ANADOLU ÜNİVERSİTESİ BİLİM VE TEKNOLOJİ DERGİSİ –A
Uygulamalı Bilimler ve Mühendislik

ANADOLU UNIVERSITY JOURNAL OF SCIENCE AND TECHNOLOGY –A
Applied Sciences and Engineering
Cilt/Vol.: 13-Sayı/No: 1 : 51-63 (2012)

ARAŞTIRMA MAKALESİ / **RESEARCH ARTICLE**

E-POSTA TRAFİĞİNİN SIFIR DEĞER AĞIRLIKLI REGRESYON YÖNTEMLERİ KULLANILARAK İNCELENMESİ

Yılmaz KAYA [1], Abdullah YEŞİLOVA [2]

ÖZ

Sayıma dayalı olarak elde edilen veriler beklenenden fazla sıfır değerine sahip olabilirler. Bu tip verilerin analizinde sıfır değerlerini dikkate alan regresyon yöntemlerinin kullanılması daha uygun olmaktadır. Beklenenden fazla sayıda sıfır değerine sahip bağımlı değişkenin modellenmesinde sıfır değer ağırlıklı Poisson (ZIP), sıfır değer ağırlıklı negatif binomial (ZINB), Poisson Hurdle (PH) veya negatif binomial Hurdle (NBH) regresyon yöntemlerinin kullanılması daha uygun yaklaşımlardır. Bu çalışmada, Yüzüncü Yıl Üniversitesi (YYÜ) e-posta sunucusundan personelin 2009 bahar eğitim öğretim döneminde yaptıkları e-posta trafiği incelenmiştir. Veri kümesinde beklenenden fazla sayıda sıfır (%78,9) değerlerin bulunmasından dolayı veri kümesine ZIP, ZINB, PH ve NBH regresyon yöntemleri uygulanmıştır. Gönderilen e-posta sayılarında hem sıfır yayılımı hem de aşırı yayılım olduğundan dolayı aşırı yayılımı ve sıfır yayılımını dikkate alan ZINB ve NBH regresyonlarının doğru sonuçlar gösterdikleri saptanmıştır. Uyum ölçütleri ve Voung istatistiklerine göre ZINB'in veri kümesini açıklayan en iyi model olduğu görülmüştür.

Anahtar Kelimeler: E-posta trafiği, Sıfır değer ağırlıklı modeller, Sıfır yayılımı, Veri madenciliği.

INVESTIGATION OF E-MAIL TRAFFIC BY USING ZERO-INFLATED REGRESSION MODELS

ABSTRACT

Based on count data obtained with a value of zero may be greater than anticipated. These types of data sets should be used to analyze by regression methods taking into account zero values. Zero-Inflated Poisson (ZIP), Zero-Inflated negative binomial (ZINB), Poisson Hurdle (PH), negative binomial Hurdle (NBH) are more common approaches in modeling more zero value possessing dependent variables than expected. In the present study, the e-mail traffic of Yüzüncü Yıl University in 2009 spring semester was investigated. ZIP and ZINB, PH and NBH regression methods were applied on the data set because more zeros counting (78.9%) were found in data set than expected. ZINB and NBH regression considered zero dispersion and overdispersion were found to be more accurate results due to overdispersion and zero dispersion in sending e-mail. ZINB is determined to be best model according to Vuong statistics and information criteria.

Keywords: E-mail traffic, Zero inflated models, Zero inflated data sets, Data mining.

[1,] Siirt Üniversitesi, Mühendislik Mimarlık Fakültesi, Bilgisayar Mühendisliği, Siirt.
E-mail: yilmazkaya1977@gmail.com Tel: (506) 488 49 90
[2,] Yüzüncü Yıl Üniversitesi, Ziraat Fakültesi, Zootekni, Van.

Geliş: 4 Aralık 2011; **Düzeltme:** 28 Haziran 2012; **Kabul:** 5 Temmuz 2012

1. GİRİŞ

Veri madenciliği, verilerin içerisindeki desenlerin, ilişkilerin, değişimlerin, düzensizliklerin, kuralların ve istatistiksel olarak önemli olan yapıların keşfedilmesidir (Feinerer ve ark., 2008). Veri madenciliğini istatistiksel bir yöntemler serisi olarak görmek mümkün olabilir. Dolayısıyla veri madenciliği çalışması esas olarak bir istatistik uygulamasıdır. Regresyon, veri madenciliğinde değişkenler arasında neden sonuç ilişkilerini inceleyen istatistiksel yöntemlerdir. Doğru sonuçların elde edilmesi için veri kümesine uygun bir regresyon yönteminin seçilmesi gerekir. Veri kümesine uygun bir modelin seçilmesindeki amaç verilerdeki değişimi en iyi şekilde açıklamak, varyasyon kaynaklarını doğru belirlemek ve sapmasız parametre tahminlerini elde etmektir (Yeşilova ve ark., 2010).

Sayıma dayalı olarak elde edilmiş veriler genellikle Poisson dağılımı (PD) gösterir ve Poisson regresyon (PR) ile analiz edilirler. PR, bağımsız değişkenler ile sayıma dayalı olarak elde edilen bağımlı değişken arasındaki ilişkiyi açıklamaktadır. PR'da bağımlı değişkenin varyansının ortalamasından büyük çıkması aşırı yayılım (overdispersion) olarak ifade edilmektedir (Wang ve ark., 2002; Yeşilova ve ark., 2007). Aşırı yayılım durumunda PR'yi uygulamak parametre tahminlerinin ve standart hatalarının sapmalı olmasına neden olmaktadır (Khoshgoftaar ve ark., 2005). Bu gibi durumlarda aşırı yayılımı dikkate alan negatif binomial (NB) regresyon modelinin kullanılması daha uygun olabilir (Jansakul, 2005; John ve ark., 2007; Jansakul ve Hinde, 2009).

Veri kümesinde beklenenden fazla sayıda sıfır değerin olması sıfır değer yayılımı (zero inflation) olarak tanımlanmaktadır (Martin ve ark., 2006; Cui ve Yang, 2009). Gözlemlerin büyük bir kısmının sıfır olduğu veri kümelerinde, sıfır değerlerinin analiz dışı tutulması doğru olmayan sonuçların elde edilmesine neden olmaktadır. Veri kümesinde beklenenden fazla sıfır değerin bulunması durumunda veri kümesinin sıfırları göz önünde bulunduran sıfır değer ağırlıklı modeller (zero-inflated models) ile analiz edilmesi daha uygun olmaktadır (Ridout ve ark., 2001). Hurdle modeller sıfır değerlerinin çok olduğu veri kümeleri için kullanılmaktadır. Hurdle modeller iki aşamadan oluşmaktadır. Birincisi, sıfır sayımlara (0) karşı pozitif sayımları (1) gösteren ikili (binary) cevaplar; ikincisi ise yalnız pozitif sayımların meydana geldiği süreçtir (Yeşilova ve ark., 2010). Binary cevaplar logit bağlantı fonksiyonu kullanılarak modellenmektedir. Pozitif sayımlar ise sıfır değer sınırlandırılmış (zero-value truncated) sayma model yani log bağlantı fonksiyonu kullanılarak modellenmektedir (Martin ve ark., 2006). Sayma kısmının Poisson dağılımı göstermesi durumunda Poisson Hurdle (PH), negatif binom dağılımı göstermesi durumunda ise NB Hurdle (NBH) model kullanılmaktadır (Gerdtham,1997).

Sıfır değer ağırlıklı Poisson (Zero Inflated Poisson=ZIP) regresyonu da, veri kümesinin beklenenden fazla sayıda sıfır değer içermesi durumunda bağımlı değişkenin modellenmesinde kullanılmaktadır. ZIP regresyonu, bağımlı değişkeninin iki farklı veri grubundan oluştuğunu varsaymaktadır. Bunlardan birincisi, sıfır değerlerini de içerebilecek Poisson dağılımlı veri grubu olurken, buna karşın ikinci grup ise daima sıfır değerlerini içermektedir (Cameron ve Trivedi, 1998). ZIP regresyonda logit fonksiyonu bağımlı değişkenin hangi veri grubuna dâhil olduğunu belirlemek için kullanılır. Poisson dağılımı gösteren grup ise PR ile modellenir. Poisson dağılımı gösteren ikinci grupta aşırı yayılım söz konusu olduğunda ZIP regresyonu yerine sıfır değer ağırlıklı negatif binomial (Zero Inflated Negative Binomial=ZINB) regresyonu kullanılması daha uygun olmaktadır (Hall, 2000). ZIP modelde olduğu gibi sıfır gözlemler ile sıfır olmayan gözlemler ayrı olarak modellenir. Ancak ZIP regresyondan farklı olarak ZINB'de sıfır olmayan gözlemler NB regresyonu ile modellenmektedir. PR, NB, PH , NBH, ZIP ve ZINB regresyon modellerinde parametre tahminleri yaygın olarak EM algoritmasını esas alan en yüksek olabilirlik (Maximum Likelihood=ML) yöntemi kullanılarak elde edilmektedir (Karen ve Kelvin, 2005). Uygun model seçiminde Akaiki (AIC) bilgi ölçütü kullanılabilir. En küçük bilgi ölçütlerine sahip model en iyi model olarak kabul edilmektedir.

Bu çalışmada e-posta trafiği sıfır değer ağırlıklı regresyon yöntemleri ile incelenmiştir.

Çalışmadaki veri kümesi Yüzüncü Yıl Üniversitesi (YYÜ) 568 akademik ve 595 idari toplam 1163 personelden elde edilmiştir. Öncelikle YYÜ e-posta sunucusundan personele ait e-posta adresleri elde edilerek, her personelin 2008-2009 öğretim yılı bahar döneminde (Mart, Nisan, Mayıs, Haziran) gönderdiği e-posta sayıları sunucu kayıt(log) dosyalarından elde edilmiştir. Daha sonra 1163 personele anket uygulanarak kişilerin çalıştığı birim, unvan, cinsiyet, medeni hal, yaş, kişilerin kullandıkları e-posta adres sayısı, en çok kullandıkları e-posta adresinin "yyu.edu.tr" uzantılı olup olmadığı, msn-icq gibi herhangi bir sohbet programını kullanıp kullanmadıkları, "yyu.edu.tr" uzantılı e-posta adresi ile herhangi bir bilimsel toplantıya katılıp katılmadıkları, ortalama bir günde kaç saat internet kullandıkları, web sitelerinin olup olmadığı, evde internet bağlantılarının olup olmadığı ve dizüstü bilgisayarlarının olup olmadığı sorularına ilişkin cevaplar elde edilmiştir. Veri kümesindeki beklenenden fazla sıfır değerlerin bulunmasından dolayı, ZIP, ZINB, PH, NBH regresyonları uygulanmıştır. Modellerin karşılaştırılması AIC bilgi ölçütü Voung istatistiklerinden yararlanılmıştır.

2. MATERYAL VE YÖNTEM

2.1 Materyal

YYÜ e-posta sunucusundan ve anket yolu ile elde edilen değişkenler Çizelge 1 de verilmiştir.

Toplam değişkeni bağımlı değişken, Personel tip, Birim kodu, Unvan kodu, Cinsiyet, Medeni hal, Yaş, E-posta adet, E-posta YYU, Sohbet programı, Bilimsel toplantı, Ortalama internet, Website, İnternet bağlantısı ve Dizüstü bilgisayar değişkenleri ise bağımsız değişkenler olarak değerlendirilmiştir. Bu çalışmada, analizler için R yazılımı kullanılmıştır. R istatistik, matematik, veri madenciliği gibi çok farklı amaçlar için kullanılabilecek bir programlama dilidir. Açık kaynaklı bir program olup GNU lisansı altında dağıtılmaktadır.

2.2 Yöntem

2.2.1 Sıfır Yayılımı

Veri kümeleri beklenenden daha fazla sıfır değerin içermesi durumunda, sıfır yayılımı (Zero inflation=ZI) meydana gelmektedir. Sıfır değer ağırlıklı dağılımlar gözlemlerin iki gruba ait olduğunu varsaymaktadır. Birinci grup (*g*1*)*, gözlemlerin doğrudan sıfır olarak gözlendiği gruptur. İkinci grup *(g*2*)* Poisson veya NB dağılımı gösteren gözlemlerin ait olduğu gruptur. *Z_i'ye (indikatör değişken)* bağlı olarak gözlemin *g*1 grubuna ait olma olasılığını belirten indikatör değişkendir ve $Z_i \sim Berneulli(p_i)$ biçiminde yazılabilir (Tin, 2008).

$$\begin{aligned} p_i &= Pr(i \in g1 \mid Z_i) \\ 1 - p_i = q &= Pr(i \in g2 \mid Z_i) \end{aligned} \tag{1}$$

olduğu varsayılsın. Bu durumda sıfır değer ağırlıklı modelin genel formu,

$$\begin{aligned} Pr(y_i \mid x_i, z_i) &= p_i + (1 - p_i) g(\mu_i) \quad \text{eğer } y_i = 0 \\ Pr(y_i \mid x_i, z_i) &= (1 - p_i) f(\mu_i) \quad \text{eğer } y_i > 0 \end{aligned} \tag{2}$$

biçiminde yazılabilir. $g(\mu_i) = Pr(y_i = 0 \mid x_i)$ bağımlı değişkenin sıfır olduğunu göstermektedir. $f(\mu_i)$ Poisson veya NB dağılımlarından herhangi birini göstermektedir (Dominique ve ark., 2005).

2.2.2 Poisson Hurdle Regresyon

Hurdle modelde verilerin elde edilişi iki farklı aşamada gerçekleşir. Birinci aşama geçiş aşaması (transation stage) olarak bilinir ve binomial dağılım gösterir. Bu aşamada bağımlı değişkenin sıfır veya sıfır olmaması belirlenir (Jansakul ve Hinde, 2009). İkinci aşama olay aşaması (event stage) olarak bilinir. Tüm sıfır değerler hariç elde edilen değerler bu aşamada modellenir.

Poisson hurdle model, soldan sınırlandırılmış sayıma dayalı olarak elde edilen pozitif gözlem değerleri *(y_i>0)* Poisson dağılımı kullanılarak modellendiğinde, Poisson hurdle model olarak adlandırılmaktadır (Rose ve ark.,2006). y_i, $i=1,2,\ldots,n$ 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* ortak değişken matrisleridir. Poisson hurdle model, Çizelge 1. Veri kümesini oluşturan değişkenler.

Kod	Değişken	Açıklama
Personel Tip	Personel Tip	1=Akademik Personel , 0=İdari Personel
Birim Kodu	Birim Kodu	Her fakülte veya yüksekokula bir kod verilmiştir.
Unvan Kodu	Unvan Kod	1=Prof. Dr., 2=Doç. Dr., 3=Yrd.Doç. Dr., 4=Öğr. Grv., 5=Okutman, 6=Araş.Grv., 7=Uzman, 0=İdari Personel
Cinsiyet	Cinsiyet	0=Bayan, 1=Bay
Medeni Hal	Medeni hal	0=Bekar, 1=Evli
Yas	Yaş	Personel yaşı
Toplam	Gönderilen e-posta sayıları	Her personelin gönderdiği e-posta sayısı
Eposta Adet	Kişinin e-posta sayısı	Her personelin e-posta adres sayısı
Eposta YYU	En çok kullanılan e-posta adresi "yyu.edu.tr" uzantılı mı?	1=Evet, 0=Hayır
Sohbet Programı	MSN, ICQ gibi sohbet programları kullanıyor musunuz?	1=Evet, 0=Hayır
Bilimsel Toplantı	YYU uzantılı e-posta ile bir bilimsel toplantıya katıldınız mı?	1=Evet, 0=Hayır
Ortalama İnternet	Ortalama günlük internet kullanımı kaç saat?	Ortalama günlük internet kullanımı (saat)
Web Site	Web siteniz var mı?	1=Evet, 0=Hayır
Internet Bağlantısı	Evde internet bağlantınız var mı?	1=Evet, 0=Hayır
Dizüstu Bilgisayar	Dizüstü bilgisayarınız var mı?	1=Evet, 0=Hayır

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

$$P(y_i = q/x,z) = \frac{p(x)exp(-\lambda(z))\lambda(z)^q}{q!\left(1 - exp(-\lambda(z))\right)} \qquad (3)$$

biçiminde yazılabilir (Min ve Agresti., 2005). Eşitlik 3'de verilen *p(x)* ve $\lambda(z)$ sırasıyla logit ve log-doğrusal fonksiyonları ile modellenmektedirler. Yani,

$$log\left(\lambda(z)\right) = x_i'\beta \qquad (4)$$

$$logit(p_i) = z_i'\alpha \tag{5}$$

biçiminde modellenmektedirler (Rose ve ark.,2006). Eşitlik 4 ve eşitlik 5'te verilen β ve α sırasıyla bilinmeyen parametre vektörleridir. β, α parametrelerinin tahmin edilmesinde ML yöntemi kullanılmaktadır. Poisson hurdle için log olabilirlik fonksiyonu,

$$\begin{aligned} L = & \sum_{y_i>0} x_i\beta - \sum_{i=1}^{n} log\left(1 + exp(x_i\beta)\right) \\ & + \sum_{y_i>0}\left[y_i z_i\alpha - exp(z_i\alpha) \right. \\ & \left. - log\left(1 - exp(-exp(z_i\alpha))\right) - log(y_i!)\right] \\ = & L(\beta) + L(\alpha) \end{aligned} \tag{6}$$

biçiminde yazılmaktadır. $L(\beta)$ *ve* $L(\alpha)$ ayrı ayrı maksimize edilerek ML tahminleri elde edilmektedir (Min ve Agresti., 2005).

2.2.3 Negatif Binom Hurdle Regresyonu

Negatif binomial Hurdle'da, sayıma dayalı olarak elde edilen bağımlı değişkenin sıfır ya da sıfır değerli olmama sonuçlarını belirleyen binomiyal olasılık modeli ile pozitif sonuçları tanımlayan sınırlandırılmış sayıma dayalı modeli için verilen log olabilirlik fonksiyonu aşağıdaki gibi yazılabilir (John ve ark., 2007),

$$\mathrm{L} = \ln(f(0)) + \left\{\ln\left[1 - f(0)\right] + \ln \mathrm{P(j)}\right\} \tag{7}$$

Eşitlik 7'te verilen *f(0)* modelin binary kısmının olasılığını göstermektedir. *P(j)* pozitif sayımın olasılığını göstermektedir. Logit model kullanılması durumunda, sıfır sayımın olasılığı,

$$f(0) = P(y = 0; x) = 1/1 + exp(x\beta_1))$$

ve

$1 - f(0)$ ise,

$$exp(x\beta_1)/(1 + exp(x\beta_1))$$

biçiminde yazılabilir. Böylece negatif binomiyal Hurdle modelin her iki kısmı için log olabilirlik fonksiyonu aşağıdaki gibi yazılabilir.

$$\begin{aligned} & cond\left\{y = 0, ln(1/1 - exp(xb1)), \right. \\ & ln\left(exp(xb1)/(1 + exp(xb1))\right) + y * ln\left(exp(xb)/(1 + exp(xb))\right) \\ & - ln(1 + exp(xb))/\alpha + ln\,\Gamma(y + 1/\alpha) - ln\,\Gamma(y+1) - ln\,\Gamma(1/\alpha) \\ & \left. - ln(1 - (1 + exp(xb))(-1/\alpha)\right\} \end{aligned} \tag{8}$$

2.2.4 Sıfır Değer Ağırlıklı Poisson Regresyon

Veri kümesinde beklenen fazla sayıda sıfır değerlerin bulunması durumunda PR ve NB modelleri yerine, fazla sayıda sıfırlardan kaynaklanan aşırı yayılımı dikkate alan ZIP modeli kullanılabilir. Fazla sayıda sıfır değerine sahip y_i açıklamak için, ZIP regresyon modeli,

$$Pr(y_i/x_i) = \begin{cases} \pi_i + (1 - \pi_i)exp(-\mu_i) & y_i = 0 \\ (1 - \pi_i)exp(-\mu_i)\mu_i^{y_i}/y_i! & y_i > 0 \end{cases} \tag{9}$$

biçiminde yazılabilir (Rose ve ark., 2006). Eşitlik 9'da, π ekstra sıfırların olma olasılığını göstermektedir. $y_i = 0$ olan gözlemler iki gruptan oluşmaktadır. Bu gruplardan biri, gözlemlerin Poisson süreci göstermediği, diğeri ise gözlemlerin,

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

olmasından dolayı μ ortalamalı Poisson dağılımı gösterdiği gruptur. ZIP regresyonda her iki grup için farklı bağlantı fonksiyonu kullanılır. Logit bağlantı fonksiyonu, potansiyel "yyu.edu.tr" uzantılı e-posta adresi kullanan kullanıcıları göstermek için kullanılırken, log bağlantı fonksiyonu ise "yyu.edu.tr" e-posta adresini kullanan kullanıcıların gönderdikleri e-posta sayıları ile bağımsız değişkenler arasındaki bağlantıyı sağlamaktadır. Yani,

$$\begin{aligned} log(\mu) &= X\beta \\ \mu &= exp(X\beta) \end{aligned} \tag{10}$$

$$\log it(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = G\gamma$$
$$\pi = \frac{exp(G\gamma)}{1+exp(G\gamma)} \quad (11)$$

bağlantı fonksiyonları kullanılarak parametre tahminleri elde edilmektedir [7]. Eşitlik 10 ve 11'de *X* ve *G* ortak değişken matrisleri β ve γ sırasıyla, *(p+1)x1 ve (q+1)x1* boyutlu bilinmeyen parametre vektörleridir. ZIP regresyon modeli için log olabilirlik fonksiyonu,

$$L(y,\beta,\gamma) = \sum_{y_i=0} \log\left(e^{G_i\gamma} + exp\left(-e^{X_i\beta}\right)\right) + \sum_{y_i>0}\left(y_iX_i\beta - e^{X_i\beta}\right) - \sum_{i=1}^{n} \log\left(1+e^{G_i\gamma}\right) - \sum_{y_i>0} \log(y_i!) \quad (12)$$

biçiminde yazılabilir (Cui ve Yang, 2009) Eşitlik 12'de G_i ve X_i, *G* ve *X* matrislerinin *i*'ninci sırasını göstermektedir. Eşitlik 12'da 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 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,

$$\begin{aligned} L(\gamma,\beta,y,Z) &= \sum_{i=1}^{n} \log\left(f(Z_i/\gamma)\right) + \sum_{i=1}^{n} \log\left(f(y_i/Z_i,\beta)\right) \\ &= \sum_{i=1}^{n}\left(Z_iG_i\gamma - \log\left(1+e^{G_i\gamma}\right)\right) \\ &\quad + \sum_{i=1}^{n}(1-Z_i)\left(y_iX_i\beta - e^{X_i\beta}\right) - \sum_{i=1}^{n}(1-Z_i)\log(y_i!) \\ &= L(\gamma;y,Z) + L(\beta,y,Z) - \sum_{i=1}^{n}(1-Z_i)\log(y_i!) \end{aligned} \quad (13)$$

biçiminde yazılabilir. ML tahminleri EM algoritması kullanılarak elde edilir. EM algoritması kullanılarak,

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

$$z_i = \begin{cases} \left(1+e^{-G_i\gamma(k) - exp\left(X_i\beta^{(k)}\right)}\right)^{-1}, & y_i = 0 \\ 0, & y_i = 1,2,... \end{cases} \quad (14)$$

biçiminde yazılabilir. Eşitlik 14'de verilen *k* EM algoritmasının iterasyon sayısını göstermektedir (Cui ve Yang, 2009).

M-aşaması: Tüm veriler için eşitlik 14'de verilen log olabilirlik fonksiyonunun maksimize edilmesiyle γ parametresi,

$$L\left(\gamma;y,Z^{(k)}\right) = \sum_{y_i=0} Z_i^{(k)}G_i\gamma - \sum_{y_i=0} Z_i^{(k)} \log\left(1+e^{G_i\gamma}\right) - \sum_{y_i=0}\left(1-Z_i^{(k)}\right)\log\left(1+e^{G_i\gamma}\right) \quad (15)$$

biçiminde tahmin edilebilir. Yukarıda verilen E ve M aşamaları yakınsama ölçütü (10^{-6}) elde edilinceye kadar devam edilir (Yeşilova ve ark., 2010; Rose ve ark., 2006).

2.2.5 Sıfır Değer Ağırlıklı Negatif Binomiyal Regresyon

Sıfır değer ağırlıklı negatif binomiyal regresyon (Zero-inflated negative binomiyal regression=ZINB) Sıfır değerlerinin çok fazla olduğu y_i bağımlı değişkeninin modellenmesinde kullanılan alternatif regresyon yöntemidir. ZINB regresyon modeli,

$$Pr(y_i/x_i) = \begin{cases} \pi_i + (1-\pi_i)(1+\alpha\mu_i)^{-\alpha^{-1}}, & y_i = 0 \\ (1-\pi_i)\dfrac{\Gamma\left(y_i+\alpha^{-1}\right)\alpha^{y_i}\mu_i^{y_i}}{y_i!\,\Gamma\left(\alpha^{-1}\right)\left(1+\alpha\mu_i\right)^{y_i+\alpha^{-1}}}, & y_i > 0 \end{cases} \quad (16)$$

biçiminde yazılabilir (Jansakul ve Hinde, 2009). Eşitlik 16'da, π_i ve μ_i parametreleri kovarietlere bağımlı ve ($\alpha \geq 0$) yayılım parametresidir. ZINB modelde α ve π sıfıra eşit olması durumunda ZINB dağılımı poisson dağılımına dönüşmektedir (Mwalili ve ark., 2008). ZINB log olabilirlik fonksiyonu,

$$L(\mu,\alpha,\pi;y) = \sum_i \Bigg(I_{y_i=0} \log\left[(1-\pi_i)(1+\alpha\mu_i)^{-\alpha^{-1}} + \pi_i\right] + I_{y_i>0} \log\left((1-\pi_i) \frac{\Gamma\left(y_i+\frac{1}{\alpha}\right)(\alpha\mu_i)^{y_i}}{y_i!\,\Gamma\left(\frac{1}{\alpha}\right)(1+\alpha\mu_i)^{y_i+\frac{1}{\alpha}}} \right) \Bigg)$$

$$= \sum_i \Bigg(I_{y_i=0} \log\left[(1-\pi_i)(1+\alpha\mu_i)^{-\alpha^{-1}} + \pi_i\right] + I_{y_i>0} \Bigg(\log(1-\pi_i) - \frac{1}{\alpha}\log(1+\alpha\mu_i) - y_i \log\left(1+\frac{1}{\alpha\mu_i}\right) + \log\Gamma\left(y_i+\frac{1}{k}\right) - \log\Gamma\left(\frac{1}{\alpha}\right) - \log y_i! \Bigg) \Bigg) \quad (17)$$

biçiminde yazılabilir. Eşitlik 17'de, $I_{(.)}$ tesadüfi bir indikatör fonksiyonudur.

2.2.6 Model Seçimi

Akaiki bilgi ölçütü (AIC) model uyumu için kullanılan uyum ölçütüdür (Min ve Agresti, 2005). AIC bilgi ölçütü;

$$AIC = -2\log L + 2r \quad (18)$$

biçiminde hesaplanmaktadır. AIC bilgi ölçütü modellerin karşılaştırılmasında kullanılabilirler. En küçük bilgi ölçütüne sahip en uygun model olarak kabul edilir (Hall, 2000).

2.2.7 Vuong İstatistiği

Voung istatistiği modelleri birbirleri ile karşılaştırmak için kullanılır (Vuong, 1989). Vuong istatistiği,

$$V = \frac{\overline{m}\sqrt{n}}{S_m} \quad (19)$$

biçiminde hesaplanır (Vuong, 1989). Burada $m_i = \left[\frac{\hat{P}_1(Y_i \mid X_i)}{\hat{P}_S(Y_i \mid X_i)}\right]$, P_S genellikle Poisson veya NB dağılımlardan biri ve P_1 ise sıfır değer ağırlıklı Poisson veya hurdle dağılımlarından biri olmaktadır. m_i istatistiği; $\overline{m}$ ortalamalı ve S_m standart sapmalıdır. *V* asimtotik olarak normal dağılımlıdır. Eğer V>1.96 ise sıfır değer ağırlıklı model uygun olmaktadır. V<=1.96 ise NB veya Poisson regresyon yöntemleri daha uygun olmaktadır.

3. BULGULAR

3.1 Aşırı ve Sıfır Yayılımlarının Belirlenmesi

YYÜ e-posta sunucusu üzerinde herhangi bir e-posta adresi olmayan personel e-posta gönderemeyeceğinden dolayı gönderilen "*toplam*" e-posta değişkeni doğrudan sıfır olacaktır. Bu sıfırlar sistematik sıfırlar (true zeros, samples zeros, systematic zeros) olarak tanımlanmaktadır. Diğer taraftan YYÜ e-posta sunucusu üzerinde bir e-posta hesabı olup da ele alınan süre içinde e-posta göndermeyenler için alınan "*toplam*" değişkeni yapısal veya şansa bağlı sıfırları (false zeros, structural zeros, random zeros) kabul edilmektedir. Sıfır kaynaklarının personel tipine ve cinsiyete göre dağılımı Çizelge 2.'te verilmiştir.

YYÜ e-posta sunucusu üzerinde e-posta hesabı olmayan 221 akademik personel ve 496 idari olmak üzere toplam 717 personel bulunmaktadır. Dolayısıyla bağımlı değişkenin 717/1163=0.616 (%61.6)'sı sistematik sıfırlardan, 202/1163 =0.173 (%17.3)'sı ise şansa bağlı sıfırlardan kaynaklanmıştır. Böylece veri kümesindeki sıfırların oranı 0.616+0.173=0.789 olarak elde edilmiştir. Veri kümesindeki sıfır yayılımının olup olmadığı,

$$ZI(\text{Zero Inflation}=\text{Sıfır Yayılım}) = 1+\log(p_0)/\mu = 1+\log(0.789)/12.033 = 0.98$$

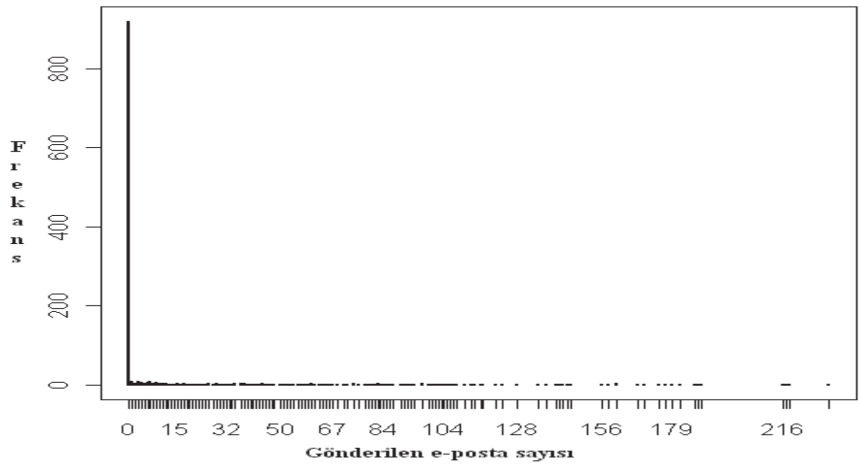
şeklinde hesaplanmıştır. ZI>0 olduğundan veri kümesinde sıfır yayılımının olduğunu göstermiştir. p_0 bağımlı değişkenindeki sıfır oranını, μ ise bağımlı değişkenin ortalamasını göstermektedir. Gönderilen e-posta sayılarının (bağımlı değişken) ortalaması 12.033 ve varyansı 1151.247 olarak elde edilmiştir. Varyansın ortalamadan çok büyük olması veri kümesinde aşırı yayılım olduğunu göstermiştir. Poisson modele göre ortalaması 12.033 olan değişken için beklenen sıfır sayısı,

$$E(Frk(Y)) = Frk(Y) * \exp(-\overline{Y}) = ne^{-\lambda} = 1163 * \exp(-12.033) \approx 8$$

olarak elde beklenilmektedir. Gönderilen e-posta sayıların dağılımı Şekil 1'de verilmiştir.

Çizelge 2. Elde edilen sistematik ve şansa bağlı sıfırların personele tipine göre dağılımı.

Sıfır Kaynağı / Cinsiyet	Akademik Personel		İdari Personel	
	Sistematik sıfırların sayısı	Şansa bağlı sıfırların sayısı	Sistematik sıfırların sayısı	Şansa bağlı sıfırların sayısı
Bay	164	121	412	55
Bayan	57	19	84	7
Toplam	221	140	496	62



Şekil 1. Gönderilen e-posta sayıların dağılımı.

Şekil 1 incelendiğinde veri kümesindeki sıfır değerlerin yoğunluğundan dolayı bağımlı değişkenin dağılımı sola doğru çarpıklık göstermiştir.

ZIP, ZINB, PH ve NBH modellere ait Akaiki bilgi ölçütlerin değerleri Çizelge 3'te verilmiştir. Hangi modelin veri kümesini en iyi açıkladığı AIC bilgi ölçütüne göre karar verilebilir.

Çizelge 3'te verilen AIC değerlerine göre veri kümesini en iyi açıklayan model ZINB model olmuştur. En küçük AIC değerine sahip model en uygun model olarak kabul edilmektedir. Kullanılan modellerin veri kümesini açıklama yüzdeleri Çizelge 4'de verilmiştir.

Çizelge 3. AIC değerleri

ZIP	ZINB	PH	NBH
8858.188	**2944.902**	8858.192	2945.325

Çizelge 4. Modellerin veri kümesini açıklama yüzdeleri.

	AIC	$\Delta_i = AIC_i - AIC_{\min}$	$w_i = \dfrac{\exp[-\frac{1}{2}\Delta_i]}{\sum_{i=1}^{m}\exp[-\frac{1}{2}\Delta_i]}$
ZIP	8858.188	5913.286	0
ZINB	**2944.902**	0	0.5526788
PH	8858.192	5913.29	0
NBH	2945.325	0.423	0.4473212

Çizelge 4'e göre veri kümesini en iyi açıklayan ZINB modelidir. Diğer modellere göre ZIB modeli %55.27, NBH %44.73, ZIP %0 ve PH %0 uygun modeller olarak ifade edilmektedir.

Vuong, iki modeli birbirleri ile karşılaştırarak hangi modelin veri setini iyi açıkladığını belirten bir testtir. ZIP, ZINB, PH ve NBH regresyonları ikili olarak karşılaştırmalar sonucu elde edilen Vuong istatistikleri Çizelge 5'te verilmiştir. Çizelge 5'te 1.sütun Model1, 1.satır Model2'yi göstermektedir.

Çizelge 5. Vuong İstatistikleri

Model 1 \ Model 2	ZIP	ZINB	PH	NBH
ZIP		V= -8.55 P= 6.31e-18 Model2>Model1	V= 0.69 P= 0.24 Model1>Model2	V= -8.55 P= 6.34e-18 Model2>Model1
ZINB			V= 8.55 P=0.001 Model1>Model2	V= 0.88 P= 0.189 Model1>Model2
PH				V= -8.55 P= 6.34e-18 Model2>Model1

Vuong istatistiklerine bakıldığında en iyi modelin ZINB model olduğu saptanmıştır. ZIP ve ZINB modelleri karşılaştırıldığında ZINB modelin ZIP modelden daha uygun (-8.55, $p<0.001$) olduğu, ZINB ve PH modeller karşılaştırıldığında ZINB modelin PH modelden daha uygun (8.55, $p<0.001$) olduğu ve ZINB model ile NBH modeller karşılaştırıldığında ZINB modelin NBH modelden daha uygun (V= 0.88, $p>0.1$) olduğu saptanmıştır. Vuong istatistiklerine göre ZINB, regresyon yöntemleri içinde en iyi sonuçları vermiştir. ZINB regresyon modeline ait parametre tahminleri ve standart hatalar Çizelge 6'da verilmiştir.

Çizelge 6. ZINB regresyonuna ait parametre tahminleri

Negatif Binomiyal Kısmı (Log)					
Değişken	**Parametre Tahminleri**	**Standart Hata**	**t değeri**	**p >|t|**	$e^{\hat{\beta}_i}$
Intercept	1.412954	0.631698	2.237	0.025302 **	4.1080720
Personel_Tip	0.386158	0.264188	1.462	0.143829	1.4713172
Birim_Kod	-0.014268	0.010553	-1.352	0.176368	0.9858337
Unvan_Kod	0.018666	0.047177	0.396	0.692353	1.0188415
Cinsiyet	0.015995	0.188744	0.085	0.932466	1.0161232
Medeni_Hal	-0.054806	0.159028	-0.345	0.730370	0.9466683
Yas	0.003772	0.011277	0.334	0.738005	1.0037792
Posta_Adet	0.028656	0.060614	0.473	0.636386	1.0290703
Posta_YYU	0.542119	0.124001	4.372	1.23e-05 ***	1.7196475
Sohbet_Prog	0.490017	0.114137	4.293	1.76e-05 ***	1.6323447
Bilimsel_Top	0.659047	0.141204	4.667	3.05e-06 ***	1.9329490
Ortalama_Int	0.119709	0.022308	5.366	8.04e-08 ***	1.1271687
WebSite	0.561504	0.125038	4.491	7.10e-06 ***	1.7533069
Internet_Bag	0.213902	0.151739	1.410	0.158638	1.2385017
Dizustu_Bil	-0.132028	0.180807	-0.730	0.465257	0.8763161
Logit Kısmı					
Değişkenler	**Parametre Tahminleri**	**Standart Hata**	**t değeri**	**p >|t|**	$e^{\hat{\beta}_i}$
Intercept	4.86511	1.09242	4.454	8.45e-06 ***	129.68542966

Çizelge 6. (Devamı) ZINB regresyonuna ait parametre tahminleri					
Personel_Tip	-2.78044	0.47147	-5.897	3.69e-09 ***	0.06201125
Birim_Kod	0.03196	0.01679	1.903	0.056995 *	1.03247948
Unvan_Kod	0.35032	0.09032	3.879	0.000105 **	1.41952713
Cinsiyet	-0.01468	0.34491	-0.043	0.966044	0.98542423
Medeni_Hal	0.20276	0.29968	0.677	0.498683	1.22477268
Yas	0.01215	0.01936	0.627	0.530338	1.01221951
Posta_Adet	-0.67190	0.11519	-5.833	5.45e-09 ***	0.51073865
Posta_YYU	-1.09429	0.25154	-4.350	1.36e-05 ***	0.33477862
Sohbet_Prog	0.46296	0.25081	1.846	0.064919 *	1.58876446
Bilimsel_Top	-2.73451	0.26358	-10.374	< 2e-16 ***	0.06492604
Ortalama_Int	-0.35153	0.04530	-7.759	8.55e-15 ***	0.70361368
WebSite	0.03288	0.29900	0.110	0.912434	1.03342765
Internet_Bag	-0.14302	0.28789	-0.497	0.619337	0.86673386
Dizustu_Bil	0.50115	0.31069	1.613	0.106737	1.65062187

***p<0.001, **p<0.05, *p<0.1

ZINB modelde parametre tahmini yorumlanırken logit ve log kısımlarını ayrı değerlendirmek gerekir. Logit kısmı için ZINB regresyonu sonuçları,

$$\left(\frac{\pi}{1-\pi}\right) = exp(4.9\text{-}2.8Personel\_Tip+0.03Birim\_Kod+0.35Unvan\_Kod\text{-}0.02Cinsiyet$$
$$+0.20Medeni\_Hal+0.01Yas\text{-}0.67Posta\_Adet\ \text{-}1.09Posta\_YYU+0.46Sohbet\_Prog$$
$$\text{-}2.74Bilimsel\_Top\text{-}0.35Ortalama\_Int+0.03WebSite\text{-}0.14Internet\_Bag+0.50Dizustu\_Bil$$

olarak elde edilmiştir. e-posta gönderme olasılığı üzerinde Birim_Kod, Unvan_Kod, Sohbet_Prog değişkenleri pozitif yönde, Personel_Tip, Posta_Adet, Posta_YYU ve Bilimsel_Top değişkenleri negatif yönde istatistiksel olarak önemli bir etki gösterdikleri saptanmıştır (p<0.05). İdari personelin akademik personele göre $(e^{-2.78} \sim 0.06)$ %94 daha az e-posta gönderme eğiliminde oldukları saptanmıştır. Birimler arası farklılık e-posta gönderme olasılığını $(e^{0.03} \sim 1.032)$ %3 değiştirmiştir. Personelin akademik ünvan olarak yükselmesi e-posta göndermeyi $(e^{0.35} \sim 1.419)$ %42 artırmıştır. Kişilerin e-posta adresi sayısındaki bir adet artış e-posta göndermeyi $(e^{-0.671} \sim 0.51)$ %49 oranında azaltmıştır. En çok “yyu.edu.tr” uzantılı e-posta adresini kullanan personelin kullanmayanlara göre e-posta göndermeyi $(e^{-1.094} \sim 0.334)$ %67 azaltmıştır. Bir sohbet programını kullanan personelin kullanmayanlara göre e-posta gönderme olasılığını $(e^{0.4629} \sim 1.588)$ %59 artırmıştır. “yyu.edu.tr” uzantılı e-posta adresi ile bir bilimsel toplantıya katılan personelin katılmayanlara göre e-posta göndermeyi $(e^{-2.73} \sim 0.065)$ %93 azaltmıştır. Günlük ortalama internet kullanımı artıkça e-posta gönderme olasılığının $(e^{-0.143} \sim 1.03)$ %3 azaldığı saptanmıştır.

ZINB regresyonun NB kısmı için elde edilen regresyon denklemi,

$$\mu = exp(1.41+0.37Personel\_Tip-0.02Birim\_Kod+0.02Unvan\_Kod+0.02Cinsiyet - 0.05Medeni\_Hal+0.01Yas+0.03Posta\_Adet +0.54Posta\_YYU+0.49Sohbet\_Prog + 0.66Bilimsel\_Top+0.12\ Ortalama\_Int+0.56WebSite+0.21Internet\_Bag-0.13Dizustu\_Bil$$

olarak elde edilmiştir. Parametre tahminlerine bakıldığında, gönderilen e-posta sayısı üzerinde Posta_YYU, Sohbet_Prog, Bilimsel_Top, Ortalama_Int, Website değişkenleri istatistiksel olarak önemli bulunmuştur (p<0.001). E-posta göndermek için en çok "yyu.edu.tr" uzantılı e-posta adresini kullanan personelin kullanmayanlara göre gönderilen ortalama e-posta sayısını $(e^{0.542} \sim 1.719)$ %72 artırmışlar. Herhangi bir sohbet programını kullanan personelin kullanmayanlara göre gönderilen ortalama e-posta sayısını $(e^{0.49} \sim 1.63)$ %63 arttırdıkları saptanmıştır. "yyu.edu.tr" uzantılı e-posta ile bir bilimsel toplantıya katılan personelin katılmayanlara göre ortalama e-posta sayısını $(e^{0.659} \sim 1.932)$ %93 artırdıkları ve Web sitesi olan personelin olmayanlara göre ortalama e-posta sayısını $(e^{0.5615} \sim 1.75)$ %75 artırdıkları saptanmıştır.

4. SONUÇ

Sayıma dayalı olarak elde edilen verilere Poisson regresyonun uygulanabilirliği veri kümesinin ortalama ile varyanslarının birbirine eşit olmasına bağlıdır. Gönderilen e-posta sayılarının (bağımlı değişken) ortalaması 12.033 ve varyansı 1151.247 olarak elde edilmiştir. Varyansın ortalamadan çok büyük olması veri kümesinde aşırı yayılım olduğunu göstermiştir.

Veri kümesinde beklenenden fazla sıfır değerlerin olması durumunda sıfır yayılımı meydana gelir. "*toplam*" değişkenin %78.9'u sıfır değerlerinden oluştuğundan veri kümesinde sıfır yayılımı da söz konusudur. Veri kümelerinde sıfır yayılımı durumunda Poisson Hurdle, NB Hurdle, ZIP ve ZINB gibi modeller yaygın olarak kullanılmaktadır.

Veri kümesini en iyi açıklayan model ZINB model olmuştur. ZINB hem aşırı hem de sıfır yayımlını dikkate aldığından en uygun model olarak gözlenmiştir. ZINB modelin en uygun model olduğu AIC bilgi kriteri (2944.902) ve Voung istatistikleri de desteklemiştir.

ZINB en iyi modelin parametre tahminlerine göre gönderilen e-posta sayısı üzerinde Posta_YYU, Sohbet_Prog, Bilimsel_Top, Ortalama_Int, Website değişkenleri istatistiksel olarak önemli bulunmuştur (p<0.001). E-posta gönderme olasılığı üzerinde ise Birim_Kod, Unvan_Kod, Sohbet_Prog, Personel_Tip, Posta_Adet, Posta_YYU ve Bilimsel_Top değişkenleri istatistiksel olarak önemli bir etki gösterdikleri saptanmıştır (p<0.05).

KAYNAKLAR

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

Cui, Y. and Yang, W. (2009). Zero-inflated generalized Poisson regression mixture model for mapping quantitative trait loci underlying count trait with many zeros. *Journal of Theoretical Biology* 256, 276–285.

Dominique, L., Simon, P.W. and John, N.I. (2005). Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis and Prevention* 37, 35–46.

Gerdtham, U.G. (1997). Equity In Health Care Utilization: Further Tests Based on Hurdle Models And Swedish Micro Data. *Health Economics* 6, 303–319.

Feinerer, I., Hornik, K. and Meyer, D. (2008). Text Mining Infrastructure in R. Journal of Statistical Software 25(5), 12-18.

Hall, D.B. (2000). Zero-inflated Poisson and negative binomial regression with random effects: A case study. *Biometrics* 56, 1030-1039.

Jansakul, N. (2005). Fitting A Zero-Inflated Negative Binomial Model via R. *20th International Workshop on Statistical Modelling*. Sydney, Australia. 277-284.

Jansakul, N. and Hinde, J.P. (2009). Score Tests for Extra-Zero Models in Zero-Inflated Negative Binomial Models. *Communications in Statistics - Simulation and Computation* 38(1), 92 -108.

John, M.W., Hung, M.L., Robert, H.L. and Allen, W.H. (2007). Power Calculations for ZIP and ZINB Models. *Journal of Data Science* 5, 519-534.

Karen, C.H.Y. and Kelvin, K.W.Y. (2005). On modeling claim frequency data in general insurance with extra zeros. *Mathematics and Economics* 36, 153–163.

Khoshgoftaar, T.M., Gao, K. and Szabo, R.M. (2005). Comparing Software Fault Predictions of Pure and Zero- inflated Poisson Regression Models. *International Journal of Systems Science* 36(11), 707-715.

Martin, S.W., Rose, C.E, Wannemuehler, K.A. and 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.

Min, Y. and Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling* 5, 1–19.

Mwalili, S.,M., Lesaffre, E. and Declerck, D. (2008). The zero-inflated negative binomial regression model with correction for misclassification: An example in caries research. *Statistical Methods in Medical Research* 17, 123–139.

Ridout, M., Hinde, J. and 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.

Rose, C.E., Martin, S.W., Wannemuehler, K.A. and Plikaytis, B.D. (2006). On the Use of Zero-Inflated and Hurdle Models for Modeling Vaccine Adverse Event Count Data. *Journal of Biopharmaceutical Statistics* 16, 463–481.

Tin, A. (2008). Modeling zero-inflated count data with underdispersion and overdispersion. *SAS Global Forum*. Paper 372.

Vuong, Q.H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 57, 307–333.

Wang, K., Kelvin, K.W.Y. and Andy, H.L. (2002). A zero-inflated Poisson mixed model to analyze diagnosis related groups with majority of same-day hospital stays. *Computer Methods and Programs in Biomedicine* 68, 195–203.

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

Yeşilova, A., Kaya, Y., Kaki, B. and Kasap, İ. (2010). Analysis of plant protection studies with excess zeros using zero-inflated and negative binomial hurdle models. *GU Journal of Science* 23(2), 131-136.

Yeşilova, A, Kaydan, M.B. and Kaya, Y. (2010). Modeling insect-egg data with excess zeros using zero-inflated regression models. *Hacettepe Journal of Mathematics and Statistics* 39(2), 273–282.