

# Conway-Maxwell-Poisson (COM-Poisson) Regresyon Modeli: Aşırı ve Az Yayılımlı Verilere Uygulaması

## Conway-Maxwell-Poisson (COM-Poisson) Regression Model: Application to Over- and Under Dispersed Data

- Bahar ÇELİK<sup>a</sup>,
- Esin AVCİ<sup>a</sup>

<sup>a</sup> İstatistik Bölümü,  
Giresun Üniversitesi  
Fen Edebiyat Fakültesi,  
Giresun, TÜRKİYE

Received: 28.06.2019  
Received in revised form: 11.09.2019  
Accepted: 11.09.2019  
Available Online: 20.12.2019

Correspondence:  
Esin AVCİ  
Giresun Üniversitesi  
Fen Edebiyat Fakültesi,  
İstatistik Bölümü, Giresun,  
TÜRKİYE/TURKEY  
esinavci@hotmail.com

Bu makale Bahar Çelik'in Giresun Üniversitesi  
Fen Edebiyat Fakültesi Matematik Anabilim  
Dalında Fen Bilimleri Enstitüsü'ne bağlı olarak  
Yüksek Lisans Tezinden türetilmiştir.

**ÖZET Amaç:** Bağımlı değişkenin sayma/kesikli olma durumuyla birçok araştırmada karşılaşmaktadır. Her ne kadar Poisson regresyon modeli sayma verilerinin analizinde sıklıkla kullanılsa da ortalama ile varyansı eşitliği varsayımı birçok deneysel veri için sağlanmamaktadır. Aşırı/az yayılım olarak tanımlanan bu durumlarda alternatif regresyon modelleriyle analiz edilmesi daha uygun olmaktadır. Hem aşırı hem de az yayılım durumu için uygun analiz olanağı veren Conway-Maxwell-Poisson Regresyon modeli ayrıntılı olarak tanıtılmaktadır. **Gereç ve Yöntemler:** Poisson, negatif binom ve COM-Poisson regresyon modelleri önce aşırı yayılım gösteren 2011-2014 yılları arasında Giresun ili Toplum Ruh Sağlığı Merkezine (TRSM) kayıtlı olan hastaların, merkeze gelme sayılarına etki eden sosyo-demografik faktörlerin modellenmesinde, daha sonra az yayılım gösteren öğretmenlerin sosyal medya hesap sayılarına etki eden okul türü, cinsiyet ve yaş grubu faktörlerinin modellenmesinde uygulanmıştır. Her iki veri seti için değişken seçim yöntemi olarak tüm olası alt kümeler yöntemi kullanılmıştır. En küçük AIC değerli model en iyi model olarak seçilmiştir. **Bulgular:** TRSM'ye geliş sayılarına cinsiyet, eğitim düzeyi, medeni durum ile sosyal güvence durumunun, sosyal medya hesabına ise okul türü ve yaş grubunun etkili olduğu saptanmıştır. **Sonuç:** Her iki veri setinde de COM-Poisson regresyon en iyi modeli oluşturmuştur.

**Anahtar Sözcükler:** Aşırı-az yayılım; Conway-Maxwell-Poisson; sayma veri; negatif binom regresyon; Poisson regresyon

**ABSTRACT Objective:** Counting/discrete dependent variable is encountered in many studies. Although the Poisson regression model is frequently used in the analysis of counting data, the assumption of the equality of mean and variance is not provided for many experimental data. It is more appropriate to analyze with alternative regression models in case of over-under dispersion. The Conway-Maxwell-Poisson Regression model which is suitable for analysis of both over-under dispersed data is presenting in detail. **Material and Methods:** Poisson, negative binomial and COM-Poisson regression models were used to determine the effect of socio-demographic factors on the number of visits Giresun Community Mental Health Center (TRSM) between 2011 and 2014. Then, three regression models applied to model school type, gender, and age group factors with the number of social media accounts of the teachers. All possible subset regression approach was used as a variable selection method for both data sets. The smallest AIC-valued model was selected as the best model. **Results:** It was determined that gender, education level, marital status and social security status were effective on the number of arrivals to TRSM and the type of school and age group on social media account. **Conclusion:** COM-Poisson regression was the best model in both datasets.

**Keywords:** Over-under dispersion; Conway-Maxwell-Poisson; count data; negative binomial regression; Poisson regression

Bir bağımlı değişken ile bir veya birden fazla bağımsız değişken arasındaki neden-sonuç ilişkisinin ortaya çıkarılmasında regresyon modelinden yararlanılmaktadır. Regresyon modelinde bağımlı değişkenin dağılımı önem kazanmaktadır. Klasik regresyon modeli bağımlı değişkenin normal dağılıma sahip olması durumunda uygulanmaktadır. Ancak birçok bilimsel çalışmada normal dağılımın dışında farklı dağılıma sahip bağımlı değişkenlerle de karşılaştırılmaktadır. Bu durumda normal dağılımının yanı sıra üstel, Gamma, Poisson, Binom vb. dağılımları da kapsayan genelleştirilmiş lineer modellerden yararlanılmaktadır. Bu modeller ayrıca bağımlı değişkenin normal dağılım göstermediği durumlar için veri dönüşümüne alternatif olarak kullanılabilir. Buradan elde edilen sonuçların veri dönüşümüne göre daha dar güven aralığına sahip olduğu ve ters dönüşüme gerek olmaması nedeniyle daha doğru sonuçlar verdiği görülmüştür.<sup>1,2</sup> Genelleştirilmiş lineer modellerde varyans, ortalamanın bir fonksiyonudur.<sup>3</sup>

Sağlık çalışmalarında, biyoistatistik, aktüeryal bilimler ile demografide sıklıkla negatif olmayan tam sayılardan oluşan sayımla elde edilen bağımlı değişkenlerle karşılaşılmaktadır. Genel olarak bu tür değişkenlerin dağılımının Poisson dağılımına sahip olduğu varsayılarak ona etki eden değişken(ler) arasındaki ilişkinin analizinde Poisson regresyon modeli sıklıkla kullanılmaktadır. Ancak uygulamada sayımla elde edilen verilerde Poisson dağılımının varsayımı olan ortalamanın varyansa eşitliği çoğu zaman ihlal edilmektedir. Bu varsayımın ihlal edilmesi regresyon modelini oluşturan katsayıların standart hatalarının yanlış olması dolayısıyla da modelde yer almaması gereken değişken(ler) modelde yer almasına ve sonuç olarak da hatalı modelin oluşturulmasına neden olmaktadır. Varyansın ortalamadan büyük olması durumuna aşırı (over-dispersion), küçük olması durumuna da az (under-dispersion) yayılım adı verilmektedir. Varyansın ortalamadan büyük olması durumunda (aşırı yayılım) negatif Binom regresyon modelleri yaygın olarak kullanılmaktadır. Bir diğer yaklaşım ise ortalama ile varyans arasındaki fonksiyonel ilişki-den yararlanan yarı (Quasi) olabilirlik yöntemidir.

1962 yılında Conway ve Maxwell tarafından tanımlanan ve Shmueli ve ark. (2005) tarafından geliştirilen Conway-Maxwell-Poisson (COM-Poisson) dağılımı, Poisson dağılımının aşırı veya az yayılım gösteren verileri için genelleştirilmiş bir dağılımdır.<sup>4,5</sup> Ayrıca yayılım göstermeyen veriler için Poisson regresyon modeline eşdeğer sonuçlar vermektedir.

Aşırı veya az yayılım gösteren ile yayılım göstermeyen veriler için hesaplama kolaylığı sağlaması nedeniyle son yıllarda Conway-Maxwell-Poisson (COM-Poisson) regresyon modeli yaygın olarak kullanılmaya başlanmıştır. Sellers ve ark. (2012), COM-Poisson regresyon modelini işletme, ulaşım ve biyoloji gibi farklı alanlarda uygulamışlardır.<sup>6</sup> Santarelli ve ark. (2016), COM-Poisson regresyon modelinin pozitron emisyon tomografisinde (PET) görüntü rekonstrüksiyon algoritmalarında ve niceliksel PET veri analizinde kullanımının uygun olduğunu belirtmişlerdir.<sup>7</sup> Zhu (2012), COM-Poisson dağılımını zaman serisinde kullanarak yeni bir INGARCH modeli önermiştir.<sup>8</sup> Avcı (2018), Giresun ilinde Toplum Ruh Sağlığı Merkezine (TRSM) kayıtlı şizofren hastaların hastaneye yatış sayısını etkileyen demografik faktörleri Poisson, negatif Binom ve COM-Poisson ile incelemiştir.<sup>9</sup> Avcı (2018), alg sayısına etki eden mevsim ve istasyonları Poisson ve COM-Poisson ile karşılaştırmıştır.<sup>10</sup>

Bu çalışmada, COM-Poisson regresyon modelinin aşırı ve az yayılım gösteren veriler için uygunluğunun ortaya konulması amaçlandığından uygulama kısmında aşırı ve az yayılım gösteren iki veri seti analiz edilmiş ve Poisson ile negatif binom regresyon sonuçları ile karşılaştırılmıştır. Öncelikle aşırı yayılım gösteren 2011-2014 yılları arasında Giresun ili Toplum Ruh Sağlığı Merkezine (TRSM) kayıtlı olan hastaların, merkeze gelme sayılarına etki eden sosyo-demografik faktörlerin saptanması amaçlanmıştır. İkinci kısmında ise, az yayılım gösteren 2015-2016 yılları arasında Giresun Üniversitesi Bilimsel Araştırma Proje desteği ile (SOS-BAP-A-200515-43) gerçekleştirilen proje verilerinin bir kısmı olan öğretmenlerin sosyal medya hesap sayılarına etki eden okul türü, cinsiyet ve yaş grubu faktörleri incelenmiştir. Her iki veri seti için

değişken seçim yöntemi olarak tüm olası alt kümeler yöntemi kullanılmıştır. En küçük AIC değerli model en iyi model olarak seçilmiştir. Analizler R programının “MASS” ve “COMPOissonReg” paketlerinden ve “glm” fonksiyonundan yararlanılarak yapılmıştır.

## GEREÇ VE YÖNTEMLER

Bu bölümde regresyon analizine kısa bir giriş yapıldıktan sonra bağımlı değişkenin sayımla elde edilen verilerden oluşması durumunda sıklıkla kullanılan Poisson ve negatif binom regresyon modellerinden kısaca ve son olarak Conway-Maxwell-Poisson (COM-Poisson) dağılımı ile regresyon modelinden ayrıntılı bahsedilecektir.

Regresyon analizi, bir bağımlı değişken ile bir veya daha fazla bağımsız değişken arasındaki nedensel ilişkiyi tanımlama, tahmin etme veya kestirmek için kullanılan istatistiksel modeller sınıfı şeklinde tanımlanabilir. Bağımlı değişken  $Y$  ve bağımsız değişkenler kümesi  $X_1, X_2, \dots, X_k$  ( $k$  bağımsız değişken sayısıdır) olarak tanımlanır.  $Y$  ve  $X_1, X_2, \dots, X_k$  arasındaki ilişki,

$$Y = f(X_1, X_2, \dots, X_k) + \varepsilon \quad (1)$$

olarak ifade edilir. Buradaki  $\varepsilon$  terimi hata vektörüdür ve uyum eksikliğini ifade etmektedir.  $Y$  ve  $X_1, X_2, \dots, X_k$  arasındaki ilişkiyi tanımlamak için  $f(X_1, X_2, \dots, X_k)$  regresyon fonksiyonu kullanılabilir ve (1) eşitliği

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \varepsilon \quad (2)$$

şeklinde bir lineer regresyon modeli yazılabilir. Burada  $\beta_1, \beta_2, \dots, \beta_k$  regresyon parametreleri veya katsayıları olarak adlandırılır ve veriden tahmin edilen bilinmeyen sabitlerdir.

## POISSON REGRESYON

Bağımlı değişkenin Poisson dağılımına sahip olması durumunda Poisson regresyon modelinin kullanılması uygun olmaktadır. Poisson dağılımı, çoğunlukla nadir olayların oluş sayısının modellenmesinde kullanılan, belirli bir zamanda, alanda veya hacimde gerçekleşen olay sayısını tanımlayan bir dağılımdır. İlk defa 1837’de yayınlanan eserinde Siméon-Denis Poisson tarafından ortaya atılmıştır.<sup>11</sup> Poisson dağılımının ortalaması ile varyansı birbirine eşittir ( $E(Y_i) = V(Y_i) = \lambda_i$ ).<sup>12</sup>

$$y \sim \text{Poisson}(\lambda) \quad (3)$$

Poisson regresyon modeli kesikli ve negatif olmayan verileri içermesinden dolayı bağımlı değişkenin logaritması, bağımsız değişkenlerin doğrusal fonksiyonu ile bağlantılıdır. Genel anlamda model,

$$\log(E(Y_i)) = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k \quad (4)$$

ile tanımlanmaktadır. Bu modele istatistikte “log-linear” model adı da verilmektedir. Poisson regresyon modelinde, bağımlı değişkenin ortalaması ve varyansının birbirine eşit ile çok az sayıda sıfır değerine sahip olması, ayrıca gözlemlerin bağımsız ve gözlem sırasının eşit olma varsayımların sağlanması gerekmektedir.

## AŞIRI-AZ YAYILIM DURUMU

Temel olarak aşırı yayılım (over-dispersion) bağımlı değişkenin varyansının, bağımlı değişken dağılımının beklediğinden daha büyük olması durumudur. Benzer olarak, az yayılım (under-dispersion) bağımlı değişkenin dağılımının beklediğinden daha küçük olmasıdır.

Aşırı-az yayılım durumunda varyans fonksiyonu bir  $\Phi$  çarpımsal yayılım etkenine sahip olduğu varsayılmaktadır. Poisson için varyans fonksiyonu aşağıdaki gibi tanımlanmaktadır.

$$Var(y) = \Phi\mu \quad (5)$$

$\Phi > 1$  ise aşırı,  $\Phi < 1$  ise az ve  $\Phi = 1$  olması durumunda eş yayılım adı verilmektedir. Aşırı-az yayılım durumunda model parametrelerinin tahmin değerleri  $\Phi$  değerinden etkilenmez. Ancak model parametrelerinin standart sapması  $\Phi$  değerinden etkilenir.  $\Phi$  parametresi değeri biliniyorsa ya da bazı veri noktalarında tekrarlanıyorsa doğrudan belirlenebilir aksi takdirde ortalama sapma değeri ya da ölçeklendirilmiş Pearson ki-kare değeri  $\Phi$ 'ın nokta tahmin edicisi olarak kullanılır.<sup>2</sup>

Varyansın ortalamadan büyük (aşırı yayılım) ve bağımlı değişkende çok fazla sıfır olması durumunda Poisson regresyon analizi yerine negatif binom regresyon analizinin kullanılması gerekmektedir.<sup>13</sup>

### NEGATİF BİNOM REGRESYON

Negatif binom regresyon modelinde bağımlı değişkenin negatif binom dağılımına sahip olduğu varsayılmaktadır. Negatif binom dağılımı, Poisson dağılım parametresi olan ortalamanın Gamma dağılımına sahip olması varsayımına dayanmaktadır. Yani,

$$y | \lambda \sim \text{Poisson}(\lambda) \quad (6)$$

$$\lambda \sim \text{Gamma}(\alpha, \beta) \quad (7)$$

olur. Burada  $\lambda$ ,  $\alpha, \beta$  ortalamalı ve  $\alpha\beta^2$  varyanslı Gamma dağılımına sahiptir. Sırasıyla ortalama ve varyans,  $E(Y_i) = \lambda_i$  ve  $Var(Y_i) = \lambda_i(1 + \alpha\lambda_i^k)$  dir.

Negatif binom regresyon modeli, (4) eşitliği ile verilen Poisson regresyon modeli gibi tanımlanır.<sup>14</sup>

Aşırı veya az yayılım gösteren veriler için son yıllarda Conway-Maxwell-Poisson (COM-Poisson) regresyon modeli yaygın olarak kullanılmaya başlanmıştır.

### CONWAY-MAXWELL-POISSON (COM-POISSON) DAĞILIMI

Conway ve Maxwell tarafından 1962 yılında tanımlanan bir dağılımdır.<sup>4</sup> Conway-Maxwell-Poisson (COM-Poisson) dağılımı, Poisson dağılımının aşırı-az yayılım gösteren verileri için genelleştirilmiş bir dağılımdır. Dağılımın olasılık fonksiyonu, Poisson dağılımından yararlanılarak elde edilir. Poisson dağılımının özelliklerinden biri ardışık olasılıkların  $y$ 'ye göre lineer olmasıdır.

$$\frac{P(Y=y-1)}{P(Y=y)} = \frac{y}{\lambda} \quad (8)$$

Ancak bazı uygulamalarda bu oran  $y$ 'ye göre lineer olarak azalmamaktadır. Bu durumda dağılım Poisson dağılımından daha kalın veya ince kuyruklu bir dağılım yapısına dönüşmektedir. Bu nedenle (8) eşitliği aşağıdaki gibi tanımlanır.

$$\frac{P(Y=y-1)}{P(Y=y)} = \frac{y^u}{\lambda} \quad (9)$$

(9) ile tanımlanmış olan eşitliğin dağılımına Conway-Maxwell-Poisson (COM-Poisson) olarak adı verilir ve olasılık fonksiyonu aşağıdaki gibi verilmektedir.<sup>15</sup>

$$P(Y=y|\lambda, v) = \frac{\lambda^y}{(y!)^v} \frac{1}{\sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^v}} \quad y=0,1,2,\dots \quad (10)$$

Burada  $\lambda > 0$  ve  $v \geq 0$ 'dır.  $\frac{\lambda^j}{(j!)^v}$  serisi herhangi bir  $\lambda > 0$  ve  $v > 0$  için yakınsaklık göstermektedir.  $j \rightarrow \infty$  iken serideki iki ardışık terim için oran  $\frac{\lambda}{j^v}$  sıfıra yaklaşır. (10) eşitliğindeki payda,

$$Z(\lambda, v) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^v} \quad (11)$$

ile tanımlanır. Burada  $Z(\lambda, v)$ 'ye normalleştirme katsayısı ve  $v$ 'ye ise yayılım parametresi adı verilmektedir. Veri  $v < 1$  olduğunda aşırı,  $v > 1$  olduğunda az yayılım göstermektedir.  $v$ 'nün değerlerine bağlı olarak COM-Poisson bilinen bazı kesikli dağılımların genelleştirilmesidir. Bu dağılımlar;

- i.  $v = 1$  olduğunda  $Z(\lambda, v) = e^\lambda$  olur. Poisson ( $\lambda$ )
- ii.  $v \rightarrow \infty$  olduğunda ise  $Z(\lambda, v) \rightarrow 1 + \lambda$  olur. Başarı olasılığı  $\frac{\lambda}{1+\lambda}$  olan Bernoulli
- iii.  $v = 0$  ve  $\lambda < 1$  iken  $Z(\lambda, v) = \sum_{j=0}^{\infty} \lambda^j = \frac{1}{1-\lambda}$  olur. Geometrik

şeklinde olmaktadır.  $v = 0$  ve  $\lambda \geq 1$  iken  $Z(\lambda, v)$  yakınsaklık göstermediğinden herhangi bir dağılım tanımlanamamaktadır. COM-Poisson; ( $v = 0$ ) iken geometrik (ilk başarı sağlanıncaya kadar başarısız deneme sayısı), ( $v = 1$ ) Poisson ve ( $v \rightarrow \infty$ ) Bernoulli dağılımları arasında sürekli bir köprü görevi görmektedir. Benzer biçimde  $n$  tane bağımsız COM-Poisson rasgele değişkeni de bu üç bilinen kesikli dağılımlar arasında sürekli bir köprü olmaktadır.<sup>6</sup>

- i.  $v = 0$  ve  $\lambda < 1$  için geometrik dağılımlı  $n$  tane değişkenin toplamı,  $n$  ve  $1 - \lambda$  parametrelili negatif binom
- ii.  $v = 1$  için toplam  $n\lambda$  parametrelili Poisson
- iii.  $v \rightarrow \infty$  için  $n$  ve  $\lambda / (1+\lambda)$  parametrelili binom dağılımına dönüşür.

Shmueli vd. (2005) iki parametrelili COM-Poisson dağılımının tahmin edilmesinde üç farklı tahmin yöntemi önermiştir. Birincisi en çok olabirlik tahmin yöntemi, ikinci ağırlıklandırılmış en küçük kareler yöntemi ve üçüncüsü ise Bayesci tahmin yöntemidir.<sup>5</sup> Khan ve Khan (2010)'da parametre tahmini için Wedderburn (1974)'nün yarı (Quasi) tahmin yöntemini geliştirerek  $\lambda$  ve  $v$  parametrelerinin yarı (Quasi) tahminlerini elde etmişlerdir.<sup>16,17</sup>

## CONWAY-MAXWELL-POISSON (COM-POISSON) REGRESYON MODELİ

Sellers ve Shmueli (2010), McCullagh ve Nelder (1997)'in genelleştirilmiş lineer model yaklaşımına dayandırarak COM-Poisson regresyon modelini log link (bağıntı) fonksiyonunu kullanarak (4) eşitliğindeki gibi tanımlamışlardır.

$E(Y)$  ile  $X'\beta$  arasında lineer olmayan bir ilişki olduğundan  $\beta$  ve  $v$ ,  $\beta^{(0)}$  ve  $v^{(0)} = 1$  başlangıç değerleri kullanılarak normal denklemler yardımıyla iterasyonlu olarak yeniden ağırlıklandırılmış en küçük kareler ile çözümlenir veya optimizasyon programı kullanarak en çok olabirlik fonksiyonu maksimizasyonu yardımıyla tahmin edilir. Tahmin edilen katsayılar a ait standart hatalar Fisher bilgi matrisinden yararlanarak elde edilir.<sup>18,19</sup>

Genelleştirilmiş lineer modellere benzer olarak her bir bağımsız değişkenin anlamlılığı, yaklaşık standart normal dağılım yaklaşımından yararlanarak elde edilir. Küçük örneklem durumunda yine Bootstrap'tan yararlanarak ilgilenilen katsayının dağılımı elde edilebilmektedir.<sup>20</sup>

Sellers ve Shmueli (2010), veri analizinde yayılım parametresinin istatistiksel olarak anlamlı olup olmadığını test etmek için bir test istatistiği geliştirmişlerdir. Başka bir ifadeyle önerilen test, veri analizinde Poisson regresyon ( $v = 1$ ) ile COM-Poisson regresyon modelinin kullanımının uygunluğunu test edilmektedir ( $v \neq 1$ ). Test istatistiği aşağıdaki gibi tanımlanmaktadır,

$$C = -2\log\Lambda = -2[\log L(\hat{\beta}^{(0)}, \hat{v} = 1) - \log L(\hat{\beta}, \hat{v})] \quad (12)$$

Burada  $\Lambda$ , olabilirlik oran test istatistiğini,  $\hat{\beta}^{(0)}$  ise  $v = 1$  yokluk hipotezi (Poisson dağılımı) altında elde edilen en çok olabilirlik tahmin edicisini ve  $(\hat{\beta}, \hat{v})$  ise COM-Poisson dağılımı altında en çok olabilirlik tahmin edicisini göstermektedir. Yokluk hipotezinin doğruluğu altında  $C$  yaklaşık olarak 1 serbestlik dereceli  $\chi^2$  dağılımına sahiptir. Küçük örneklem için test istatistiğinin dağılımı Bootstrap yöntemi ile tahmin edilebilmektedir.<sup>18</sup>

COM-Poisson regresyonda, ortalamayla bağımsız değişkenler arasındaki ilişkinin ne toplamsal ne de çarpımsal olmasından dolayı, yani koşullu ortalamasının oranı bağımsız değişkenlerle lineer olmayan bir ilişki içerdiğinden ( $E(Y^v) = \lambda$ ) katsayıların yorumlanmasında ilgili katsayının  $v$ 'ye bölünerek yorumlanması önerilmektedir.<sup>18</sup>

#### AKAIKE BİLGİ KRİTERİ (AIC)

1973 yılında Akaike tarafından ortaya atılan bilgi kriteri yaklaşımına göre bir modelin uyumunu ölçmek için aşağıdaki formül kullanılmaktadır. Log-olabilirlik ve parametre sayısına dayanan bilgi kriteri ise,

$$AIC = -2\log L - 2q \quad (13)$$

şeklinindedir. Burada  $L$  tahmin edilen model için olabilirlik fonksiyonunun maksimum değeri,  $q$  ise modeldeki parametre sayısıdır. Bu yaklaşımda fazla parametre sayısına sahip model cezalandırılmaktadır. Uygun modelin seçiminde Akaike bilgi kriteri sıklıkla kullanılır.<sup>21</sup>

Çalışmada kullanılan veriler, 2011-2014 yılları arasında Giresun ili Toplum Ruh Sağlığı Merkezine (TRSM) kayıtlı olan tüm hasta bilgileri Giresun İli Kamu Hastaneleri Birliği Genel Sekreterliğinin 42991614/770 sayılı izini ve 2015-2016 yılları arasında Giresun Üniversitesi Bilimsel Araştırma Proje desteği ile (SOS-BAP-A-200515-43) gerçekleştirilen proje verileri, Giresun İl Millî Eğitim Müdürlüğü'nün 6315158 sayılı onayı ile ortaokul ve liselerden tabakalı örnekleme ile anket yöntemiyle derlenmiştir. Bilimsel araştırma tekniklerinden gözlemsel-tanımlayıcı araştırma tasarım düzeni uygulanmıştır. Analizler R programının "MASS" ve "COMPOissonReg" paketlerinden ve "glm" fonksiyonundan yararlanılarak yapılmıştır.

#### BULGULAR

Bu çalışmada sırasıyla aşırı yayılım gösteren 2011-2014 yılları arasında Giresun ili Toplum Ruh Sağlığı Merkezine (TRSM) kayıtlı olan hastaların, merkeze gelme sayılarına etki eden sosyo-demografik faktörlerin saptanması ve az yayılım gösteren 2015-2016 yılları arasında Giresun Üniversitesi Bilimsel Araştırma Proje desteği ile (SOS-BAP-A-200515-43) gerçekleştirilen proje verilerinin bir kısmı olan öğretmenin sosyal medya hesap sayılarına cinsiyet, yaş grupları (45 yaş altı) ve okul türü faktörlerinin incelenmesi amaçlanmıştır.

## TRSM VERİLERİ

2011-2014 yılları arasında Giresun ili Toplum Ruh Sağlığı Merkezinden (TRSM) 244 kayıtlı hasta bulunmaktadır. Bu hastaların TRSM'ye geliş sayısına etki ettiği düşünülen bağımsız değişkenler,

$X_1$ : Cinsiyet

$X_2$ : Eğitim Düzeyi

$X_3$ : Yaş

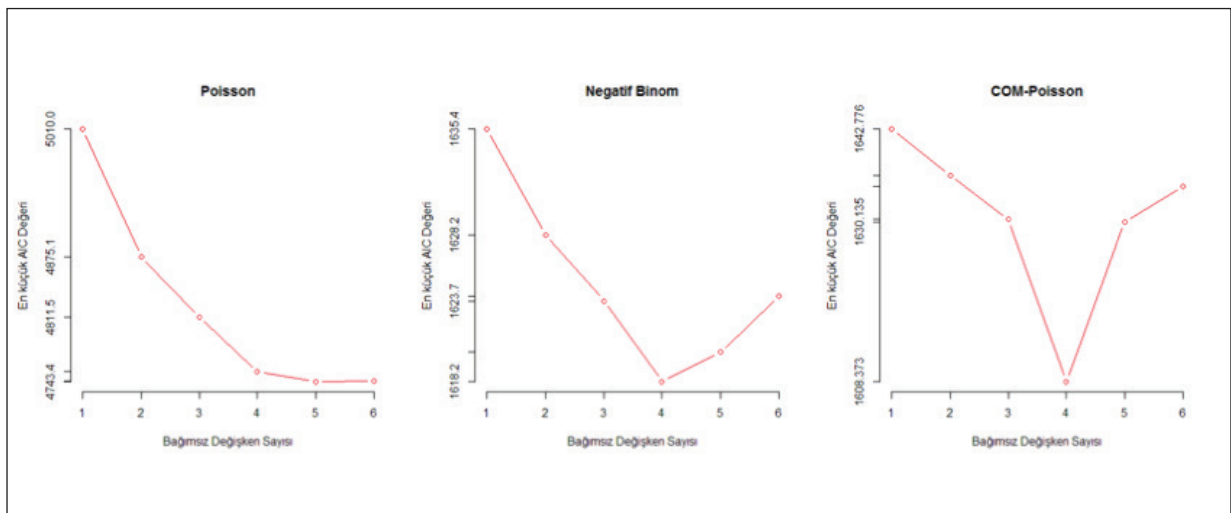
$X_4$ : Medeni Durum

$X_5$ : Sosyal Güvencesi

$X_6$ : Madde Bağımlılığı

şeklinde ele alınmıştır. TRSM'ye geliş sayısı ve yaş değişkeni dışında diğer değişkenler kategoriktir. TRSM'ye geliş sayısı ve yaş değişkeni ortanca ile kategorik değişkenler ise yüzde ile özetlenmiştir. Kategorik olan değişkenler sırasıyla cinsiyet 2, eğitim düzeyi 6, medeni durum 4, sosyal güvencesi 2 ve madde bağımlılığı 4 kategoriden oluşmaktadır. Geliş sayılarının ortancası 5 ve aralığı ise 0 ile 161 arasında, yaş değişkeni için ise ortanca değeri 44 ve aralığı ise 17 ile 78 arasında değiştiği saptanmıştır. TRSM'yi ziyaret eden hastaların yaklaşık olarak %63'ü erkek %37'si kadın, %9'u okur-yazar değil, %45'i ilkokul, %16'i ortaokul, %19'i lise, %4'ü önlisans ve %7'si lisans eğitim düzeyine sahip, %50'si bekar, %35'ü evli, %7'si boşanmış ve %8'i dul, %3'ünün sosyal güvencesi bulunurken ve %97'sinin sosyal güvencesi bulunmadığı, %42'sinin herhangi bir madde bağımlılığı yokken, %52'si sigara, %0,08'i alkol, %5'i hem sigara hem alkol kullandığı görülmüştür.

TRSM'ye geliş sayıları için ele alınacak bağımsız değişkenler, adimsal yöntemlere göre daha güvenilir ve araştırmacıya en iyi alt kümeyi seçme olanağı sağlaması nedeniyle tüm olası alt kümeler yöntemi tercih edilmiştir. Tüm olası alt kümeler yöntemi ele alınarak aşağıdaki modeller oluşturulmuştur. Bu modellerle sırasıyla Poisson, negatif binom ve COM-Poisson regresyon uygulanarak en uygun olanı seçilmiştir. Tüm olası altküme yöntemine göre modeldeki değişken sayısının en küçük AIC değerine karşı çizimi görsel olarak değişimi göstermektedir. Sırasıyla Poisson (Şekil 1.a), negatif binom (Şekil 1.b) ve COM-Poisson (Şekil 1.c) için grafikler aşağıda verilmiştir.



ŞEKİL 1: TRSM verisi için tüm olası altküme yöntemine göre modeldeki değişken sayısının en küçük Akaike Bilgi Kriteri (AIC) değerine karşı grafiği.

Şekil 1.a Poisson regresyon modeli için tüm olası altküme yöntemine göre modeldeki değişken sayısının en küçük AIC değerine karşı grafiğinden görüldüğü gibi modelde bağımsız değişken sayısı arttıkça AIC değerinde azalma görülmüştür. Azalma modelde 5 bağımsız değişken olduğunda en küçük AIC değerine ulaşmakta daha sonra yine artmıştır. Bu nedenle en uygun model 5 bağımsız değişkeni içeren model olmuştur. En küçük AIC değerini veren 5 bağımsız değişkenli model,

$$\ln E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 \quad (13)$$

olarak tanımlanmıştır. AIC değeri 4743.4 olarak elde edilmiştir.

Şekil 1.b’de negatif binom regresyon modelinde bağımsız değişken sayısı arttıkça AIC değerinde azalma olduğu ve bu azalmanın modelde 4 bağımsız değişken olduğunda en küçük AIC değerine ulaştığı daha sonra yine arttığı görülmüştür. En küçük AIC (1618.2) değerini veren 4 bağımsız değişkenli model;  $X_1$ ,  $X_2$ ,  $X_4$  ve  $X_5$  bağımsız değişkenlerini içeren model olmuştur. Poisson regresyonda anlamlı bulunan madde bağımlılığı ( $X_6$ ) negatif binom regresyonda anlamlı bulunmamış ve AIC değeri 4743.4’ten 1618.2’ye düşmüştür.

COM-Poisson regresyon modelinde Şekil 1.c’den görüldüğü gibi 4 bağımsız değişkenli modelin en küçük AIC değerine ulaştığı daha sonra yine arttığı saptanmıştır. En küçük AIC değerini veren modeli  $X_1$ ,  $X_2$ ,  $X_4$  ve  $X_5$  bağımsız değişkenlerinden oluşmaktadır. AIC değeri 1608.4’e düşmüştür.

İncelenen üç regresyon modeline ait katsayıların anlamlılıklarındaki farklılığın gözlemlenebilmesi için modele göre katsayıların sonuçları [Tablo 1](#)’de verilmiştir.

TABLO 1: TRSM verisi için regresyon katsayılarının sonuçları.						
Değişkenler	Poisson		Negatif Binom		COM-Poisson	
	Tahminler	Standart Hata	Tahminler	Standart Hata	Tahminler	Standart Hata
(Sabit)	1.263	0.113*	1.232	0.292*	-0.203	0.043*
Cinsiyet Erkek (referans)	-	-	-	-	-	-
Kadın	0.350	0.045*	0.514	0.159*	0.024	0.011*
Eğitim Düzeyi Okur-yazar değil (referans)	-	-	-	-	-	-
İlkokul	0.844	0.108*	0.798	0.286*	0.097	0.043*
Ortaokul	1.188	0.114*	1.200	0.325*	0.126	0.044*
Lise	1.274	0.111*	0.418	0.314*	0.129	0.043*
Önlisans	0.469	0.168*	0.539	0.469	0.057	0.063
Lisans	1.217	0.123*	1.335	0.379*	0.125	0.046*
Medeni Durum Bekar (referans)	-	-	-	-	-	-
Evli	-0.399	0.048*	-0.395	0.162*	-0.035	0.015*
Boşanmış	0.475	0.071*	0.535	0.303	0.026	0.016
Dul	0.078	0.076	0.016	0.286	0.008	0.021
Sosyal Güvence Var (referans)	-	-	-	-	-	-
S. G. Yok	-1.587	0.269*	-1.632	0.512*	-0.078	0.137*
Madde Bağımlılığı Yok (referans)	-	-	-	-	-	-
Sigara	0.1267	0.0439*	-	-	-	-
Alkol	-0.5037	0.2643*	-	-	-	-
Sigara ve Alkol	-0.0472	0.1036	-	-	-	-
Yayılim Parametresi	16.904		0.869		0.074	
Akaik Bilgi Kriteri (AIC)	4743.4		1618.2		1608.4	

\* %5 anlamlılık düzeyine göre istatistiksel olarak anlamlı, Conway-Maxwell-Poisson (COM-Poisson)



**Tablo 1**'den en küçük AIC değerini veren COM-Poisson modeli olduğu görülmüştür.  $\nu < 1$  olduğundan aşırı yayılım göstermiştir. Yayılım parametresi, anlamlılık testinde (3127.258,  $p < 0.05$ ) anlamlı bulunmuştur. %5 anlamlılık düzeyine göre COM-Poisson regresyon modeli için cinsiyet, eğitim düzeyi (önlisans alt kategorisi hariç), medeni durum (Boşanmış ve dul alt kategorileri hariç) ile sosyal güvence istatistiksel olarak anlamlı olduğu saptanmıştır.

Katsayılar, ilgili katsayının  $\nu$ 'ye (0.074) bölünüp üsteli alınmasıyla yorumlanmaktadır. Buna göre; kadınlar erkeklerle göre yaklaşık olarak 1,5 kat daha fazla TRSM'ye gelmiştir. Eğitim düzeyi bakımından karşılaştırıldığında referans kategori olan okur-yazar olmayanlara göre tüm eğitim düzeylerinin daha fazla TRSM'ye geldiği gözlenmiştir. Ancak Poisson regresyondan farklı olarak en fazla "Lise" (yaklaşık 6 kat), en az "İlkokul" (4 kat) mezunlarının daha sık TRSM'ye geldiği saptanmıştır. Poisson regresyondakine benzer olarak medeni durumu evli olanların bekarlara göre yaklaşık olarak yarı yarıya daha az TRSM'ye geldiği saptanmıştır. Sosyal güvencesi olmayanların olanlara göre yaklaşık olarak 3 kat daha az TRSM'ye geldikleri görülmüştür.

### SOSYAL MEDYA VERİLERİ

2015-2016 yılları arasında Giresun Üniversitesi Bilimsel Araştırma Proje desteği ile (SOS-BAP-A-200515-43) gerçekleştirilen proje verilerinin bir kısmı kullanılmıştır. Burada 231 öğretmenin sosyal medya hesap sayıları; cinsiyet, yaş grupları (45 yaş altı) ve okul türüne göre incelenmiştir. Bağımsız değişkenler aşağıdaki gibi tanımlanmıştır;

$X_1$ : Okul Türü

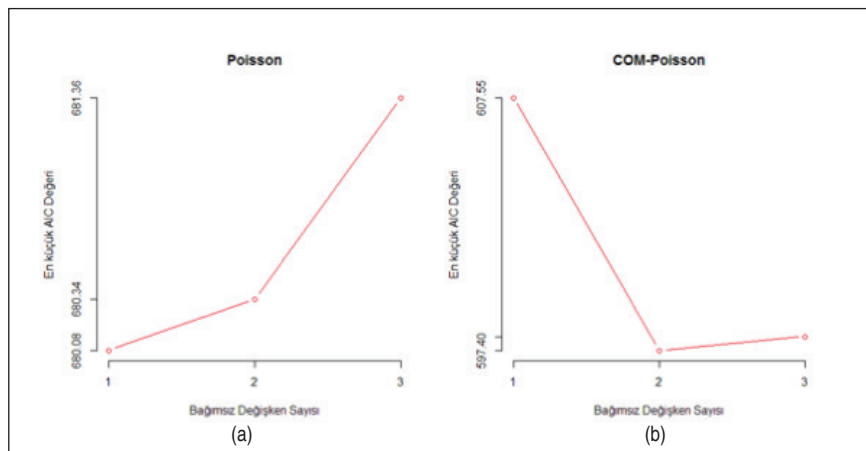
$X_2$ : Cinsiyet

$X_3$ : Yaş Grupları

İncelenen tüm bağımsız değişkenler kategoriktir ve yüzde ile özetlenmiştir. Öğretmenlerin yaklaşık olarak %99'u devlet, %1'i özel okulda, %39'u erkek, %61'i kadın, %3'ü "18-24", %22'si "25-29", %26'sı "30-34", %28'i "35-39" ve %21'i "40-44" yaş grubunda olduğu saptanmıştır.

Veri az yayılım gösterdiğinden tüm olası alt kümeler yöntemi sadece Poisson ve COM-Poisson regresyon modelleri için değişken sayısının en küçük AIC değerine karşı grafiği çizilmiştir (**Şekil 2**).

**Şekil 2.a**'dan görüldüğü gibi Poisson regresyon modeli için bağımsız değişken sayısı arttıkça AIC değerinde artma görülmüştür. En küçük AIC değerini veren sadece  $X_1$  bağımsız değişkenli model olmuş ve AIC değeri 680.08 olarak elde edilmiştir.



**ŞEKİL 2:** Sosyal medya verisi için tüm olası altküme yöntemine göre modeldeki değişken sayısının en küçük Akaike Bilgi Kriteri (AIC) değerine karşı grafiği.

**TABLO 2:** Sosyal medya verisi için regresyon katsayılarının sonuçları.

Değişkenler	Poisson		COM-Poisson	
	Tahminler	Standart Hata	Tahminler	Standart Hata
(Sabit)	0.660	0.048*	3.848	0.524*
Okul Türü Devlet (referans)	-	-	-	-
Özel	0.844	0.337	2.194	0.608*
Yaş Grubu 18-24 (referans)	-	-	-	-
25-29	-	-	-1.029	0.398*
30-34	-	-	-1.354	0.405*
35-39	-	-	-1.356	0.402*
40-44	-	-	-1.686	0.423*
Yayılm Parametresi	0.444		3.075	
Akaïke Bilgi Kriteri (AIC)	680.08		597.40	

\* %5 anlamlılık düzeyine göre istatistiksel olarak anlamlı, Conway-Maxwell-Poisson (COM-Poisson)

Şekil 2.b'den görüldüğü gibi modelde bağımsız değişken sayısı arttıkça AIC değerinde azalma görülmüş ve bu azalma modelde 2 bağımsız değişken olduğunda en küçük AIC değerine ulaşmakta daha sonra yine arttığı saptanmıştır. En küçük AIC değerini  $X_1$  ve  $X_3$  bağımsız değişkenli model olmuştur. Poisson regresyonda anlamsız bulunan yaş grupları ( $X_3$ ) COM-Poisson regresyonda anlamlı bulunmuş ve AIC değeri 680.08'den 597.40'e düşmüştür.

İki regresyon modeline ait katsayıların anlamlılıklarındaki farklılığın gözlemlenebilmesi için modellere göre katsayıların sonuçları Tablo 2'de verilmiştir.

Tablo 2'den en küçük AIC değerini veren COM-Poisson modeli olduğu görülmüştür.  $\nu > 1$  olduğundan az yayılım göstermiştir. Yayılm parametresi, anlamlılık testinde (84.379,  $p < 0.05$ ) anlamlı bulunmuştur. %5 anlamlılık düzeyine göre okul türü ve yaş grubu istatistiksel olarak anlamlıdır.

Katsayılar, ilgili katsayının  $\nu$ 'ye (3,075) bölünüp üsteli alınarak yorumlanmıştır. Buna göre özel okulda görev yapan öğretmenlerin devlet okulunda görev yapan öğretmenlere göre yaklaşık 2 kat daha fazla sosyal medya hesabı bulunduğu saptanmıştır. Yaş grubu bakımından karşılaştırıldığında referans kategori olan 18-24'a göre tüm yaş gruplarının daha az sosyal medya hesabı bulunduğu saptanmıştır. Yaş grubu arttıkça sosyal medya hesap sayısı azalığı görülmüştür. 40-44 yaş grubundakilerin 18-24 yaş grubuna göre sosyal medya hesap sayısı yaklaşık yarı yarıya düşmüş, 25-29 yaş grubundakilerin 18-24 yaş grubuna göre sosyal medya hesap sayısı yaklaşık %30 azalmıştır.

## TARTIŞMA VE SONUÇ

Poisson regresyon modeli, bağımlı değişkenin sayıma verisi olması durumunda sıklıkla kullanılmaktadır. Ancak birçok uygulamada ortalama ile varyans eşitliği varsayımı ihlal edilmektedir. Varyansın ortalama-dan büyük (aşırı) veya küçük (az) olması parametre tahmini üzerinde çok az bir etkisi olmasına rağmen standart hata anlamlılık testleri ve güven aralıklarının oluşturulmasında oldukça etkilidir. Aşırı yayılım durumunda sıklıkla negatif binom regresyon modeli kullanılmaktadır. Son yıllarda hem aşırı hem de az yayılım durumunda etkin sonuçlar veren Conway-Maxwell-Poisson (COM-Poisson) regresyon modeli uygulamaları yaygınlaşmaktadır.

Bu çalışmada aşırı ve az yayılım gösteren iki veri seti kullanılmıştır. Aşırı yayılım gösteren 2011-2014 yılları arasında Giresun ili Toplum Ruh Sağlığı Merkezine (TRSM) kayıtlı olan hastaların, merkeze gelme sayılarına etki eden sosyo-demografik faktörlerin saptanması amaçlanmıştır. TRSM'ye geliş sayılarına

cinsiyet, eğitim düzeyi, medeni durum ile sosyal güvence durumunun etkili olduğu saptanmıştır. İkinci veri seti ise, az yayılım gösteren 2015-2016 yılları arasında Giresun Üniversitesi Bilimsel Araştırma Proje desteği ile (SOS-BAP-A-200515-43) gerçekleştirilen proje verilerinin bir kısmı olan öğretmenlerin sosyal medya hesap sayılarına etki eden okul türü, cinsiyet ve yaş grubu faktörleri incelenmiş ve sosyal medya hesabına ise okul türü ve yaş grubunun etkili olduğu saptanmıştır.

Her iki veri seti için değişken seçim yöntemi olarak tüm olası alt kümeler yöntemi kullanılmıştır. En küçük AIC değerli model en iyi model olarak seçilmiştir. Uygulama yapılan iki veri setinde COM-Poisson regresyon en iyi modeli oluşturmuştur.

### **Finansal Kaynak**

*Bu çalışma sırasında, yapılan araştırma konusu ile ilgili doğrudan bağlantısı bulunan herhangi bir ilaç firmasından, tıbbi alet, gereç ve malzeme sağlayan ve/veya üreten bir firma veya herhangi bir ticari firmadan, çalışmanın değerlendirme sürecinde, çalışma ile ilgili verilecek kararı olumsuz etkileyebilecek maddi ve/veya manevi herhangi bir destek alınmamıştır.*

### **Çıkar Çatışması**

*Bu çalışma ile ilgili olarak yazarların ve/veya aile bireylerinin çıkar çatışması potansiyeli olabilecek bilimsel ve tıbbi komite üyeliği veya üyeleri ile ilişkisi, danışmanlık, bilirkişilik, herhangi bir firmada çalışma durumu, hissedarlık ve benzer durumları yoktur.*

### **Yazar Katkıları**

**Fikir/Kavram:** Esin Avcı; **Tasarım:** Bahar Çelik; **Denetleme/Danışmanlık:** Esin Avcı; **Veri Toplama ve/veya İşleme:** Esin Avcı; **Analiz ve/veya Yorum:** Esin Avcı, Bahar Çelik; **Kaynak Taraması:** Esin Avcı, Bahar Çelik; **Yazım:** Esin Avcı, Bahar Çelik; **Eleştirel İnceleme:** Esin Avcı.

## **KAYNAKLAR**

- Lewis SL, Montgomery DC, Myers RH. Examples of designed experiments with nonnormal responses. J Qual Technol. 2001;33(3):265-78. [Crossref]
- Mccullagh P, Nelder JA. Background an outline of generalized linear models. Generalized Linear Models. 2<sup>nd</sup> ed. London: Chapman & Hall; 1989. p.30-5. [Crossref] [PMC]
- Nelder AJ, Wedderburn RWM. Generalized linear models. J R Stat Soc A. 1972;135(3):370-84. [Crossref]
- Conway RW, Maxwell WL. A queueing model with state dependent service rates. Journal of Industrial Engineering. 1962;12:132-6.
- Shmueli G, Minka TP, Kadane JB, Borle S, Boatwright P. A useful distribution for fitting discrete data: revival of the conway-maxwell-poisson distribution. Appl Statist. 2005;54(1):127-42. [Crossref]
- Sellers KF, Borle S, Shmueli G. The COM-poisson model for count data: a survey of methods and applications. Appl Stoch Models Bus Ind. 2012;28:104-16. [Crossref]
- Santarelli MF, Della Latta D, Scipioni M, Positano V, Landini L. A conway-maxwell-poisson (CMP) model to address data dispersion on positron emission tomography. Comput Biol Med. 2016;77:90-101. PMID: 27522237. [Crossref] [PubMed]
- Zhu F. Modelling the series of counts with COM-poisson INGARCH models. Math Comput Model. 2012;56(9-10):191-203. [Crossref]
- Avcı E. Using count regression models to determine the factory which effects the hospitalization number of people with schizophrenia. Journal of Data Science. 2018;16(3):511-30.
- Avcı E. Flexibility of using COM-poisson regression model for count data. Statistics Optimization and Information Computing. 2018;6:278-85. [Crossref]
- Poisson SD. Recherches. Probabilité des Jugements en Matière Criminelle et en Matière Civile, Précédées des Règles Générales du Calcul des Probabilités. 1<sup>st</sup> ed. Paris, France: Bachelier; 1837. p.206.
- Deniz Ö. [Poisson regression analysis]. İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi. 2005;7:59-72.
- Sümbüloğlu K, Akdağ B. [Poisson regression analysis]. Hatipoğlu G, editör. İleri Biyoistatistiksel Yöntemler. 1. Baskı. Ankara: Hatipoğlu Yayınları; 2009. p.275-85.
- Saraçbaşı T, Aktaş Altunay S. Sayımla elde edilmiş veriler için regresyon modelleri. Kategorik Veri Çözümüleme. 1. Baskı. Ankara: Hacettepe Üniversitesi Yayınları; 2016. p.159.
- Winner G, Altmann G. Thesaurus of Univariate Discrete Probability Distributions. 1<sup>st</sup> ed. Germany: STAMM, Essen; 1999. p.865.
- Khan NM, Khan MHM. Model for analyzing counts with over-, equi- and under-dispersion in actuarial statistics. J Math & Stat. 2010;6(2):92-5. [Crossref]
- Wedderburn RWM. Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. Biometrika. 1974;61(3):439-47. [Crossref]
- Sellers KF, Shmueli G. A flexible regression model for count data. Ann Appl Stat. 2010;4(2):943-61. [Crossref]
- Sellers KF, Shmueli G. Data dispersion: Now you see it... Now you don't. Communication in Statistics: Theory and Methods. 2013;42(17):3134-47. [Crossref]
- Collett D. Strategy for model selection. Modelling Binary Data. 2<sup>nd</sup> ed. London: Chapman & Hall; 1991. p.240-4.
- Olsson M. Generalized Linear Models: An Applied Approach. Sweden: Student Literature; 2002. p.46-7.