IAAOJ, Scientific Science, 2013, 1(2), 3-7

AŞIRI YAYILIMLI VERİLER İÇİN GENELLEŞTİRİLMİŞ POİSSON KARMA MODELLERİN HAVA KİRLİLİĞİ ÜZERİNE BİR UYGULAMASI

Haydar KOÇ[1], M. Ali CENGİZ[1], Tuba KOÇ[1], Emre DÜNDER[1]

[1] Ondokuz Mayıs Üniversitesi, Fen-Edebiyat Fak., İstatistik Bölümü, Samsun, Türkiye
e-posta: haydarkoc@omu.edu.tr

ÖZET

Sayım verisi analizi içeren birçok çalışmada, hava kirliliği nedeniyle hastaneye başvuranların sayısı gibi, Poisson regresyon analizi kullanılan en temel tekniktir. Bununla birlikte regresyon modellerinde aşırı yayılım çok sık görülen bir durumdur. Poisson regresyon modelinde aşırı yayılım bağımlı değişkeninin varyansı ortalamasından büyük olduğu durumlarda ortaya çıkmaktadır. Bu durum tahminlerin standart hatalarının olduğundan daha düşük çıkmasına neden olmaktadır. Genelleştirilmiş Poisson Regresyon (GPR) model aşırı yayılım problemini giderebilmek için sayım verilerinin modellenmesinde kullanılır. Bu çalışmada aşırı yayılımlı bir sayım verisini modellemek için gerçek bir veri seti kullanılarak GPR model üzerinde duruldu.

Anahtar kelimeler: Genelleştirilmiş Poisson model, sayım verisi, aşırı yayılım

APPLICATION OF GENERALIZED POISSON MIXED MODEL FOR OVERDISPERSED COUNT DATA ON AIR POLLUTION

ABSTRACT

Many studies often involve the analysis of count data, such as the number of hospitalizations caused by air pollution, where Poisson Regression (PR) is the Standard basic technique. However, overdispersion is widely seen in this regression model. Overdispersion in this model occurs when the response variance is greater than the mean. This may cause standard errors of the estimates to be deflated or underestimated. The Generalized Poisson Regression (GPR) model is used to model dispersed count data to handle the overdispersion problem. In this study, we focus on Generalized Poisson Mixed Model for Overdispersed Count Data with real data set.

Keywords: Generalized Poisson model, count data, overdispersion.

1. Giriş

Regresyon modelleri bir değişken seti ile bir bağımlı değişken arasındaki ilişkiyi modellemek için kullanılan en yaygın araçlardır. Birçok uygulamada, ilgilenilen açıklayıcı değişken negatif değerler almayan tamsayılardan oluşan bir sayım verisidir. Sayım verisi için kullanılan en yaygın regresyon modeli Poisson regresyon modelidir. Sayım verisi olan bir bağımlı değişkenin beklenen değerini açıklayıcı değişkenlerin bir fonksiyonu olarak ifade eden Poisson regresyon modeli, Poisson dağılımından türetilir. Poisson regresyon modelinin en önemli özelliği bağımlı değişkenin beklenen değer ve varyansının eşit olduğu anlamına gelen eşit yayılım özelliğidir. Pratikte ortalama ve varsayın eşitliği nadiren gerçekleşen bir durumdur. Eğer Poisson modelde varyans ortalamadan daha büyükse yani beklenenden daha fazla bir değişkenlik varsa bu duruma aşırı yayılım adı verilir. Aşırı yayılım varsa ve bu durum hesaba katılmadan parametre tahminleri yapılırsa bu durum parametre tahminlerinin standart hatalarının olduğundan daha düşük çıkmasına neden olur ve bunun sonucunda model için açıklayıcı değişkenlerin seçiminde yanlışlık yapılmış olur.

Aşırı yayılımı hesaba katmanın bir yolu, Poisson dağılımından daha fazla yayılım gösteren bir olasılık dağılımı türetmektedir. Bir Poisson sürecinde, Poisson parametresinin gamma dağıldığı varsayıldığında Negatif binom dağılımı elde edilir ve sonuçta elde edilen dağılım Poisson'a göre aşırı yayılımlıdır. Joe ve Zhu (2005) Genelleştirilmiş Poisson dağılımının (GP) Poisson karışımı olarak ele

alınabileceğini ve dolayısıyla negatif binom (NB) dağılımına bir alternatif olduğunu göstermiştir. NB gibi GP dağılımı da bir ölçek parametresine sahiptir.

2. Materyal ve Metot

(y_i, x_i) ikilisi bir veri setinden gözlemlerimiz olsun. Burada y_i ve x_i sayıları sırasıyla bağımlı değişken ve bağımsız değişkenlerin bir vektörüdür. x_i verildiğinde y_i bağımlı değişkeninin aşağıda verilen olasılık fonksiyonu ile Poisson dağıldığı varsayılır.

$$P(y_i|\lambda_i, x_i) = \frac{e^{-\lambda_i}\lambda^{y_i}}{y_i!} \qquad y_i = 0,1,2,\dots \tag{1}$$

Burada λ_i, y_i bağımlı değişkeninin ortalamasıdır. y_i nin beklenen değer ve varyansı Eşitlik (2)' de verildiği gibidir.

$$E(y_i|x_i) = Var(y_i|x_i) = \lambda_i \tag{2}$$

Poisson regresyon modelinin önemli bir özelliğini olan, ortalama ve varyansın birbirine eşit olduğu eşit yayılım özelliği Eşitlik(2) ile görülmektedir.

Y'nin beklenen değerinin negatif değerler almamasını garanti etmek için, beklenen değer ve bağımsız değişkenler arasındaki ilişkiyi gösteren link fonksiyonu Eşitlik (3)'de verilen formda olmalıdır (Cameron ve Trivedi,1998).

$$\lambda_i = E(y_i|x_i) = e^{X_i'\beta} \tag{3}$$

Burada $\beta = [b_0, b_1, \dots, b_k]$ bilinmeyen parametre vektörünü göstermektedir ve X_i' ile de bağımsız değişkenler vektörünün transpozu ifade edilmektedir. (1) ve (3) eşitliği birlikte Poisson regresyon modelini ifade etmektedir.

En çok olabilirlik fonksiyonu (MLE) regresyon modelleri için en çok kullanılan tekniktir. Bir gözlem seti verildiğinde Poisson regresyon modelin log-olabilirlik fonksiyonu aşağıda verildiği gibidir:

$$l(\lambda) = l(\lambda|y) = \sum\{y_i \ln(\lambda_i) - \lambda_i - \ln(y_i!)\} \tag{4}$$

En çok olabilirlik yönteminde genellikle Newton-Raphson iterasyon tekniği kullanılır.

Genelleştirilmiş Poisson dağılımına sahip Y_i bağımlı değişkeninin olasılık fonksiyonu Eşitlik (5)'de ki gibidir:

$$P(y_i|\lambda_i, \alpha) = \frac{\lambda_i(\lambda_i+\alpha y_i)^{y_i-1}e^{-\lambda_i-\alpha y_i}}{y_i!} \qquad y_i = 0,1,2,\dots \tag{5}$$

Burada $\lambda_i > 0$ ve $0 < \alpha < 1$ 'dir. (Joe ve Zhu, 2005) Y_i'nin ortalama ve varyansı;

$$\mu_i = E(y_i) = \frac{\lambda_i}{1-\alpha} \tag{6}$$

$$V(y_i) = \frac{\lambda_i}{(1-\alpha)^3} = \frac{\lambda_i}{(1-\alpha)^2}E(y_i) = \phi E(y_i) \tag{7}$$

şeklindedir.

Bu gösterimde $\phi = {}^{\lambda_i}/_{(1-\alpha)^2}$ terimi yayılım faktörü olarak ifade edilir. Açıkça görüldüğü gibi $\alpha = 0$ olduğunda GP dağılımı λ_i parametreli klasik Poisson dağılımına indirgenmektedir ve $\alpha > 0$ olduğunda aşırı yayılım elde edilmektedir.

GP dağılımına dayalı olarak açıklayıcı değişkenler, Eşitlik (8)'de ki gibi bir log-link fonksiyonu yardımıyla regresyon modelinde birleştirilir.

$$\frac{\lambda_i}{1-\alpha} = \mu_i = E(y_i|x_i) = e^{X_i'\beta} \tag{8}$$

GP modelin log olabilirlik modeli ortalama μ ve ölçek parametresi α cinsinden Eşitlik (9)'da ki gibi yazılabilir.

$$l(\mu,\alpha|y) = log\{\mu(1-\alpha)\} + (y-1)log\{\mu-\alpha(\mu-y)\} - \big(\mu-\alpha(\mu-y)\big) - log\{y_i!\} \qquad (9)$$

GLMM birçok istatistiksel model sınıfını içermektedir. "Genelleştirme" kelimesi ile bağımlı değişkenin normal dağılmadığı, "Karma" kelimesi ile de modelde ki sabit etkilere rastgele etkilerin ilave edilmesi anlatılmak istenilmiştir.(Işık, F.,2011) Genelleştirilmiş Lineer Karma Modeller (GLMM), Genelleştirilmiş Lineer Modellerin (GLM) genişletilmiş halidir. Genelleştirilmiş Lineer Karma Modeller (GLMM) Breslow ve Clayton (1993), McGilchrist (1994) , Lee ve Nelder (1996) tarafından bulunmuştur. GLMM, GLM de sabit etkileri içeren lineer tahmin ediciye rastgele etkilerin eklenmesiyle oluşturulur. Bir Genelleştirilmiş Lineer Karma Modelde rastgele etkiler $\eta = X\beta + Zu$ lineer tahmin edicisinin bir parçasıdır ve verinin şartlı ortalaması ile lineer tahmin edicileri bir lineer form ile bağlıdır. y, $(n \times 1)$ boyutlu bağımlı değişken vektörü ve u, $(q \times 1)$ boyutlu rastgele etkiler vektörü olmak üzere,

$$E(y|u) = g^{-1}(X\beta + Zu) \qquad (10)$$

şeklinde yazılır. Burada,

- $g(.)$ türevlenebilen monoton link fonksiyonu ve $g^{-1}(.)$ fonksiyonu $g(.)$ fonksiyonunun tersi,
- X, $(n \times p)$ boyutlu modeldeki sabit etkilere ilişkin tasarı matrisi,
- β $(p \times 1)$ boyutlu sabit etkiler vektörü,
- $Z, (n \times q)$ boyutlu modeldeki rastgele ilişkilere ilişkin tasarı matrisi,
- $u, (q \times 1)$ boyutlu rastgele etkiler vektörüdür.

3. Uygulama

Bu çalışmada, Terzi ve Cengiz (2009), Cengiz (2012) ve Cengiz ve Terzi (2012) çalışmalarında kullanılan veri seti kullanılmıştır. Afyon ilindeki farklı hastanelere başvuran öksürüklü hasta sayıları ile şehir merkezinden elde edilen hava kirliliği ölçümleri arasındaki ilişki Poisson Karma ve Genelleştirilmiş Poisson Karma model yöntemleri kullanılarak ortaya konuldu. 2008-2010 tarihleri arasında rastgele seçilen 5 hastanenin haftalık hasta kayıtlarından öksürüklü hasta sayıları ve SO_2 (Sülfür dioksit) ve PM10 (Parçacıklı madde) değerleri aynı tarihlerde Afyon çevre işleri müdürlüğü hava kirliliği biriminden alınarak veri seti oluşturulmuştur. Verilerin analizi SAS 9.3 yazılımından GLIMMIX prosedürü kullanılarak yapılmıştır. Burada öksürüklü hasta sayısı bağımlı değişken ve hastaneler rastgele etki olarak alınmıştır.

Tablo1. Poisson Karma Modeller için Parametrelerin Tahmini

Sabit etki için çözüm					
Etki	**Tahmin**	**Standart hata**	**DF**	**t**	**Pr > |t|**
Intercept	0,1507	0,1805	3	0,83	0,4651
SO2	0,001482	0,00137	95	1,08	0,2835
PM10	0,00448	0,00136	95	3,3	0,0014
kovaryans Parametere tahmini					
Cov Parm		**Subject**	**Tahmin**	**Standart hata**	
Intercept		hospital	0,01246	0,024	
Şartlı dağılım için uyum istatistikleri					
-2 log L(y | r. effects)			380,47		
Pearson Ki-kare			151,72		
Pearson Ki-kare / DF			1,67		

Tablo 1' de Ölçeklendirilmiş Pearson istatistik değeri (1,5) 1'den büyük olduğu için aşırı yayılım olduğu görülmektedir. ''Sabit etki için çözüm'' kısmında parametrelerin tahminleri, standart hataları ve önemlilik testleri verilmiştir. Değerler incelendiğinde sadece PM10 değişkeninin anlamlı olduğu görülmektedir.

Tablo 2. Genelleştirilmiş Poisson Karma Modeller için Parametrelerin Tahmini

Sabit etki için çözüm					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Sabit	0,1695	0,2201	3	0,77	0,4972
SO2	0,001254	0,00165	95	0,76	0,4487
PM10	0,004472	0,00185	95	2,41	0,018
Kovaryans Parametere tahmini					
Cov Parm		**Subject**	**Tahmin**	**Standart hata**	
Sabit		hospital	0,000324	0,00037	
ölçek			0,2322	0,08027	
Şartlı dağılım için uyum istatistikleri					
-2 log L(y | r. effects)			373,01		
Pearson Chi-Square			100,71		
Pearson Chi-Square / DF			1,00		

Tablo 2' de ölçeklendirilmiş Pearson istatistik değerinin 1' e düştüğü görülmektedir. Dolayısıyla aşırı yayılım problemi ortadan kalkmıştır. Tablo 1 ile karşılaştırıldığında parametrelerin standart hatalarının arttığı da görülmektedir. Ayrıca ''Kovaryans parametre tahmini'' kısmında ölçek parametresinin (0,2322) sıfırdan farklı olduğu görülmektedir.

Tablo 3. İki model için Uyum iyiliği istatistikleri

Y	**AIC**	**AICC**	**BIC**	**CAIC**	**HQIC**	**-2log Likelihood**
Poisson Mixed Model	390,76	391,01	388,92	391,92	386,72	384,76
Generalized Poisson Mixed Models	381,01	381,42	378,55	382,55	375,62	373,01

4. Sonuç ve Tartışma

Bu çalışmada öksürüklü hasta sayılarının modellenmesinde iki farklı model kullanıldı ve parametre tahminleri açısından benzer sonuçlar elde edilse de verideki aşırı yayılım nedeniyle tahminlerin standart hataları incelendiğinde PR modelde GP modele göre daha düşük olduğu görülmektedir. Sonuçlar incelendiğinde öksürüklü hasta sayısına PM10 değişkeninin önemli bir etkisi olduğu ($p< 0,05$) ancak SO_2 nin %5 önem seviyesinde önemli bir etkiye sahip olmadığı anlaşılmaktadır. Tablo 3'te ki sonuçlara göre PR modelin uyum istatistikleri GP modele göre daha yüksektir. Bu sonuçlara göre GP model öksürüklü hasta sayılarının modellenmesinde PR modele göre daha uygun bir modeldir.

Kaynaklar

Breslow, N.E. and Clayton, D.G., (1993). Approximate Inference in Generalized Linear Mixed Models, Journal of the American Statistical Association, 88, 9−25.

Cameron, A.C. and Trivedi, P. K., (1998) Regression Analysis of Count Data: Cambridge University Press.

Cengiz, M.A. Terzi, Y., (2012).Comparing models of the effect of air pollutants on hospital admissions and symptoms for chronic obstructive pulmonary disease, Central European Journal of Public Health, 20 (4), 282-286.

Cengiz, M.A., (2012). Zero–Inflated regression models for modeling the effect of air pollutants on hospital admissions, Polish Journal Environmental Studies 21(3), 565–568.

Işık, F., (2011). Generalized Linear Mixed Models: An Introduction for Tree Breeders and Pathologists, Fourth International Workshop on the Genetics of Host- Parasite Interactions in Forestry, USA.

Joe, H. and Zhu, R., (2005). Generalized Poisson Distribution: the Property of Mixture of Poisson and Comparison with Negative Binomial Distribution, Biometrical Journal 47(2),219-229.

Lee, Y. and Nelder, J.A., (1996). Hierarchical generalized linear models (with discussion) Journal of the Royal Statistical Society, Series B, 58, 619-678.

Mcgilchrist, C.A., (1994). Estimation in Generalised Mixed Models. J. Roy. Statist. Soc. B., 56: 61-69.

Terzi, Y, ve Cengiz, M.A., (2009). Using of generalized additive model for model selection in multiple poisson regression fora ir pollution data, Scientific Research and Essays, Vol.4 (9),pp.867-871,September.