Y. Y. Ü. Fen Bilimleri Enstitüsü Dergisi
Yıl : 2008, Cilt : 13, Sayı : 1, Sayfa : 14-18

Sıfır Değer Ağırlıklı Regresyon Yöntemleri

Abdullah YEŞİLOVA[1] **Barış KAKİ**[1]
[1] Yüzüncü Yıl Üniversitesi, Ziraat Fakültesi, Zootekni Bölümü, 65080, VAN

Özet: Sayıma dayalı olarak elde edilen verilerin analizinde, Poisson ve negatif binomial regresyon yöntemleri kullanılmaktadır. Poisson dağılımına sahip sayıma dayalı olarak elde edilen veriler, beklenenden daha fazla sayıda sıfır değerine sahip olabilir. Fazla sayıda sıfır değerine sahip bağımlı değişkenin modellenmesinde, sıfır ağırlıklı poisson (ZIP) regresyonunun kullanılması uygun bir yaklaşımdır. ZIP, veri kümesinin iki farklı tip veriden oluştuğunu varsaymaktadır. Bunlardan birincisi, sıfır değerlerine sahip olabilen ve Poisson dağılımına sahip sayıma dayalı veriler, buna karşın ikinci tip ise daima sıfır değerleri alan veriler olmaktadır. Sıfır değer ağırlıklı negatif binomial (ZINB) regresyon ve hurdle model sıfır değerlerinin çok olduğu veri kümeleri için kullanılan alternatif modellerdir. ZIP regresyonunda olduğu gibi, ZINB regresyonunda da sıfır olan ve olmayan gözlemler iki faklı şekilde modellenmektedir. Hurdle model ise iki aşamadan oluşmaktadır. Birincisi, sıfır sayımlara (0) karşı pozitif sayımları(1) gösteren binary cevaplar; ikincisi ise yalnız pozitif sayımların meydana geldiği süreçtir. Binary cevaplar binary model kullanılarak modellenmektedir. Pozitif sayımlar ise sıfır değer sınırlandırılmış (zero-truncated) sayma model kullanılarak modellenmektedir. PR, NB, ZIP, ZINB ve hurdle modellerinde Parametre tahminlerinin elde edilmesinde, EM algoritmasını esas alan en yüksek olabilirlik (ML) yöntemi kullanılmaktadır. Bu çalışmada, sıfır değer ağırlıklı regresyon modellerin teorik özelliklerinin incelenmesi amaçlanmıştır.

Anahtar Sözcükler: Aşırı yayılım, en yüksek olabilirlik, hurdle model, sıfır değer ağırlıklı veri

Zero-Inflated Regression Methods

Abstract: Poisson regression (PR) and negative binomial (NB) regression methods are used in the analysis of based count data. The observation obtained based on count data having Poisson distribution may have more zero values rather than expected. In this case, use of zero-inflated Poisson (ZIP) regression is a suitable approach to model the dependent variable having excessive zero values. ZIP assumes that data set consists of two different types of data. First type is count data based on Poisson distribution which may have zero values, whereas second type is data having only zero values. Zero-inflated negative binomial (ZINB) regression and hurdle model are alternative models used for data set having excessive zero-values. As in ZIP regression, in ZINB regression, observations with and without zero are modeled in two different ways. On the other hand, Hurdle model consists of two stages. First is binary response showing positive counts (1) against zero counts (0). Second is process with only positive counts. Binary responses are modeled by using binary model. Positive counts are modeled by using zero-truncated model. Binary part uses logit and probit, and count part uses Poisson, geometric, and negative binomial. Generally, for positive count part, Poisson hurdle (HP) and negative binomial hurdle (HNB) are used. In PR, NB, ZIP, ZINB, and hurdle models, Maximum likelihood (ML) based on EM algorithm is used for estimation of parameters. In the present study aimed to investigate the theoretical properties of zero-inflated regression models.

Key words: Hurdle model, maximum likelihood, overdispersion, zero-inflated data

Giriş

Poisson regresyon (PR) sayıma dayalı olarak elde edilen verilerin analizinde yoğun olarak kullanmaktadır (Frome ve ark., 1973; Agresti, 1997; Cameron ve Trivedi, 1998; Böhning, 1994; Stokes ve ark., 2000). PR, bağımsız değişkenler ile sayıma dayalı olarak elde edilen bağımlı değişken arasındaki ilişkiyi açıklamaktadır. PR'de bağımsız değişkenlerin doğrusal yapısını bağımlı değişkenin beklenen değerine bağlayan bağlantı fonksiyonu, logaritmik dönüşüm ile verilmektedir (Nelder ve Wedderburn, 1972; McCullagh ve Nelder, 1989; Breslow, 1990; Long ve Freese, 2006).

Bilindiği gibi, Poisson dağılımında ortalama ile varyans birbirine eşittir. Ancak uygulamada bu eşitliği sağlamak her zaman mümkün değildir. Varyansın ortalamadan büyük çıkması aşırı yayılım (overdispersion), küçük çıkması da az yayılım (underdispersion) olarak tanımlanmaktadır (Breslow, 1990; Böhning, 1994; SAS, 2007; Yeşilova ve ark., 2007). Veri kümelerinde genellikle aşırı yayılım, nadiren de az yayılım ile karşılaşılmaktadır. Böyle durumlarda PR'yi uygulamak, yanlı parametre tahminlerinin elde edilmesine neden olur (Cox, 1983; Breslow, 1990). Veri setinde aşırı yayılım söz konusu olduğunda negatif binomial (NB) regresyon yaklaşımının kullanılması daha uygun olmaktadır (Lawles, 1987; Jansakul, 2005). NB regresyon modeli, aşırı yayılımdan kaynaklanan etkiyi dikkate alarak parametre tahmini yapmaktadır.

Poisson dağılımı gösteren sayıma dayalı olarak elde edilen veriler, beklenenden daha fazla sayıda sıfır değerine sahip olabilir. Böyle bir durumda fazla sayıda sıfır değerine sahip bağımlı değişkenin modellenmesinde, sıfır değer ağırlıklı Poisson (ZIP) regression modelinin kullanılması daha uygun bir yaklaşımdır (Lambert, 1992; Böhning, 1998; Böhning ve ark. 1999; Ridout ve ark. 2001; Yau ve Lee, 2001). ZIP, üzerinde çalışılan örneğin iki farklı tip veriden oluştuğunu varsaymaktadır. Bunlardan birincisi, sıfır değerlerine sahip olabilen Poisson dağılımlı sayıma dayalı veriler olurken, ikinci tip ise daima sıfır değerleri alan verilerdir. Bununla birlikte, yukarıda bahsedilen aşırı yayılım durumu, sıfır değerlerinin çok olduğu veri kümelerinde de söz konusudur. Böyle durumlarda, sıfır değer ağırlıklandırılmış negatif binomial (ZINB) regresyonu alternatif bir yöntemdir (Lawless, 1987; Hall, 2000; Ridout ve ark., 2001; Jansakul, 2005; Long ve Freese, 2006; Yeşilova ve ark., 2007). ZINB regresyonu, aşırı yayılımdan kaynaklanan kısmı, α ile modele dahil etmektedir. ZIP regresyonunda olduğu gibi, ZINB regresyonunda da sıfır olan ve olmayan gözlemler faklı şekilde modellenmektedir. Bunun yanı sıra hurdle model sıfır değerlerinin çok olduğu veri kümeleri için kullanılmaktadır. Hurdle model iki

aşamadan oluşmaktadır. Birincisi, sıfır sayımlara (0) karşı pozitif sayımları(1) gösteren binary cevaplar; ikincisi ise yalnız pozitif sayımların meydana geldiği süreçtir. Binary cevaplar binary (ikili) model kullanılarak modellenmektedir. Pozitif sayımlar ise sıfır değer sınırlandırılmış (zero-value truncated) sayma model kullanılarak modellenmektedir (Martin ve ark., 2006; Hilbe, 2007). Binary kısım logit, probit veya complememtary loglog kullanırken, sayma kısmı ise Poisson, geometrik ve negatif binomial dağılım kullanmaktadır. Pozitif sayımlar kısmı için Poisson hurdle (PH) ya da negatif binomial hurdle (NBH) kullanılmaktadır.

PR, NB, ZINB, PH ve NBH regresyon yaklaşımlarında, parametre tahminleri ML yöntemi kullanılarak elde edilmektedir (Long ve Freese, 2006). Uygun model seçiminde Akaiki bilgi ölçütü ile Bayesian bilgi ölçütü kullanılabilir. En küçük uyum ölçütlerine sahip model en iyi model olarak kabul edilmektedir (Dalrymple ve ark., 2003). Bu çalışmada, sıfır değerlerinin çok olduğu veri kümelerinde, sıfır değerlerini farklı bir şekilde modelleyen regresyon yöntemlerin teorik özelliklerinin incelenmesi amaçlanmıştır.

Bu çalışmada, sıfır değer ağırlıklı verilerin analizinde kullanılan yöntemlerin teorik özelliklerinin incelenmesi amaçlanmıştır. Bu amaçla, PR, NB, ZINB, PH ve NBH yöntemleri için model tanımlaması yapılarak, log-olabilirlik fonksiyonları ve en yüksek olabilirlik (ML) tahminleri elde edilmiştir.

Yöntem

Bu bölümde, Poisson regresyonu, negatif binomial regresyon, sıfır değer ağırlıklı Poisson regresyon, sıfır değer ağırlıklı negatif binomial regresyon, Poisson hurdle ve negatif binomial hurdle modellerinin teorik özellikleri incelenecektir.

Poisson regresyon (Poisson regression): PR'de ilgilenilen olayın gözlenen sayısı olan y_i bağımlı değişkenin Poisson dağılışına sahip olduğu varsayılmaktadır. Poisson ortalaması olan μ'nün logaritmasının, bağımsız değişkenlerin bir doğrusal fonksiyonu olduğu varsayılmaktadır (Cheung, 2002; Yesilova ve ark., 2007; SAS, 2007). Log bağlantı fonksiyonlu Poisson regresyon modeli (Lee ve Wang, 2001),

$$\Pr(y_i/\mu_i, x_i) = \exp(-\mu_i)\mu_i^{y_i}/y_i!, \quad y_i = 0,1,... \tag{1}$$

biçiminde verilmektedir (Cameron ve Trivedi, 1998; Long ve Freese, 2006; Yeşilova ve ark., 2007). PR modelinde, ML kullanılarak parametre tahminleri elde edilmektedir. Bağımsız gözlemler verilmişken, PR modeli için log olabilirlik fonksiyonu,

$$L(\beta/y_i, x_i) = \sum_{i=1}^{n}\left[y_i x_i'\beta - \exp(x_i'\beta) - \ln y_i!\right] \tag{2}$$

biçiminde yazılabilir (Khoshgoftaar ve ark., 2005; SAS, 2007). Eşitlik 2'de β bilinmeyen parametre vektörünü göstermektedir. β bilinmeyen parametre vektörü, log-olabilirlik fonksiyonun maksimize edilmesiyle tahmin edilmektedir.

Negatif binomial regresyon (Negative binomial regression): NB regresyon modeli, bağımlı değişken ile bağımsız değişkenler vektörü arasında, log bağlantı fonksiyonunu kullanmaktadır. NB regresyon modeli (Hall, 2000),

$$\Pr(Y = y_i/x_i) = \frac{\Gamma\left(y_i + \frac{1}{\alpha}\right)(\alpha\mu_i)^{y_i}}{y_i!\Gamma\left(\frac{1}{\alpha}\right)(1+\alpha\mu_i)^{y_i+\frac{1}{\alpha}}} \quad \alpha > 0 \tag{3}$$

biçiminde verilmektedir. Eşitlik 3'te, α aşırı yayılım derecesini gösteren yardımcı parametredir. NB regresyon modeline ilişkin log olabilirlik fonksiyonu (Lawles, 1987),

$$L(\beta,\alpha,y) = \sum_{i=1}^{n}\left[\frac{1}{\alpha}\log(1+\alpha\mu_i) - y_i\log\left(1+\frac{1}{\alpha\mu_i}\right) + \log\Gamma\left(y_i+\frac{1}{\alpha}\right) - \log\Gamma\left(\frac{1}{\alpha}\right) - \log y_i!\right] \tag{4}$$

biçiminde yazılabilir.

Sıfır değer ağırlıklı poisson regresyon (Zero-inflated Poisson regression) : y_i ekstra sıfırların sayısını açıklamak için, ZIP regresyon modeli,

$$\Pr(y_i/x_i) = \begin{cases} \pi_i + (1-\pi_i)\exp(-\mu_i) & y_i = 0 \\ (1-\pi_i)\exp(-\mu_i)\mu_i^{y_i}/y_i! & y_i > 0 \end{cases} \tag{5}$$

biçiminde yazılabilir (Ridout ve ark., 2001; Cheung, 2002). Eşitlik 5'te, π_i ekstra sıfırların olma olasılığını göstermektedir. Bundan dolayı $y_i = 0$ olan bireyler, iki gruptan oluşmuş şekilde tanımlanır. Bu gruplardan biri, deneğin Poisson süreci göstermediği, diğeri ise deneklerin

$$\exp(-\mu_i)\mu_i^0/0! = \exp(-\mu_i)$$

olmasından dolayı, sıfır değerleri alan μ ortalamalı Poisson dağılımına sahip olduğunu gösterir (Böhning, 1998).

ZIP regresyon model,

$$\log(\mu) = B\beta \tag{6}$$

ve

$$\log\left(\frac{\pi}{1-\pi}\right) = G\gamma \tag{7}$$

bağlantı (link) fonksiyonları kullanılarak elde edilebilir (Lambert, 1992). Eşitlik 6 ve 7'de B ve G kovariet (eş değişken) matrisleri, β ve γ sırasıyla, (p+1)x1 ve (q+1)x1 boyutlu bilinmeyen parametre vektörleridir. ZIP regresyon modeli için log olabilirlik fonksiyonu,

$$L(y,\beta,\gamma) = \sum_{y_i=0}\log\left(e^{G_i\gamma} + \exp\left(-e^{B_i\beta}\right)\right) + \sum_{y_i>0}\left(y_i B_i\beta - e^{B_i\beta}\right) - \sum_{i=1}^{n}\log\left(1+e^{G_i\gamma}\right) - \sum_{y_i>0}\log(y_i!) \tag{8}$$

biçiminde yazılabilir (Lambert, 1992). Eşitlik 8'de G_i ve B_i, G ve B matrislerinin i'ninci sırasını göstermektedir. Eşitlik 8'de verilen log olabilirlik fonksiyonundaki üssel terimlerin

maksimize edilmesi oldukça karmaşıktır. Bu nedenle söz konusu log olabilirlik fonksiyonun maksimize edilmesi için farklı bir yol izlenmektedir. Bunun için sıfır ve bir değerlerini alan ve tesadüfi olduğu varsayılan z_i indikatör (belirleyici) değişkeni modele dahil edilir. Y_i değeri sıfır olduğunda $z_i = 1$ ve Y_i Poisson durumda (sıfırdan büyük değerler aldığında) $z_i = 0$ olduğu varsayılır. Bu durumda, tüm veriler için log olabilirlik fonksiyonu,

$$\begin{aligned} L(\gamma,\beta,y,z) &= \sum_{i=1}^{n}\log\left(f(z_i/\gamma)\right)+\sum_{i=1}^{n}\log\left(f(y_i/z_i,\beta)\right) \\ &= \sum_{i=1}^{n}\left(z_i G_i\gamma - \log\left(1+e^{G_i\gamma}\right)\right) \\ &\quad +\sum_{i=1}^{n}(1-z_i)\left(y_i B_i\beta - e^{B_i\beta}\right)-\sum_{i=1}^{n}(1-z_i)\log(y_i!) \\ &= L(\gamma;y,z)+L(\beta,y,z)-\sum_{i=1}^{n}(1-z_i)\log(y_i!) \end{aligned} \tag{9}$$

biçiminde yazılabilir (Lambert, 1992). EM algoritması kullanılarak, E ve M aşamaları aşağıdaki gibi yazılabilir.

E-aşaması: gözlenmiş veriler verilmişken, z_i tesadüfi indikatör değişkeni,

$$z_i = \begin{cases} \left(1+e^{-G_i\gamma(K)-\exp\left(B_i\beta^{(k)}\right)}\right)^{-1}, & y_i = 0 \\ 0, & y_i = 1,2,... \end{cases} \tag{10}$$

biçiminde yazılabilir. Eşitlik 10'da verilen k, EM algoritmasının iterasyon sayısını göstermektedir.

M-aşaması: β parametresi ağırlıklandırılmış log doğrusal Poisson regresyon model kullanılarak tahmin edilmektedir. Tüm veriler için eşitlik 9'da verilen log olabilirlik fonksiyonunun maksimize edilmesiyle γ parametresi,

$$\begin{aligned} L\left(\gamma;y,Z^{(k)}\right) &= \Sigma_{y_i=0}Z_i^{(k)}G_i\gamma - \Sigma_{y_i=0}Z_i^{(k)}\log\left(1+e^{G_i\gamma}\right) \\ &\quad -\Sigma_{y_i=0}\left(1-Z_i^{(k)}\right)\log\left(1+e^{G_i\gamma}\right) \end{aligned}$$

biçiminde tahmin edilebilir. Yukarıda verilen E ve M aşamaları yakınsama ölçütü (10^{-6}) elde edilinceye kadar devam edilir (Lambert, 1992). Yakınsama ölçütü olarak Newton-Raphson algoritması kullanılmaktadır.

Sıfır değer ağırlıklı binomial regresyon (Zero-inflated negative binomial regression): Sıfır değerlerinin çok fazla olduğu y_i bağımlı değişkeninin modellenmesinde alternatif regresyon yöntemi, ZINB'dir. ZINB regresyon modeli,

$$\Pr(y_i/x_i) = \begin{cases} \pi_i + (1-\pi_i)(1+\alpha\mu_i)^{-\alpha^{-1}}, & y_i = 0 \\ (1-\pi_i)\dfrac{\Gamma\left(y_i-\alpha^{-1}\right)\alpha^{y_i}\mu_i^{y_i}}{y_i!\Gamma\left(\alpha^{-1}\right)\left(1+\alpha\mu_i\right)^{y_i+\alpha^{-1}}}, & y_i > 0 \end{cases} \tag{11}$$

biçiminde yazılabilir (Ridout ve ark., 2001; Jansakul, 2005). Eşitlik 11'de, α $(\alpha \geq 0)$ yayılım parametresini göstermektedir. Bağımsız y_i gözlem değerleri için ZINB log olabilirlik fonksiyonu aşağıdaki gibi yazılabilir (Jansakul, 2005),

$$\begin{aligned} L(\mu,\alpha,\pi;y) &= \sum_i\Bigg(I_{y_i=0}\log(1-\pi_i)(1+\alpha\mu_i)^{-\alpha^{-1}} + \\ &\quad I_{y_i>0}\log\left((1-\pi_i)\frac{\Gamma\left(y_i+\frac{1}{\alpha}\right)(\alpha\mu_i)^{y_i}}{y_i!\Gamma\left(\frac{1}{\alpha}\right)(1+\alpha\mu_i)^{y_i+\frac{1}{\alpha}}}\right)\Bigg) \\ &= \sum_i\Bigg(I_{y_i=0}\log(1-\pi_i)(1+\alpha\mu_i)^{-\alpha^{-1}} \\ &\quad + I_{y_i>0}\Bigg(\log(1-\pi_i)-\frac{1}{\alpha}\log(1+\alpha\mu_i) \\ &\quad - y_i\log\left(1+\frac{1}{\alpha\mu_i}\right)+\log\Gamma\left(y_i+\frac{1}{k}\right) \\ &\quad -\log\Gamma\left(\frac{1}{\alpha}\right)-\log y_i!\Bigg)\Bigg) \end{aligned} \tag{12}$$

Eşitlik 12'de, $I_{(.)}$ tesadüfi bir indikatör fonksiyonudur. Lambert (1992) tarafından verilen model tanımlaması,

$$\log(\mu) = X\beta \tag{13}$$

ve

$$\log\left(\frac{\pi}{1-\pi}\right) = G\gamma \tag{14}$$

biçiminde yazılabilir. Eşitlik 13 ve 14'de verilen X ve G kovariet matrisleri β ve γ sırasıyla, (p+1)x1 ve (q+1)x1 boyutlu bilinmeyen parametre vektörleridir. β, α ve γ için ML tahminleri EM algoritması kullanılarak elde edilebilir.

ZINB için EM algoritması: $Z_i \approx \text{Bernoulli}(w_i)$ ile gösterilen indikatör tesadüfi değişkeni olsun ve aşağıdaki biçimde verilsin (Jansakul, 2005).

$$Z_i = \begin{pmatrix} 1, & \text{sıfır} \\ 0 & \text{Poisson} \end{pmatrix} \text{ ve } Z_i = \begin{pmatrix} (Y_i/Z_i=1), & \equiv 0 \\ (Y_i/Z_i=0), & \approx \text{NB}(\mu_i,\alpha) \end{pmatrix}$$

olup log olabilirlik fonksiyonu,

$$\begin{aligned} L &= \sum_i\left[z_i\log\left(\frac{\pi_i}{1-\pi_i}\right)+\log(1-\pi_i)\right]+\sum_i(1-z_i) \\ &= \left[y_i\log\alpha-\left(y_i+\alpha^{-1}\right)\log(1+\alpha\mu_i)+d\log(y_i,\alpha^{-1})-\log y_i!\right] \\ &= L(w,z)+L(\lambda,\alpha;y,z) \end{aligned} \tag{15}$$

biçiminde yazılabilir (Hall, 2000; Ridout ve ark., 2001; Jansakul, 2005). Daha sonra EM algoritması kullanılarak bilinmeyen parametrelerin tahminleri elde edilir.

E-aşaması: E aşamasında, gözlenmiş veriler verilmişken z_i'nin koşullu bekleneni Bayes teoremi kullanılarak aşağıdaki gibi elde edilir.

$$\hat{z}_i = E(Z_i/Y_i,\lambda_i,\alpha,\pi_i)$$

$$= \begin{cases} \dfrac{\pi_i}{\pi_i + (1-\pi_i)(1+\alpha\mu_i)^{-\alpha^{-1}}} & y_i = 0 \\ 0 & y_i > 0 \end{cases} \quad (16)$$

M-aşaması: β parametreleri ağırlıklandırılmış NB regresyon model kullanılarak, tüm veriler için eşitlik 15'de verilen log olabilirlik fonksiyonundaki $l(\lambda,\alpha;y,z)$'nın, maksimize edilmesiyle tahminlenebilir. γ parametresi, $\hat{z}$ cevaplı ağırlıklandırılmış lojistik regresyon kullanılarak, eşitlik 15'de verilen $l(w,z)$'nin maksimize edilmesiyle tahminlenebilir. E ve M aşamaları yakınsama ölçütü (10^{-6}) elde edilinceye kadar devam edilir (Lambert, 1992). Yakınsama ölçütü olarak Newton-Raphson algoritması kullanılmaktadır.

Poisson hurdle model: Sınırlandırılmış sayıma dayalı olarak elde edilen pozitif gözlem değerleri ($y_i > 0$) Poisson dağılımı kullanılarak modellenmesi, Poisson hurdle model olarak adlandırılmaktadır. y_i, i=1,2,....,n birbirinden bağımsız sayıma dayalı olarak elde edilen gözlem değerleri olsun. $y_i = 0$ olma olasılığı $1-p(x)$ ve $y_i \approx$ sınırlandırılmış Poisson$(\lambda(z))$ olma olasılığı $p(x)$ olsun. Burada x ve z kovariet matrisleridir. Poisson hurdle model (Dalrymple ve ark., 2003),

$$P(y_i = 0/x) = 1 - p(x)$$

$$P(y_i = q/x,z) = \frac{p(x)\exp(-\lambda(z))\lambda(z)^q}{q!(1-\exp(-\lambda(z)))} \quad q = 1,2,... \quad (17)$$

biçiminde yazılabilir. Eşitlik 17'de verilen $p(x)$ ve $\lambda(z)$ sırasıyla logit ve log-doğrusal fonksiyonları ile modellenmektedirler. Yani,

$$\log(\lambda(z)) = x_i'\beta \quad (18)$$

$$\text{logit}(p_i) = z_i'\alpha \quad (19)$$

biçiminde modellenmektedirler. Eşitlik 18 ve eşitlik 19'da verilen β ve α sırasıyla bilinmeyen parametre vektörleridir. β, α parametrelerinin tahmin edilmesinde ML yöntemi kullanılmaktadır. Poisson hurdle için log olabilirlik aşağıdaki gibi yazılabilir.

$$\begin{aligned} L = & \sum_{y_i>0} x_i\beta - \sum_{i=1}^{n}\log(1+\exp(x_i\beta)) \\ & + \sum_{y_i>0}\left[y_i z_i\alpha - \exp(z_i\alpha)\right. \\ & \left. - \log(1-\exp(-\exp(z_i\alpha))) - \log(y_i!)\right] \\ = & L(\beta) + L(\alpha) \end{aligned} \quad (20)$$

Eşitlik 20'de verilen ve lojistik modeli esas alan $L(\beta)$ olabilirliği, bilinen genelleştirilmiş doğrusal model (generalized linear model) kullanılarak uyumu yapılabilir. Bununla birlikte $L(\alpha)$ olabilirlik fonksiyonun maksimize edilmesiyle α bilinmeyen parametre vektörünün en yüksek olabilirlik tahmini elde edilmektedir.

Negatif binomial hurdle model: Negatif binomial hurdle'da, sayıma dayalı olarak elde edilen bağımlı değişkenin sıfır ya da sıfır değerli olmama sonuçlarını belirleyen binomial olasılık modeli ile pozitif sonuçları tanımlayan sınırlandırılmış sayıma dayalı modeli için verilen log olabilirlik fonksiyonu aşağıdaki gibi yazılabilir (Long ve Freese, 2006; Hilbe, 2007),

$$L = \ln(f(0)) + \left\{\ln\left[1 - f(0)\right] + \ln P(j)\right\} \quad (21)$$

Eşitlik 21'de verilen $f(0)$ modelin binary kısmının olasılığını göstermektedir. $P(j)$ pozitif sayımın olasılığını göstermektedir. Logit model kullanıldığı durumda, sıfır sayımın olasılığı,

$$f(0) = P(y = 0;x) = 1/1 + \exp(xb1))$$

ve

$1 - f(0)$ ise,

$$\exp(xb1)/(1 + \exp(xb1))$$

biçiminde yazılabilir. Böylece negatif binomial hurdle modelin her iki kısmı için log olabilirlik fonksiyonu aşağıdaki gibi yazılabilir.

$$\begin{aligned} & \text{cond}\{y = 0, \ln(1/1 - \exp(xb1)), \\ & \ln(\exp(xb1)/(1 + \exp(xb1))) + y * \ln(\exp(xb)/(1 + \exp(xb))) \\ & -\ln(1 + \exp(xb))/\alpha + \ln\Gamma(y + 1/\alpha) - \ln\Gamma(y + 1) - \ln\Gamma(1/\alpha) \\ & -\ln(1 - (1 + \exp(xb))(-1/\alpha)\} \end{aligned} \quad (22)$$

Model seçimi: Akaiki bilgi ölçütü (AIC) ve Bayesian bilgi ölçütü (BIC) model uyumu için kullanılan uyum ölçütleridir. Birçok Monte-Carlo simülasyonu BIC, AIC uyum ölçütlerinin birlikte kullanılması gerektiğini göstermektedir (Dalrymple ve ark., 2003; Yeşilova ve ark., 2007). En küçük uyum ölçütlerine sahip model, en iyi model olarak kabul edilir. Genel olarak;

$$AIC = -2\log L + 2r \quad (23)$$

ve

$$BIC = -2\log L + r\ln(n) \quad (24)$$

biçiminde tanımlanır. Eşitlik 23 ve eşitlik 24'de, logL karışımlı Poisson regresyon modelinde iterasyon bittikten sonra elde edilen log olabilirlik değerini, *r* parametre sayısını ve n örnek büyüklüğünü göstermektedir.

Sonuç: Sıfır değer ağırlıklı verilerde, gözlem değerlerinin çoğunun sıfır olması verilerin dağılımının sağa doğru çarpık olmasına neden olmaktadır. Bu tip verilere uygulanan dönüşümler, özellikle sürekli verilerde, çoğunlukla yetersiz kalmaktadır. Veri kümesinde atılan gözlem değerleri ise, bilgi kaybına neden olabilir. Sıfır değer ağırlıklı veriler her hangi bir işleme (dönüşüm, aşırı uç değerlerin atılması gibi) tabi tutulmadan, verilerin orijinal yapısı esas alan yöntemler uygulanabilir. Bu çalışmada, sıfır değerlerinin çok olduğu sayıma dayalı olarak elde edilen verilerin analizinde, yaygın olarak kullanılan PR, NB, ZIP, ZINB ve hurdle yaklaşımlarının teorik özellikleri incelenmiştir. Özellikle ZIP, ZINB, hurdle regresyonları, sıfır değerlerini farklı bir şekilde modellediklerinden dolayı yoğun olarak kullanılmaktadırlar. Sıfır değer ağırlıklı verilerin analizi ile ilgili yapılan

çalışmalarda, ZINNB ve negatif binomial hurdle regresyonların, diğer regresyonlara göre, daha iyi sonuç verdiği saptanmıştır (Hall, 2000; Ridout ve ark., 2001; Jansakul, 2005; Long ve Freese, 2006; Yeşilova ve ark., 2007).

Kaynaklar

Agresti, A., 1997. Categorical Data Analysis. New Jersey, Canada; John and Wiley & Sons, Incorporation.

Böhning, D., 1994. A Note on a Test for Poisson Overdispersion. Biometrika, 81, 418-419.

Böhning, D., 1998. Zero- Inflated Poisson Models and C.A.MAN: A Tutorial Collection of Evidence. Biometrical Journal, 40(7), 833-843.

Böhning, D., Dietz, E., Schlattmann, P., 1999. The Zero-Inflated Poisson Model and the Decayed, Missing and Filled Teeth Index in Dental Epidemiology. Journal of Royal Statistical Society, A, 162, 195–209.

Breslow, N., 1990. Tests of Hypotheses in Overdispersed Poisson Regression and Other Quasi-Likelihood Models. Journal of American Statistical Association, 85(410), 565-571.

Cameron, A.C., Trivedi, P.K., 1998. Regression Analysis of Count Data. New York: Cambridge University Pres.

Cheung, Y.B., 2002. Zero-Inflated Models for Regression Analysis of Count Data: A Study of Growth and Development. Statistics in Medicine, 21, 1461-1469.

Cox, R., 1983. Some Remarks on Overdispersion. Biometrika,70, 269-274.

Dalrymple, M.L., Hudson, I.L., Ford, R.P.K., 2003. Finite Mixture, Zero-Inflated Poisson and Hurdle Models with Application to SIDS. Computational Statistics & Data Analysis 41, 491-504.

Frome, E.D., Kutner, M.H., Beauchamp, J.J., 1973. Regression Analysis of Poisson- Distributed Data. Journal of American Statistical Association, 68(344), 935-940.

Hall, D.A., 2000. Zero-Inflated Poisson and Negative Binomial Regression with Random Effects: A Case Study. Biometrics, 56, 1030-1039.

Hilbe, J.M., 2007. Negative Binomial Regression.Cambridge,UK.

Khoshgoftaar, T.M., Gao, K., Szabo, R.M., (2005). Comparing Software Fault Predictions of Pure and Zero- inflated Poisson Regression Models. International Journal of Systems Science, Vol. 36, No. 11, 2005, p 707-715.

Lambert, D., 1992. Zero-Inflated Poisson Regression, with an Application to Defects in Mnaufacturin. Technometrics, 34(1), 1-13.

Lawles, J.F., 1987. Negative Binomial and Mixed Poisson Regression. The Canadian Journal of Statistcs, 15(3), 209-225.

Lee, A.H., Wang, K., 2001. Analysis of Zero-Inflated Poisson Data Incorporating Extent of Exposure. Biometrical Journal, 43(8), 963-975.

Long, J.S., Freese, J., (2006). Regression Models for Categorical Depentent Variable Using Stata. A Stata Pres Publication, StataCorp LD Collage Station, Texas, USA.

Jansakul, N., (2005). Fitting a Zero-inflated Negative Binomial Model via R. In Proceedings 20th International Workshop on Statistical Modelling. Sidney, Australia, p. 277-284.

Martin, S.W., Rose, C.E, Wannemuehler, K.A., Plikaytis, B.D., (2006). On the of Zero-inflated and Hurdle Models for Medelling Vaccine Adverse event Count Data. Journal of Biopharmaceutical Statistics, 16: 463-481.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models. Second Edition, London, UK, Chapmann and Hall.

Nelder, J.A., Wedderburn, R.W.M., 1972. Generalized Linear Models. Journal of Royal Statistical Society A, 135(3), 370-384.

Ridout, M., Hinde, J., Demetrio, C.G.B., 2001. A Score Test for a Zero-Inflated Poisson Regression Model Against Zero-Inflated Negative Binomial Alteratves. Biometrics, 57, 219-233.

SAS., 2007. SAS/Stat. Software. Hangen and Enhanced, USA: SAS, Institute. Incorporation.

Stokes, M.E., Davis, C.S., Koch, G.G., 2000. Categorical Data Analysis Using the SAS System. USA; John and Wiley & Sons, Incorporation.

Yau, K.K.W., Lee, A.H., 2001. Zero-Inflated Poisson Regression with Random Effects to Evaluate an Occupational Injury Prevention Programme. Statistics in Medicine, 20, 2907-2920.

Yeşilova, A., Kaki, B., Kasap, İ., 2007. Sıfır Değer Ağırlıklı Sayıma Dayalı Olarak Elde edilen Bağımlı Değişkenin Modellenmesinde Kullanılan Regresyon Yöntemler. İstatistik Araştırma Dergisi, 5(1), 1-9.