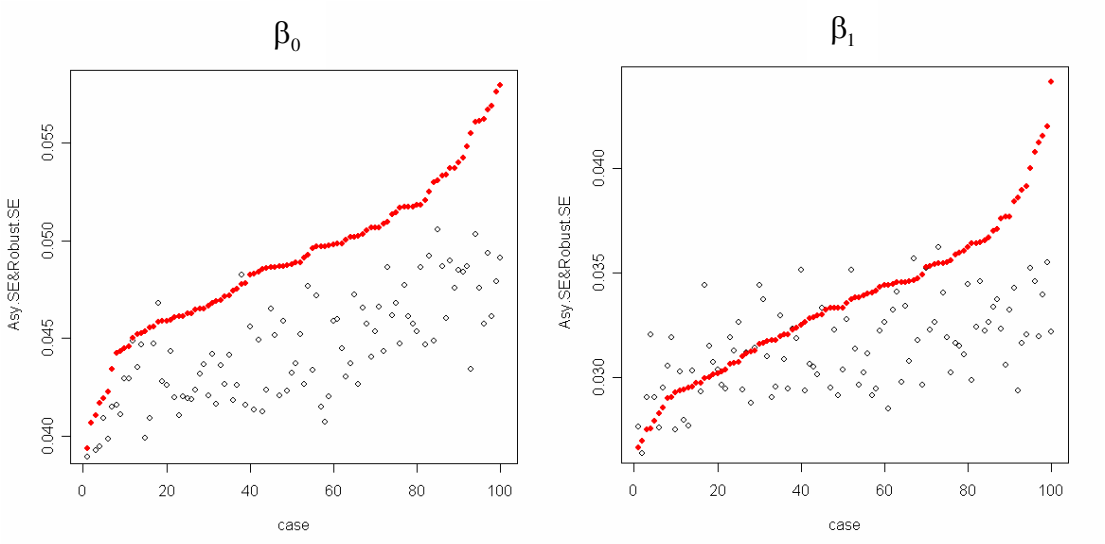
n	$\hat{\underset{\sim}{\beta}}, \hat{\alpha}$	ข้อมูลที่มีการแจกแจง NB2 Fit NB2				ข้อมูลที่มีการแจกแจง NB1 Fit NB2			
		P.est	**A.SE**	**R.SE**	**U.SE**	**P. est**	**A.SE**	**R.SE**	**U.SE**
25	β_0	2.43379	0.38215	0.35116	0.42637	2.48925	0.12966	0.13013	0.14074
	β_1	-0.13130	0.17572	0.15125	0.19879	-0.14972	0.05618	0.05236	0.05847
	β_2	0.20686	0.50218	0.45359	0.57328	0.25114	0.16542	0.16120	0.17341
	α	1.32091	0.39762	0.38658	0.46214	0.08539	0.04631	0.04566	0.04607
50	β_0	2.45929	0.27942	0.26673	0.29487	2.49642	0.09431	0.09594	0.10075
	β_1	-0.14357	0.12724	0.11650	0.13462	-0.15003	0.04102	0.03884	0.04084
	β_2	0.23922	0.36635	0.34598	0.39118	0.24808	0.12059	0.11862	0.12514
	α	1.41253	0.29501	0.29003	0.33456	0.09414	0.03369	0.03281	0.03484
100	β_0	2.48138	0.20093	0.19592	0.20991	2.49698	0.06770	0.06911	0.06997
	β_1	-0.14698	0.09131	0.08699	0.09593	-0.15040	0.02949	0.02813	0.02861
	β_2	0.23958	0.26329	0.25547	0.27537	0.25055	0.08663	0.08542	0.08710
	α	1.45901	0.21339	0.21082	0.23655	0.09861	0.02446	0.02431	0.02498

ตารางที่ 3.9 Parameter estimates (P.est), Asymptotic standard error (A.SE), Robust standard error (R.SE) และ Unbiased standard error estimates (U.SE) เปรียบเทียบระหว่างข้อมูลที่มีการแจกแจง NB2 fit NB2 กับข้อมูลที่มีการแจกแจง NB1 fit NB2 ภายใต้ตัวแบบ $\ln(\mu) = 3.25 - 0.65x_1 + 0.75x_2 + 0.25x_1x_2$

n	$\hat{\beta}, \hat{\alpha}$	ข้อมูลที่มีการแจกแจง NB2 Fit NB2				ข้อมูลที่มีการแจกแจง NB1 Fit NB2			
		P.est	**A.SE**	**R.SE**	**U.SE**	**P. est**	**A.SE**	**R.SE**	**U.SE**
25	β_0	3.08633	0.50714	0.42012	0.62027	3.23885	0.11395	0.12526	0.12834
	β_1	-0.65079	0.71312	0.60361	0.87610	-0.66363	0.14865	0.16027	0.16346
	β_2	0.80746	0.58648	0.50474	0.71696	0.76016	0.12704	0.13816	0.14167
	β_3	0.26405	0.73373	0.62375	0.62375	0.26626	0.15210	0.16332	0.16844
	α	1.27131	0.34736	0.34427	0.42774	0.02657	0.01499	0.01579	0.01421
50	β_0	3.16514	0.37491	0.34268	0.41832	3.24272	0.08139	0.09107	0.09912
	β_1	-0.65804	0.52799	0.48579	0.58300	-0.62240	0.10652	0.11397	0.12199
	β_2	0.78464	0.43408	0.40410	0.48682	0.75502	0.09088	0.10012	0.10776
	β_3	0.26663	0.54309	0.50076	0.60328	0.25436	0.10896	0.11618	0.12459
	α	1.38669	0.26348	0.26007	0.33073	0.02787	0.01092	0.01165	0.01209
100	β_0	3.21538	0.26925	0.25669	0.28488	3.24655	0.05846	0.06678	0.06983
	β_1	-0.64561	0.37941	0.36515	0.39949	-0.65237	0.07679	0.08298	0.08703
	β_2	0.75725	0.31182	0.30103	0.33026	0.75233	0.06539	0.07299	0.07620
	β_3	0.24902	0.39020	0.37603	0.41005	0.25198	0.07857	0.08452	0.08840
	α	1.42938	0.19053	0.18885	0.26869	0.02978	0.00809	0.00877	0.00875

จากตารางที่ 3.6 – 3.9 ได้ว่าถ้าใช้ตัวแบบที่ถูกต้องกับข้อมูล คือข้อมูลจากการแจกแจง NB2 แล้วfit NB2 จะให้ค่าเฉลี่ยของ $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$ และ $\hat{\alpha}$ ค่าใกล้เคียงกับค่าพารามิเตอร์ที่กำหนดไว้ในแต่ละ Working model แต่ถ้าใช้ตัวแบบไม่เหมาะสมกับข้อมูล เช่น fit NB2 กับข้อมูลจากการแจกแจง NB1 ให้ค่าเฉลี่ยของ $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$ มีค่าใกล้เคียงกับค่าพารามิเตอร์ที่กำหนดไว้ในแต่ละ Working model ยกเว้นค่า $\hat{\alpha}$ และเมื่อพิจารณา Asymptotic standard error, Robust standard error และ Unbiased standard error estimates พบว่าข้อมูลที่จำลองภายใต้การแจกแจง NB2 แล้ว fit NB2 ให้ค่าคลาดเคลื่อนมาตรฐานทั้ง 3 ชนิดใกล้เคียงกัน ส่วนข้อมูลจากการแจกแจง

NB1 แล้ว fit NB2 จะให้ค่า Robust standard error และ Unbiased standard error estimates ใกล้เคียงกัน แต่มีค่ามากกว่า Asymptotic standard error นอกจากนี้ ภาพที่ 3.1 ซึ่งเป็นการ เปรียบเทียบระหว่าง Asymptotic standard error กับ Robust standard error ของ β_0 และ β_1 ซึ่งคำนวณได้จากข้อมูลที่จำลองภายใต้การแจกแจง NB1 ที่มี $\ln(\mu_i) = 2.5 - 0.5x_{1i}$ และ α เท่ากับ 1.5 จำนวน 100 ชุด แต่ละชุดใช้ตัวอย่างขนาด 100 แล้วทำการ fit NB2 พบว่าค่า Asymptotic standard error และค่า Robust standard error แตกต่างกัน และส่วนใหญ่จะมีค่า Robust standard error มากกว่าค่า Asymptotic standard error



ภาพที่ 3.1 Simulated asymptotic standard error (○) และ Simulated robust standard error (●) ของ β_0 และ β_1 สำหรับข้อมูลจากการแจกแจง NB1ที่มี $\ln(\mu_i) = 2.5 - 0.5x_{1i}$ และ $\alpha = 1.5$ แล้ว fit NB2

ดังนั้นผลจาก Simulation study จึงเป็นหลักฐานเพียงพอที่จะสรุปได้ว่าสามารถใช้ Estimated robust standard error ของ NB2 คือ $\hat{\Lambda}_{\hat{\beta},\hat{\alpha}}$ เป็นเครื่องมือในการตรวจสอบว่า Overdispered Poisson counts นั้นมีการแจกแจงแบบ NB2 หรือไม่

3.6 การวินิจฉัยตัวแบบ (Model Diagnostics)

การวินิจฉัยตัวแบบโดยใช้ half normal plot with simulated envelope สำหรับ normal regression model ซึ่งเสนอโดย Atkinson (1985) เป็นการตรวจสอบการแจกแจงของเศษตกค้าง (Residual) โดยการนำค่าสัมบูรณ์ของ Pearson residual ที่ถูกเรียงลำดับ (น้อย $\rightarrow$ มาก) จากตัวแบบที่ได้ และค่าต่ำสุด ค่าสูงสุด และค่าเฉลี่ยของค่าสัมบูรณ์ของ Pearson residual ที่ดำเนินการในทำนองเดียวกันจากตัวแบบที่ได้จากข้อมูลที่ได้จากการจำลอง 19 ชุด มาวาดกราฟกับ Half-normal scores $\phi^{-1}\left(\frac{i+n}{n}\right)$ โดยที่ ϕ คือฟังก์ชันการแจกแจงสะสมของการแจกแจงปรกติมาตรฐาน ต่อมา Demètrio and Hinde (1997) ได้ปรับเป็น Half-normal scores ข้างต้นเป็น $\phi^{-1}\left(\frac{i+n-0.125}{2n+0.5}\right)$ เพื่อให้ได้ค่าที่ใกล้เคียงกับค่าคาดหวังของสถิติลำดับมากขึ้น ถ้าค่าของ Pearson residual ทุกค่าของตัวแบบที่เราตรวจสอบ อยู่ใน envelope แสดงว่าตัวแบบนั้นถูกต้อง ต่อมา Williams (1987) ได้แสดงให้เห็นว่าสามารถใช้ Half normal plot with simulated envelope ตรวจสอบความถูกต้องของตัวแบบเชิงเส้นวางนัยทั่วไป โดยค่าที่นำมาใช้ในการตรวจสอบคือ ค่าสัมบูรณ์ของ Standardized deviance residual ที่ถูกเรียงลำดับ (น้อย $\rightarrow$ มาก) ของตัวแบบ นอกจากนี้เขายังกล่าวว่า Half normal plot with simulated envelope สามารถใช้ตรวจสอบได้ทั้งข้อสมมติเบื้องต้นของตัวแบบ และ overdispersion (ทั้งค่าน้อยและค่ามากของ Standardized deviance residual ของตัวแบบ อยู่นอก upper envelope) ได้ด้วย

สำหรับตัวแบบ NB2 Standardized deviance residual คือ

$$r^{\bullet}_{Dnb2,i} = \frac{r_{Dnb2,i}}{\sqrt{(1+h_{ii})}} \tag{3.7}$$

เมื่อ

$$r_{Dnb2,i} = \operatorname{sgn}(y_i-\hat{\mu}_i)\sqrt{2\left\{y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right)-\left(y+\frac{1}{\hat{\alpha}}\right)\ln\left(\frac{1+\hat{\alpha}y_i}{1+\hat{\alpha}\hat{\mu}_i}\right)\right\}} \tag{3.8}$$

h_{ii} คือ Leverage ซึ่งเป็นสมาชิกบนแนวทแยงมุมหลักของ $\hat{W}^{\frac{1}{2}}X(X^T\hat{W}X)^{-1}X^T\hat{W}^{\frac{1}{2}}$ ของการถดถอย NB2 $\hat{W}$ คือ Weight matrix ที่มี $\hat{w}_{ii}=(1+\hat{\alpha}\hat{\mu}_i)\hat{\mu}_i$ และ $\operatorname{sgn}(y_i-\hat{\mu}_i)$ คือฟังก์ชันที่ทำให้ $r_{Dnb2,i}$ มีเครื่องหมายเช่นเดียวกับ $y_i-\hat{\mu}_i$

Half normal plot with simulated envelope สำหรับตรวจสอบตัวแบบ NB2 มีขั้นตอนการสร้าง ดังนี้

1) Fit ตัวแบบ NB2 คำนวณค่าสัมบูรณ์ของ $r^{\bullet}_{Dnb2,i}$ แล้วเรียงลำดับจากน้อยไปมาก สมมติแทนด้วย $r^{\bullet}_{Dnb2,(i)}$

2) จำลองค่าของตัวแปร NB2 ภายใต้ข้อสมมติว่าตัวแบบในข้อ 1) ถูกต้อง จำนวน 19 ชุด โดยให้ $e_{0j,i} \sim \Gamma(\hat{\alpha}^{-1},1)$, $j=1,2,...,19$; $i=1,2,...,n$ คำนวณ $e_{ji} = e_{0j,i} \times \hat{\alpha}$ แล้วจำลอง $Y_{ji} \sim Pois(\hat{\mu}_i e_{ji})$ แล้ว Y_{ji} ที่ได้คือ $Y_{ji} \sim NB2(\hat{\mu}_i, \hat{\alpha})$

3) Fit ตัวแบบโดยใช้ตัวแปรตามจากข้อ 2) และใช้ตัวแปรอิสระชุดเดียวกับตัวแบบในข้อ 1) คำนวณค่าสัมบูรณ์ของ Standardized deviance residual แล้วเรียงลำดับจากน้อยไปมาก และแทนด้วย $r^{*}_{j(D,i)}$, $j=1,2,...,19$, $i=1,2,...,n$

4) คำนวณค่าต่ำสุด ค่าสูงสุด และค่าเฉลี่ย ของ $r^{*}_{j(D,i)}$

5) พล็อตค่า $r^{\bullet}_{Dnb2,(i)}$ ค่าต่ำสุด ค่าสูงสุด และค่าเฉลี่ย ของ $d^{*}_{k(i)}$ กับค่า Half-normal scores $\phi^{-1}\left(\frac{i+n-0.125}{2n+0.5}\right)$ (Demètrio and Hinde, 1997)

การตรวจสอบความถูกต้องของตัวแบบโดย Half normal plot with simulated envelope พิจารณากราฟ Half normal plot ของ NB2 ถ้ามีค่า Standardized deviance residual ที่ได้จากการ fit ตัวแบบทุกค่าอยู่ภายในเส้นปะของค่าเฉลี่ยของ Standardized deviance residual ที่ได้จากการ simulate ไม่มีค่า Standardized deviance residual ที่ได้จากการ fit ตัวแบบตัวใดอยู่นอกเส้น Upper envelope และ lower envelope จะได้ว่า NB2 เป็นตัวแบบที่เหมาะสมกับข้อมูลมากที่สุด