Y. Y. Ü. Fen Bilimleri Enstitüsü Dergisi
Yıl : 2008, Cilt : 13, Sayı : 1, Sayfa : 19-23

Kategorik Veriler için Karışımlı Poisson ve Karışımlı Lojistik Regresyon Yöntemlerin Teorik Özelliklerinin İncelenmesi

Abdullah YEŞİLOVA[1] Hayrettin OKUT[1] Barış KAKİ[1]

[1] Yüzüncü Yıl Üniversitesi, Ziraat Fakültesi, Zootekni Bölümü, 65080, VAN

Özet: Karışımlı model yaklaşımı veri kümesinin heterojen bir yapı gösterdiğini varsaymaktadır. Söz konusu heterojenlikten dolayı veri kümesinde ekstra-varyasyon meydana gelmektedir. Veri kümesi kendi içerisinde homojen alt populasyonlara bölünerek, heterojenlik giderilmektedir. Böylece alt populasyonlar içi homojenlik sağlanırken, alt populasyonlar arası heterojenlik ortaya konmaya çalışılır. Kategorik veriler için Karışımlı Poisson ve lojistik regresyon modelleri kullanılmaktadır. Karışımlı Poisson regresyon sayıma dayalı olarak elde edilen verilerin analizinde, karışımlı lojistik regresyon ise binary, sıralayıcı ve sınıflayıcı verilerin analizinde kullanılmaktadır. Karışımlı Poisson ve lojistik regresyon modellerinde parametre tahminleri, EM algoritması kullanılarak en yüksek olabilirlik yöntemi ile elde edilmektedir. EM algoritmasının, E aşamasında bilinmeyen alt populasyon sayısı eksik gözlem olarak kabul edilip, bunların sayısı tahmin edilir. M aşamasında ise olabilirlik fonksiyonu maksimize edilip bilinmeyen parametreler en yüksek olabilirlik yöntemi ile elde edilir.

Anahtar kelimeler: Aşırı yayılım, EM algoritması, en yüksek olabilirlik, sayıma dayalı veriler

Investigation of Theoretical Properties of Mixture Poisson and Logistic Regressions for Categorical Data

Abstract: Mixture model approach assumes that data set has heterogeneous variation. Extra variation occurs in data set due to this mentioned heterogeneity. The heterogeneity is solved by dividing the data set into homogeneous sub-populations. By doing this, homogeneity is obtained in sub-populations and heterogeneity is situated among sub-populations. Mixture Poisson and logistic regression models are used for categorical data. Mixture Poisson regression is used for the analysis of based count data, whereas mixture logistic regression is used for the analysis of binary, ordinal and nominal data. Parameter estimations of Mixture Poisson and mixture logistic regression models are obtained by maximum likelihood method using expectation and maximization (EM) algorithms. In E step of EM algorithm, the number of unknown sub-populations is considered as missing observation, and their numbers are estimated. In M step, maximum likelihood estimations of unknown parameters are obtained by maximizing log-likelihood function.

Key Words: Count data, EM algorithm, maximum likelihood, overdispersion

Giriş

Karışımlı modellemede (mixture modelling), veri kümesinin tek bir populasyondan değil birden fazla alt populasyondan elde edilmiş heterojen bir yapı gösterdiği varsayılır. Başka bir deyişle veri kümesi, heterojen bir populasyon özelliğini göstermektedir. (Breslow, 1992; Wang ve ark., 1996; Muthen ve Muthen., 2002). Böylece veri kümesi kendi içerisinde homojen alt populasyonlara ayrılarak, söz konusu heterojenlik dikkate alınıyor. Karışımlı modellerde amaç, gözlenen değişkenlerin kaç alt populasyona ait olduğunun belirlenmesi ve hangi gözlem değerinin hangi alt populasyonda bulunması gerektiğine karar verilmesidir. Alt populasyonlara, gözlenemeyen sınıflar (latent class) adı verilmektedir (Wang ve ark., 1998; Muthen ve Muthén, 2002; Okut ve ark., 2002; Yeşilova ve Atlıhan, 2007). Dolayısıyla bütün değişkenler için tek bir parametre tahmini yerine, her alt populasyon için ayrı parametre tahmini yapılmaktadır (Yeşilova, 2003). Yani her alt populasyon için parametre tahmin değerleri değişmektedir. Sınırlı sayıda alt populasyondan oluştuğu varsayılan veri kümelerinin modellenmesinde sonlu karışımlı (Finite mixture model) modeller kullanılır.

Karışımlı modeller bütün dağılışlarda kullanılabilmekte ve Poisson dağılışında kullanılması durumunda, karışımlı Poisson model olarak adlandırılmaktadır. Karışık Poisson regresyon modeli olan karışımlı Poisson regresyon (Mixture Poisson Regression=MPR), Poisson dağılımı gösteren bağımlı değişkende aşırı yayılım olduğu durumlarda kullanılmaktadır. Aşırı yayılım Poisson dağılımının, varyansının ortalamasından büyük olması olarak tanımlanmaktadır. MPR'de, aşırı yayılım genellikle gözlenemeyen heterojenliğin (latent heterogeneity) neden olduğu bir durumdur.

Kategorik verilerin analizinde kullanılan karışımlı lojistik regresyon yöntemi (Mixture Logistic regression=MLR), veri kümesinde ekstra binomiyal varyasyon olduğu durumlarda kullanılmaktadır (Wang ve ark., 1996; Wang ve ark., 1998; Wang ve Putterman, 1998; Yeşilova, 2003). Lojistik regresyonda, gözlenen varyansının beklenen varyansdan büyük olması aşırı yayılım ya da ekstra-binomiyal varyasyon olarak tanımlanmaktadır (Dean, 1992).

MPR ve MLR yöntemlerinde parametre tahminleri, EM (Expectation Maximization) ve QN (Quasi-Newton) algoritmalarını esas alan tahminleme yöntemleri kullanılarak elde edilmektedir. EM yaklaşımında, E ve M aşamalarını kullanarak, bilinmeyen parametre vektörünün en yüksek olabilirlik (Maximum likelihood=ML) tahminleri elde edilmektedir. E aşamasında, gözlenmiş veriler üzerinde koşullu beklenen değerler kullanılarak eksik verilerin tahmini yapılmaktadır. Burada eksik veriler, gözlenemeyen (latent) sınıflardır. M aşamasında, parametre tahminleri, log olabilirlik fonksiyonun beklenen değerinin maksimize edilmesi ile elde edilmektedir (Lambert, 1992; Wang ve ark., 1996; Jansen, 1993; Dalrymple ve ark., 2003; Yeşilova ve Atlıhan, 2007).

Bu çalışmada, karışımlı Poisson ve karışımlı lojistik regresyon yöntemlerinin teorik özelliklerinin incelenmesi

amaçlanmıştır. İlk olarak her iki regresyona ilişkin model tanımlanması ve log olabilirlik fonksiyonlarının nasıl elde edildiği verilmiştir. İkinci aşamada, regresyon modellerine ait log olabilirlik fonksiyonları üzerinden, EM algoritması kullanılarak, bilinmeyen parametrelerin ML tahminlerinin elde edilmesi incelenmiştir.

Karışımlı Poisson regresyon modeli: Poisson regresyonunda model,

$$E(Y_i \mid X_i) = \lambda_i = \exp(X_i\beta), i = 1,2,... \quad (1)$$

biçiminde yazılmaktadır (Yeşilova, 2003; SAS,2007; Yeşilova ve Atlıhan, 2007). Eşitlik 1'de, Poisson ortalaması (λ_i) ile ortak değişkenler arasındaki ilişki, bir bağlantı (link) fonksiyonu ile verilmiştir. Karışımlı model için kesikli karışım dağılışı,

$$p(y) = \sum_{k=1}^{K} P(y/v_k \exp(\beta' x))\pi_k \quad (2)$$

biçiminde yazılabilir (Wang ve ark., 1996; Wang ve ark., 1998). Eşitlik 2'de, β regresyon katsayılarına ait vektörü, v, gamma dağılımlı tesadüfi etkiyi gösterirken, y_i gözlem değerleri ise $v_k \exp(\beta' x)$ ortalamalı Poisson dağılımından elde edilmektedir. Bununla birlikte Eşitlik 2'de K alt populasyon sayısını ve π_k, k'ıncı alt populasyonun karışma olasılığını gösterir (Wang ve ark., 1996; Wang ve ark., 1998; Yeşilova, 2003).

$y_i\{y_1,...,y_n\}$ Poisson dağılımına sahip gözlem değerlerinden oluşan veri kümesi homojen tek bir populasyonu temsil etmeyip birden fazla alt populasyondan oluşan heterojen bir veri kümesi olabilir. Başka bir ifade ile veri kümesinde gözlenemeyen alt populasyonlar bulunabilir. Bu durumda y, değerlerine ilişkin marjinal olasılık fonksiyonu (Okut ve ark., 2002; Yeşilova, 2003),

$$f(y) = \sum_{k=1}^{K} P(C = k)P(Y = y \mid C = k) = \sum_{k=1}^{K} \pi_k f(y, \lambda_k)$$

şeklinde yazılabilir. Poisson dağılışlı veri setinin, K kadar alt populasyona ait heterojen bir örnek olması durumunda k'ıncı alt populasyona giren i'inci şans değişkeninin olasılığı,

$$\pi_{ik} = P(c_{ik} = k)$$

biçiminde verilebilir. Bu durumda,

$$\sum_{k=1}^{K} \pi_{ik} = 1$$

olmaktadır. Bütün veriler için log olabilirlik fonksiyonu,

$$L = \sum_{i=1}^{n}\sum_{k=1}^{K} c_{ik} \log \pi_k + \sum_{i=1}^{n}\sum_{k=1}^{K} c_{ik} \log Po(y_i/\lambda_i) \quad (3)$$

biçiminde verilebilir (Dempster ve ark., 1977; Wang ve ark., 1996; Wang ve ark., 1998). Eşitlik 3'te, Po, Poisson dağılımını, C gözlenemeyen gözlemler (alt populasyon sayısı) olup,

$$C = \{c_{ik},\ i = 1,2,...,n,\ k = 1,2,...,K\}$$

Burada c_{ik},

$$\begin{cases} c_{ik} = 1, & c_{ik} \in K \\ c_{ik} = 0, & \text{diğer durumlarda} \end{cases}$$

olarak verilebilir.

Karışımlı Poisson regresyonda aşırı yayılım: Karışımlı Poisson regresyonun ortalaması ve varyansı (Wang ve ark., 1996; Wang ve ark., 1998),

$$E(y_i) = E(E(y_i \mid K = k) = \sum_{k=1}^{K} \pi_k \lambda_k \quad (4)$$

$$Var(Y_i) = E\{Var(Y_i \mid K = k)\} + Var\{E(Y_i \mid K = k)\} \quad (5)$$
$$= \sum_{k=1}^{K} \pi_k \lambda_k \left\{ \sum_{k=1}^{K} \pi_k \lambda_k^2 - \left(\sum_{k=1}^{K} \pi_k \lambda_k \right)^2 \right\}$$

biçiminde verilebilir. 4 ve 5 numaralı eşitliklerde varyans ortalamadan büyük olduğundan dolayı aşırı yayılım söz konusu olur. Veri setinin heterojen olmadığı veya aşırı yayılım göstermediği durumlarda, ortalama ile varyans arasındaki ilişki,

$$E(Y_i) = Var(Y_i)$$

ve 5 numaralı eşitlikteki varyans formülünde,

$$Var(Y_i \mid K = k) = 0$$

olur. Böylece ortalama ile varyans arasındaki eşitlik sağlanmış olur.

Karışımlı Poisson regresyon için EM algoritması ve en yüksek olabilirlik yöntemi: Karışımlı Poisson regresyon modeli için EM algoritmasının aşamaları aşağıdaki gibi verilebilir (Wang ve ark., 1996; Wang ve ark., 1998; Dalrymple ve ark., 2003).

Birinci aşamada, $\beta^{(0)}$ ve $\pi_k^{(0)}$ başlangıç değerleri ε ve ε_0 tolerans değerlerine göre belirlenir.

E aşamasında, $\beta^{(0)}$ ve $\pi^{(0)}$ başlangıç değerleri verildiğinde gözlenmiş veriler (X, Y) ve parametrelerin başlangıç değerleri üzerinden, C eksik gözlemleri elde edilir. $\hat{C}_{ik}(\beta^{(0)}, \pi^{(0)})$ kullanılarak c_i 'nin k'ıncı unsurunun koşullu olasılığı,

$$\hat{c}_{i,k} = \left(\beta^{(0)}, \pi_k^{(0)}\right) = \frac{\pi_k f_k\left(y_i/x_i, \beta_k^{(0)}\right)}{\sum_{k=1}^{K} \pi_k f_k\left(y_i/x_i, \pi_k^{(0)}\right)},\ k = 1,2,....K \quad (6)$$

biçiminde verilebilir.

M aşamasında,

$$\left\{c_i\left(\beta^{(0)}, \pi_k^{(0)}\right) = \left(z_{i,1},...,z_{i,K}\right)';\ i = 1,2,...,n\right.$$

koşullu olasılıkları verilmişken, parametre tahminleri, eşitlik 3'te verilen log olabilirlik fonksiyonun β ve π'ye göre maksimize edilmesi ile,

$$Q = \left(\beta^{(1)}.p \mid \beta^{(0)}, \pi_k^{(0)}\right) = E\left\{(L(Y.C,\beta,p,X))/Y, X.\beta^{(0)}, \pi_k^{(0)}\right\} \quad (7)$$

$$Q = Q_1 + Q_2$$

ve buradan, Q_1 ve Q_2,

$$Q_1 = \sum_{i=1}^{n}\sum_{k=1}^{K} c_{i,k}\left(\beta^{(0)}, \pi_k^{(0)}\right)\log(\pi_k) \quad (8)$$

$$Q_2 = \sum_{i=1}^{n}\sum_{k=1}^{K} c_{i,k}\left(\beta^{(0)}, \pi_k^{(0)}\right)\log(y_i/\lambda_k) \quad (9)$$

elde edilir. Eşitlik 8 ve 9'da verilen $\hat{\beta}$ ve $\hat{\pi}$ tahmin edicileri, Q_1 ve Q_2 eşitliklerinin π ve β'ya göre türevlerinin alınası ile,

$$\frac{\partial Q_1}{\partial \pi_K} = \sum_{i=1}^{n}\left(\frac{\hat{c}_{i,k}}{\hat{\pi}_K} - \frac{\hat{c}_{i,K}}{\hat{\pi}_K}\right) = 0\,,\ k = 1,\ldots,K-1 \quad (10)$$

$$\frac{\partial Q_2}{\partial \beta} = \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{c}_{i,k}\frac{\partial}{\partial c}P(y_i/\lambda_k) = 0 \quad (11)$$

biçiminde elde edilir. Eşitlik 10 kullanılarak $\hat{\pi}_k$,

$$\hat{\pi}_k = \frac{1}{n}\sum_{i=1}^{n}\hat{c}_{i,k}\ ,\ k = 1,\ldots,K-1 \quad (12)$$

biçiminde elde edilmektedir (Wang ve Putterman., 1998; Wang ve ark., 1996; Wang ve ark., 1998). Yukarıda verilen eşitlik 11'de kapalı formunun çözümünün zor olmasından dolayı, parametre tahminleri için Quasi-Newton yaklaşımı kullanılarak E ve M aşamaları,

1. aşamada, $\beta^{(0)} = \left(\beta_1^{(0)},\ldots,\beta_k^{(0)}\right)$ ve $\pi^{(0)} = \left(\pi_1^{(0)},\ldots,\pi_k^{(0)}\right)$ başlangıç değerlerinin ε ve ε_0 tolerans değerlerine göre belirlenmesi,

2. aşamada (E- aşaması), eşitlik 6 kullanılarak

$$\hat{c}_i = \left(\hat{c}_{i,1},\ldots,\hat{c}_{i,K}\right)^{-1}\ i = 1,2,\ldots,n \quad (13)$$

değerleri hesaplanılır. $\hat{c}_{i,k}$'nın hesaplanmasında aşırı taşmayı engellemek için eşitlik 6'da verilen fonksiyonun pay paydası, payda toplamının en büyük değerine bölünür.

3. aşamada (M- aşaması),

a) Eşitlik 10 kullanılarak $\hat{\pi}$ parametresinin hesaplanması

b) yarı- Newton algoritması kullanılarak 11 numaralı eşitliğin çözümünden $\hat{\beta}$ parametresinin hesaplanması.

4. aşamada, aşağıdaki koşullardan en az biri doğru ise,

$\beta^{(0)} = \hat{\beta}$, $\pi^{(0)} = \hat{\pi}$ olur ve 1. asamaya gidilir, aksi durumda c'ye gidilir.

1) $\left\|\hat{\beta} - \beta^{(0)}\right\| \geq \varepsilon$ (14)

2) $\left\|\hat{\pi} - \pi^{(0)}\right\| \geq \varepsilon$ (15)

3) $\left|L\left(Y,X,\hat{\beta},\hat{\pi}\right) - L\left(Y,X,\beta^{(0)},\pi^{(0)}\right)\right| \geq \varepsilon_0$ (16)

c) β parametreleri için, $\hat{\pi}$'ler başlangıç değerleri olarak alınır ve yarı-Newton algoritması kullanılarak gözlenmiş $L(Y,X,\beta,\pi)$ log olabilirlik fonksiyonu maksimize edilip, işlem sonlandırılır. (Wang ve ark., 1996,1998; Wang ve ark., 2001; Dalrymple ve ark., 2002).

Karışımlı lojistik regresyon modeli: Karışımlı lojistik model için kesikli karışım dağılımı (Wang ve Putterman, 1996),

$$P(y) = \sum_{k=1}^{K}\beta_i\left(y/v_k\exp(\beta' x)\right)\pi_k \quad (17)$$

biçiminde yazılabilir. Eşitlik 17'de π_k, k'ıncı alt populasyonun karışma olasılığı; y, cevap değişkeni; x, açıklayıcı değişken vektörü; β, bilinmeyen parametre vektörü; v, gamma dağılışına sahip tesadüfi bir etki olmaktadır. Y_i, binom dağılış gösterir ve,

$$P(Y_i = y_i/p_i) = \binom{n_i}{y_i}p_i^{y_i}(1-p_i)^{n_i - y_i} \quad (18)$$

biçiminde yazılır. Logit bağlantı fonksiyonu,

$$\text{logit}(\pi) = \beta' x$$

olarak yazılabilir (Frome ve ark., 1973; SAS,2007). Bütün veriler için log-olabilirlik fonksiyonu,

$$L(Y,X,\beta,\pi) = \sum_{i=1}^{n}\sum_{k=1}^{K} c_{ik}\log\pi_k + \sum_{i=1}^{n}\sum_{k=1}^{K} c_{ik}\log bi(y_i/\beta_k, x) \quad (19)$$

biçiminde yazılabilir. Eşitlik 19'da bi, binom dağılımını göstermektedir.

Karışımlı lojistik regresyonda aşırı yayılım: Karışımlı lojistik regresyon modeli için Y_i bağımsız gözlemlerin ortalaması ve varyansı sırası ile (Wang ve Putterman, 1996),

$$E(Y_i) = E(E(Y_i \mid \Pi_i)) = n_i = \left(\Sigma_{k=1}^{K} p_{ij}\pi_{ij}\right) \equiv n_i\hat{\pi}_i$$

$$\text{Var}(Y_i) = E(\text{Var}(Y_i \mid \Pi_i)) = n_i\hat{\pi}_i(1-\hat{\pi}_i) \quad (20)$$

biçiminde verilebilir. İkili bağımlı değişkenleri arasında korelasyon olması durumunda gözlenen varyans,

$$\begin{aligned}\text{Var}(Y_i) &= E(\text{Var}(Y_i \mid \Pi_i)) + \text{Var}(E(Y_i \mid \Pi_i)) \\ &= n_i\sum_{k=1}^{K} c_{ik}\pi_{ik}\left(1 - \sum_{k=1}^{K} c_{ik}\pi_{ik}\right) + ((n_i - 1)/n_i)\text{Var}(E(Y_i \mid \Pi_i)) \\ &= n_i\hat{\pi}_i(1-\hat{\pi}_i) + ((n_i - 1)/n_i)\text{Var}(E(Y_i \mid \Pi_i))\end{aligned} \quad (21)$$

biçiminde gösterilir ve

$$\text{Var}(E(Y_i \mid \Pi_i)) = n_i^2\left\{\sum_{i=1}^{k} c_{ik}\pi_{ik}^2 - \left(\sum_{i=1}^{k} c_{ik}\pi_{ik}\right)^2\right\}$$

biçiminde yazılabilir. Bu durumda Eşitlik 20'de verilen beklenen varyans ile Eşitlik 21'de verilen gözlenen

varyans birbirine eşit değildir. Gözlenen varyans beklenen varyanstan ya büyük ya da küçük olmaktadır. Gözlenen varyansın beklenen varyanstan büyük çıkması aşırı yayılım olarak adlandırılmaktadır.

Karışımlı lojistik regresyon modeli için EM algoritması ve en yüksek olabilirlik yöntemi: MLR modeli için EM algoritmasının aşamaları aşağıdaki gibi verilebilir (Wang ve Putterman, 1996).

Birinci aşamada, $\beta^{(0)}$ ve $\pi_k^{(0)}$ başlangıç değerleri ε ve ε_0 tolerans değerlerine göre belirlenir.
E aşamasında, $\beta^{(0)}$ ve $\pi^{(0)}$ başlangıç değerleri verildiğinde gözlenmiş veriler (X, Y) ve parametrelerin başlangıç değerleri üzerinden, C eksik gözlemleri elde edilir. $\hat{C}_{ik}(\beta^{(0)},\pi^{(0)})$ kullanılarak c_i'nin k'ıncı unsurunun koşullu olasılığı,

$$\hat{c}_{i,k}=\left(\beta^{(0)},\pi_k^{(0)}\right)=\frac{\pi_k bi\left(y_i/x_i,\beta_k^{(0)}\right)}{\sum_{k=1}^{K}\pi_k bi\left(y_i/x_i,\pi_k^{(0)}\right)},\ k=1,2,...,K \qquad (22)$$

biçiminde verilebilir.
M aşamasında,

$$\left\{c_i\left(\beta^{(0)},\pi_k^{(0)}\right)=\left(z_{i,1},...,z_{i,K}\right)';\ i=1,2,...,n\right.$$

koşullu olasılıkları verilmişken, parametre tahminleri, eşitlik 19'da verilen log olabilirlik fonksiyonun β ve π'ye göre maksimize edilmesi ile,

$$Q=\left(\beta^{(0)},p\,|\,\beta^{(0)},\pi_k^{(0)}\right)=E\left\{(L(Y,C,\beta,p,X))/Y,X,\beta^{(0)},\pi_k^{(0)}\right\} \qquad (23)$$

$$Q=Q_1+Q_2$$

ve buradan, Q_1 ve Q_2,

$$Q_1=\sum_{i=1}^{n}\sum_{k=1}^{K}c_{i,k}\left(\beta^{(0)},\pi_k^{(0)}\right)\log\left(\pi_k\right) \qquad (24)$$

$$Q_2=\sum_{i=1}^{n}\sum_{k=1}^{K}c_{i,k}\left(\beta^{(0)},\pi_k^{(0)}\right)\log bi(y_i/n_i,\pi_{ik}) \qquad (25)$$

elde edilir. Eşitlik 24 ve 25'de verilen $\hat{\beta}$ ve $\hat{\pi}$ tahmin edicileri, Q_1 ve Q_2 eşitliklerinin π ve β'ya göre türevlerinin alınası ile,

$$\frac{\partial Q_1}{\partial \pi_K}=0,\ k=1,....,K-1 \qquad (26)$$

$$\frac{\partial Q_2}{\partial \beta}=0 \qquad (27)$$

biçiminde elde edilir. Eşitlik 26 kullanılarak $\hat{\pi}_k$,

$$\hat{\pi}_k=\frac{1}{n}\sum_{i=1}^{n}\hat{c}_{i,k},\ k=1,....,K-1 \qquad (28)$$

biçiminde elde edilmektedir (Wang ve Putterman., 1998; Wang ve ark., 1996; Wang ve ark., 1998). Yukarıda verilen eşitlik 27'de kapalı formunun çözümünün zor olmasından dolayı, parametre tahminleri için Quasi-Newton yaklaşımı kullanılarak E ve M aşamaları,

1. aşamada, $\beta^{(0)}=\left(\beta_1^{(0)},...,\beta_k^{(0)}\right)$ ve $\pi^{(0)}=\left(\pi_1^{(0)},...,\pi_k^{(0)}\right)$ başlangıç değerlerinin ε ve ε_0 tolerans değerlerine göre belirlenmesi,

2. aşamada (E- aşaması), eşitlik 22 kullanılarak

$$\hat{c}_i=\left(\hat{c}_{i,1},...,\hat{c}_{i,K}\right)^{-1}\ i=1,2,...,n \qquad (29)$$

değerleri hesaplanılır. $\hat{c}_{i,k}$'nın hesaplanmasında aşırı taşmayı engellemek için eşitlik 22'de verilen fonksiyonun pay paydası, payda toplamının en büyük değerine bölünür.

3. aşamada (M- aşaması),

a) Eşitlik 26 kullanılarak $\hat{\pi}$ parametresinin hesaplanması

b) yarı- Newton algoritması kullanılarak 27 numaralı eşitliğin çözümünden $\hat{\beta}$ parametresinin hesaplanması.

4. aşamada, aşağıdaki koşullardan en az biri doğru ise,

$\beta^{(0)}=\hat{\beta}$, $\pi^{(0)}=\hat{\pi}$ olur ve 1. asamaya gidilir, aksi durumda c'ye gidilir.

$$1)\ \left\|\hat{\beta}-\beta^{(0)}\right\|\geq\varepsilon \qquad (30)$$

$$2)\ \left\|\hat{\pi}-\pi^{(0)}\right\|\geq\varepsilon \qquad (31)$$

$$3)\ \left|L\left(Y,X,\hat{\beta},\hat{\pi}\right)-L\left(Y,X,\beta^{(0)},\pi^{(0)}\right)\right|\geq\varepsilon_0 \qquad (32)$$

c) β parametreleri için, $\hat{\pi}$'ler başlangıç değerleri olarak alınır ve yarı-Newton algoritması kullanılarak gözlenmiş $L(Y,X,\beta,\pi)$ log olabilirlik fonksiyonu maksimize edilip, işlem sonlandırılır. (Wang ve ark., 1996,1998; Wang ve ark., 2001; Dalrymple ve ark., 2002).

Sonuç: Bağımlı değişkenin kategorik olduğu durumlarda, uygulanan regresyon yönteminin doğruluğu bakımından, aşırı yayılımın belirlenmesi büyük önem taşımaktadır. Bu bağlamda, veri kümesinde aşırı yayılım olduğu durumlarda karışımlı model yaklaşımı yaygın olarak kullanılmaktadır. Bu çalışmada, kategorik bağımlı değişkenin modellenmesinde kullanılan karışımlı Poisson ve karışımlı lojistik regresyon yöntemlerinin teorik özellikleri incelenmiştir. Her iki yöntemde, veri kümesinde oluşan aşırı yayılımı, veri kümesini kendi içerisinde homojen alt populasyonlara ayırarak gidermektedir.

Kaynaklar

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

Dalrymple., Hudson, I. L., Ford, R. P. K., 2002. Finite Mixture, Zero-Inflated Poisson and Hurdle Models with Application

to SIDS. University of Canterbury, Christchurch, New Zealand. 19.

Dean, C. B., 1992. Testing for Overdispersion in Poisson and Binomial Regression Models. Journal of American Statistical Association, 87(418):451-457.

Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum Likelihood from Incomplete Data via the EM Algrithm. Journal of Royal Statisticial Society, 39: 1-18.

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.

Jansen, R. C., 1993. Maximum Likelihood in a Generalized Linear Finite Mixture Model by Using the EM Algorithm. Biometrics, 49(1):227-231.

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

McCullagh, P., Nelder, J. A., 1989. Generalized Linear Models. Second Edition, Chapmann and Hall, London, 486.

Muthén, L. K., Muthén, B., 2002. Mplus: User's guide. Los Angeles, CA: Muthén & Muthén

Okut, H., Duncan, E. T., Duncan, C. S., Strycker, A. L., 2002. Latent Variable Mixture Modelling: Analyzing Mixture and the Structural Portion of Model. Joint Satatistical Meetings (JSM). 11-15, August, 2002 New York City.

SAS, 2007. SAS/STAT Software:Hangen and Enhanced. SAS, Inst. Inc., USA

Wang, P., Cockburn, I. M., Puterman, M. L., 1998. Analysis of Patent Data- Mixed Poisson Regression Model Approach. Journal of Business and Economic Statistics, 16(1):27-41.

Wang, P., Puterman, M. L., Cockburn, I. M., Le, N., 1996. Mixed Poisson Regression Models with Covariate Dependent Rates . Biometrics, 52:381-400.

Wang, P., Putterman, M. L., 1998. Mixed Logistic Regression Models. Journal of Agriculture, Biological and Environmental Statistics, 3(2):175-200.

Yeşilova, A. (2003). Biyolojik Çalışmalardan Elde Edilen Kategorik Verilere Karışık Poisson Regresyon Analizinin Uygulanması (Doktora Tezi, Basılmamış), Y.Y.Ü. Fen bilimleri Enstitüsü, Van.

Yeşilova, A., Atlıhan, R. (2007). Farklı Sıcaklıkların Scymnus Subvillosus'un Bıraktığı Yumurta Sayıları Üzerine Etkilerinin Karışımlı Poisson Regresyon İle Analiz Edilmesi. Y. Y. Ü. Tarım Bilimleri Dergisi.