Karadeniz Fen Bilimleri Dergisi, 9(2), 359-376, 2019. DOI: 10.31466/kfbd.644229

KFBD

Karadeniz Fen Bilimleri Dergisi
The Black Sea Journal of Sciences
ISSN (Online): 2564-7377

Araştırma Makalesi / Research Article

Sansürlü ve Sansürsüz Poisson Regresyon Modellerinin Karşılaştırılması

Öznur İŞÇİ GÜNERİ[1], Burcu DURMUŞ[1*]

[1]Muğla Sıtkı Koçman Üniversitesi, İstatistik Bölümü, Muğla, Türkiye

[*]**Sorumlu Yazar:** burcudurmus@mu.edu.tr

Geliş Tarihi: 07.11.2019
Kabul Tarihi: 13.12.2019

Öz

Poisson regresyon modeli, belli bir zaman periyodunda meydana gelen olaylara uygulanan bir regresyon modelidir. Bu modelde bağımlı değişken kesikli yani sayma verilerinden oluşur. Bu bakımdan regresyon modellerinin özel bir türüdür. Bunun yanı sıra Poisson regresyon modeli genelleştirilmiş doğrusal modeller arasında yer alır ve uygulamalarda en sık kullanılan yöntemlerden biridir. Bu model eşit yayılım gösteren veriler için uygulanmaktadır. Ancak çoğu zaman veri setleri Poisson modelinin varsayımlarını sağlamamaktadır. Bazen de veri seti hastalık, gözlemlenen kişinin ya da nesnenin kaybolması gibi nedenlerden dolayı sansürlü hale gelmektedir. Bu gibi bağımlı değişkenin sansürlü olması durumunda fazla veya az yayılım gösteren sayım verilerinin modellenmesi için sansürlü regresyon modelleri uygundur. Bu çalışmada sansürlü ve sansürsüz Poisson regresyon modelleri ele alınmıştır. Her iki model IRR (insidans oranı), uyum iyiliği ve bilgi kriterleri yardımıyla karşılaştırılmıştır. Çalışmanın sonucunda sansürleme yapılacak noktanın iyi seçilmesi durumunda sansürlü Poisson regresyon modelinin daha iyi sonuç verdiği gösterilmiştir.

Anahtar Kelimeler: Sayma Verileri, Sansürleme, Poisson Regresyon Modeli, Sansürlü Poisson Regresyon Modeli.

Comparison of Censored and Uncensored Poisson Regression Models

Abstract

Poisson regression model is a regression model applied to events that occur in a certain period of time. In this model, the dependent variable consists of discrete count data. In this respect, it is a special type of regression models. Besides, Poisson regression model is one of the generalized linear models and is one of the most commonly used methods in applications. This model is applied for data showing equal spread. However, often the data sets do not meet the assumptions of the Poisson model. Sometimes the data set becomes censored for reasons such as illness, loss of the person or object being observed. If the dependent variable is censored, censored regression models are suitable for modeling over- or under-dispersed count data. In this study, Poisson regression models uncensored and censored are discussed. Both models were compared with IRR (incidence rate ratio), goodness of fit and information criteria. As a result of the study, it is shown that the censored Poisson regression model gives better results if the point to be censored is selected well.

Keywords: Count Data, Censored, Poisson Regression Model, Censored Poisson Regression Model.

1. Giriş

Sayıma dayalı olarak elde edilen, gerçekleşme sayısı negatif olmayan tamsayı değerlerden oluşan veriler 'sayma verisi' olarak adlandırılır (Karaca ve Olmuş, 2018). Poisson regresyon modeli, bağımlı değişkenin sayma verilerinden oluştuğu durumlarda doğrusal regresyon analizine alternatif olabilen bir modeldir. Poisson regresyon modeli doğası gereği sağa çarpık bir dağılım göstermektedir. Bu durum, değişen varyans sorununu ortaya çıkarır. Poisson regresyonuna uyan bir modelde klasik regresyon analizinin uygulanması tahmin edilen katsayıların yansız olmasına neden olmaktadır (King, 1988). Dağılımın çarpıklığı bu durumun ortaya çıkmasının temel sebebidir.

Poisson regresyon modeli sayma verileri için en sık kullanılan ve en basit olan yöntemdir (Akın, 2002). Bu model ile sayımın olasılığı, Poisson dağılımı ile belirlenir. Sansürsüz sayma verilerinde, örnek ortalaması ile örnek varyansı arasındaki ilişki, dağılımının iyi bir ölçüsüdür. Eğer sayma verisinde ortalama ile varyans eşit ise Poisson regresyon analizi uygulanmaktadır. Bu durum ise gerçek hayatta nadir karşılaşın bir durumdur. Sansürlü bir sayma verisinde ise, ortalama ile varyans arasındaki ilişkinin bilinmesi çok düşük bir ihtimaldir. Sansürlü verilere geleneksel Poisson regresyon modelinin uygulanması, yanlı ve tutarsız tahminler üretecektir (Brännäs, 1992).

Çoğu zaman sayma verileri, varyansın ortalamadan daha küçük veya daha büyük olduğu yani sırasıyla az yayılım veya aşırı yayılım olarak sınıflandırıldığı önemli varyasyonlar sergiler. Az ya da aşırı yayılım durumlarıyla ilgili çeşitli modeller önerilmiştir (Wang ve Famoyea, 2002). Bağımlı değişkeni sayma verilerinden oluşan bir regresyon modelinde, bağımlı değişkenin değişim aralığı herhangi bir şekilde sınırlandırıldığında bağımsız değişkenler gözlenebiliyorsa sansürlü model söz konusu olmaktadır. Eğer bu sınırlama yoksa standart model şekline dönüşür. Böylece verinin ortalaması ve varyansı arasındaki ilişki durumuna göre sansürsüz durumda Poisson, negatif Binom ya da genelleştirilmiş Poisson regresyon modeli uygulanırken; sansürlü olması durumunda bunların sansürlü modelleri uygulanabilir. Tablo 1'de seçilmiş sayma modelleri için ortalama ve varyans arasındaki ilişki verilmiştir (Hilbe, 2014).

Tablo 1. Sayma Modeli Seçimi için Ortalama-Varyans İlişkisi

Model	Ortalama	Varyans
Poisson regresyon	λ	λ
Negatif Binom (NB1) regresyon	λ	$\lambda + \alpha\lambda$
Negatif Binom (NB2) regresyon	λ	$\lambda + \alpha\lambda^2$
Poisson Ters Gauss	λ	$\lambda + \alpha\lambda^3$
Negatif Binom-P regresyon	λ	$\lambda + \alpha\lambda^p$
Genelleştirilmiş Poisson regresyon	λ	$\lambda + 2\alpha\lambda^3\alpha^2\lambda^3$

Sansürleme literatürde yaşam analizi çalışmalarında yaygın olarak kullanılmaktadır. Genel olarak maliyet, süre gibi nedenlerden dolayı gözlenemeyen veya kesin olarak bilinemeyen hastaların yok olarak sayılması, gözden çıkarılması bu duruma örnek olarak verilebilir. Sansürleme durumunda belirli bir olay başarısızlık, ölüm, tepki, nüksetme, belirli bir hastalık gelişimi veya ayrılma sürelerinin ölçümlerinden oluşabilir.

Sansürlenmiş sayma verisi bağımsız değişkenin bilinen değerlerine karşılık, bağımlı değişkenin gözlemlerinin bazılarının gözlenememesidir. Sansürleme genel olarak sağdan ve soldan sansürleme olarak iki ana gruba ayrılır. Sağdan ve soldan sansürlemeler kullanılarak elde edilen aralık sansürlemesi ve ikili sansürleme çeşitleri de bulunmaktadır. Kesme noktasına göre sağdan ve soldan sansürleme aşağıdaki gibi tanımlanabilir (Hilbe, 2014):

- Sol sansürleme: $y \leq c$; örneğin c=3 ise sayma verilerinde 3'ten küçük veya ona eşit olan değerlere sansür uygulandığı anlamına gelir.
- Sağ sansürleme: $y \geq c$; örneğin c=15 ise sayma verilerinde herhangi bir cevabın 15'e eşit veya 15'den büyük olduğu kabul edilir. Bu değerin üstüne sansür uygulanarak c değerine göre yeniden hesaplama yapılır.

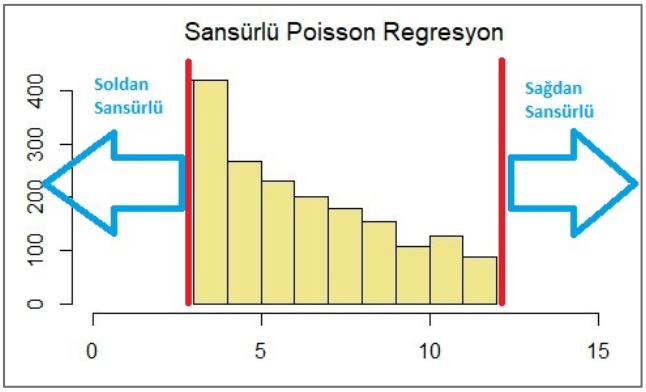
Sezgisel olarak veriler sağdan sansürlü olduğunda, bağımlı değişkenin büyük değerleri küçük olarak kodlanır ve bağımlı değişkenin koşullu ortalaması ile marjinal etkiler zayıflar (Raciborski, 2011). Sabit bir sansür eşiği ile yukarıdan sansürlenen bir sayım verisinde gözlenen ortalama, gerçek ortalamanın altında olacaktır. Ayrıca gözlemlenen varyans, gerçek varyanstan daha az olacaktır. Gözlenen ortalama, gözlenen varyanstan daha azsa gerçek ortalama gerçek varyanstan daha az veya daha fazla olabilir. Bu nedenle, sansürlü bir sayım verisinde dağılım türü bilenemeyebilir (Wang ve Famoyea, 2002). Sansürlü sayma verilerindeki ortalama ile varyans arasındaki gerçek ilişkinin bilinmemesi nedeni ile sansürlü sayma verilerinin analizi için sansürlü Poisson regresyon modeli önerilmiştir.

Son yıllarda yapılan çalışmalar incelendiğinde, Poisson regresyon analizinin farklı alanlarda uygulamalarının olduğu görülmektedir. Topaçoğlu ve Göztaşı (2019) sarıçam tohum bahçesinde ile yaprak alan indeksi ile göğüs çapı, kozalak verimi ve kalıtsallık ilişkisinin belirlenmesine yönelik yaptıkları çalışmada kozalak sayısı ile yaprak alan indeksi ve klonlar arasında istatistiki olarak anlamlı düzeyde bir ilişki olduğunu Poisson modeli ile belirlemişlerdir. Çalışmanın sonucu olarak tohum ve kozalak veriminin yüksek olması istenen tohum bahçelerinde yaprak alan indeksi değerini arttırıcı budama çalışmalarının tekniğinin ortaya konması gerektiğini ortaya koymuşlardır. Çankaya ve ark. (2017) midye parazit sayılarındaki sıfır yoğunluğunun kesikli regresyon modelleriyle açıklanması üzerine yaptıkları çalışmada, parazitlerin oluşma olasılıklarını en iyi sıfır ağırlıklı Poisson regresyon (Zero Inflated Poisson) ve engelli negatif Binom (Hurdle Negative Binom) modelleriyle ifade edildiğini göstermişler ve modellerde etkili faktörlerin bölgelerin çevresel

farklılıklarıyla ilişkili olduğunu ifade etmişlerdir. Taşkın ve ark. (2017) ise Poisson regresyon analizi ile su sıcaklığının gökkuşağı alabalıklarının büyümesi üzerindeki etkilerinin tahminini araştırmışlardır. Araştırmada, büyüme geriliği belirtileri gösteren balıklar belirlemişler ve Poisson regresyon analizi ile sıcaklık seviyelerinin balık büyümesi üzerinde olumsuz etkiler yarattığı sonucuna ulaşmışlardır. 2013 yılında Koç ve ark. tarafından yapılan bir başka çalışmada, aşırı yayılımlı veriler için genelleştirilmiş Poisson karma modellerin hava kirliliği üzerine bir uygulaması yapılmıştır. Çalışma sonuçlarına göre genelleştirilmiş Poisson modeli öksürüğü olan hasta sayılarının modellenmesinde Poisson regresyon modeline göre daha uygun bir model olarak bulunmuştur.

Literatürde sansürlü Poisson regresyon yönteminin kullanıldığı bazı çalışmalar şunlardır; Viwatwongkasem (2016) eksik veya gizli değerleri etkinleştirmek için yararlı olan beklenti-maksimizasyon (EM) algoritmasını kullanarak, frekans sayım verileri arasındaki maksimum olasılık tahminini (MLE) bulmanın iyi bir yöntem olacağını ileri sürmüştür. Çalışmasında ilaç kullanımında popülasyonun büyüklüğünü tahmin etmek için hem sıfır hem de kesikli (Truncated) ve sağdan sansürleme durumundaki verileri örneklemiştir. Sonuçlar, kesikli ve sansürlü bir Poisson olasılığının, sayısal olarak kararlı bir yakınsama, monoton artan bir olasılık ve yerel maksimum sağlayan EM algoritmasına karşılık gelen iyi tahminlerle iyi bir performans gösterdiğini, dolayısıyla MLE' nin beklenen genel maksimum değerinin başlangıç değerine bağlı olduğunu göstermiştir. Terza (1985) sabit eşik değerini kullanarak sayma verileri için sansürlü Poisson regresyon analizi uygulamış ve Newton-Raphson metodunu kullanarak en çok olabilirlik (ML) tahmincilerini elde etmiştir. Winkelmann ve Zimmermann (1995), sayma verilerinin modellenmesinde istatistiksel teknikleri ele almıştır. Bu tekniklerden bazıları Poisson, engelli (Hurdle) Poisson, kesikli (Truncated) Poisson ve negatif Binom modelleridir. Caudill ve Mixon (1995), değişken eşik değerinin temel olduğunu düşünmüştür. Sayma verileri aşırı dağıldığında sansürlü negatif Binom regresyon modelinin kullanılmasını önermiştir. Saffari, Adnan ve Greene (2012) sansürlenmiş örneklemler için Hurdle Poisson regresyon model parametrelerinin ML tahmin edicilerini elde etmiştir.

Şekil 1'de sağdan ve soldan sansürleme histogram üzerinde gösterilmiştir.



Şekil 1. Sağdan ve Soldan Sansürleme

Uygulamalarda en fazla karşılaşılan sansürleme türü sağdan sansürlemedir. Bu nedenle çalışmada sağdan sansürleme üzerinde durulmuştur. Örnek veri seti üzerinde sansürsüz ve sansürlü Poisson regresyon modeli için katsayı tahminleri, insidans oranı (IRR) ve farklı bilgi kriterleri ile modeller karşılaştırılmıştır.

2. Materyal ve Metot

2.1. Sansürlü Poisson Regresyon Modeli

Geleneksel Poisson regresyon modelinde, bağımlı değişken y_i negatif olmayan rastgele değişken ve $i = (1, 2, \dots, n)$ şeklinde değerler alır. Bu modelde y_i'nin Poisson dağılımı gösterdiği varsayılmaktadır. Poisson dağılımı için olasılık yoğunluk fonksiyonu Eşitlik 1'de verildiği gibidir:

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

Bu ifade de y_i, olayların meydana gelme sayısı, λ ise olayların tekrarlanmasının zaman birimi başına oranıdır. Başka bir deyişle λ, dağılımın ortalamasını verir. Poisson regresyon modelinin en belirgin özelliği ortalama ile varyansın birbirine eşit olmasıdır (Eşitlik 2).

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

Ancak uygulamada sayma değişkenler genellikle ortalamadan daha büyük varyansa sahip olduklarından aşırı yayılım gösterirler. Burada $\lambda_i = \exp(x_i{}^T\beta)$ sayma sayılarının ortalaması $\beta_1, \beta_2, \dots, \beta_k$ ise k boyutlu regresyon parametreleri vektörünü gösterir. Katsayıların tahmin edilmesi için farklı yaklaşımlar vardır. Bunlardan en yaygın kullanılanı en çok olabilirlik yaklaşımıdır;

$$L(\beta|(y,x)) = \sum_{i=1}^{n} P(y_i|\lambda_i) = \prod_{i=1}^{n} \frac{exp(-\lambda_i)\lambda_i^{y_i}}{y_i!} \tag{3}$$

Eşitlik 3'de verilen fonksiyonun logaritması alınırsa,

$$L(\beta) = \sum_{i=1}^{n} \{y_i \lambda_i - \lambda_i - lny_i!\} = \sum_{i=1}^{n} [-\exp(x_i{}^T \beta + y_i x_i{}^T \beta y_i!\} \quad (4)$$

Eşitlik 4 elde edilir ve bu fonksiyonun β katsayılarına göre türevi alınırsa Eşitlik 5 elde edilir.

$$\partial \ln(\beta) / \partial \beta_j = \sum_{i=1}^{n} [-\exp(x_i{}^T \beta) x_{ij} + y_i \, x_{ij}] \quad (5)$$

Buradan elde edilen k adet doğrusal olmayan Eşitlikler Newton-Raphson yöntemiyle veya yinelemeli ağırlıklı en küçük kareler yöntemi ile çözülerek katsayılar tahmin edilir.

Sayma verileri için uygulanan diğer bir yöntemde Genelleştirilmiş Poisson Regresyon (GPR) modelidir. GPR modelinde, bağımlı değişken y_i genelleştirilmiş Poisson rastgele değişkeni olsun. y_i etkileyen değişkenler $(x_1, x_2, \dots, x_{k-1})$ iken açıklayıcı değişkenler $(x_i = 1, x_1 x_2, \dots, x_{k-1})$'dir. y_i değişkeninin olasılık fonksiyonu şu şekildedir (Eşitlik 6) (Famoye, 1993; Wang ve Famoye, 1997);

$$f(y_i) = \left(\frac{\lambda_i}{1+\alpha\lambda_i}\right) \frac{(1+\alpha\lambda_i)^{y_i - 1}}{y_i!} \exp[-\frac{\lambda_i(1+\alpha y_i)}{1+\alpha\lambda_i}] \qquad y_i = 0, 1, 2, \dots \quad (6)$$

y_i'nin ortalama ile varyansı ise Eşitlik 7 ve 8'deki gibidir.

$$E(y_i|x_i) = \lambda_i \quad (7)$$

$$Var(y_i|x_i) = \lambda_i(1+\alpha\lambda_i) \quad (8)$$

Yukarıda verilen Eşitlik 6 ve 8'deki α, yayılım parametresi olarak adlandırılır. $\alpha = 0$ ise eşit yayılım söz konusudur. Pratikte bu durumla nadir olarak karşılaşılır. Eğer ortalama varyansa eşit değilse Poisson regresyon modeli ile elde edilen sonuçlar yanlı tahmin edilmiştir (Famoye ve ark., 2004).

$\alpha > 0$ olduğunda, varyans ortalamanın üzerindedir ve bu durum için, GPR modeli aşırı dağılıma sahip sayma verilerini temsil eder.

$\alpha < 0$ olduğunda, varyans ortalamanın altındadır ve bu durum için GPR modeli eksik dağılıma sahip sayma verilerini temsil eder.

Geleneksel Poisson regresyon modelinde tüm y_i değerleri gözlenir. Bununla birlikte sansürlü Poisson modelinde sansürleme noktası c_i'nin yalnızca altında kalan y_i^* değerleri gözlemlenir (Eşitlik 9). Sansürleme mekanizmasında sağ sansürlemenin anlamı sağdaki gözlemlerin kesildiğini gösterir. Burada sayımın c_i veya daha küçük olduğu, ancak c_i'den küçük sayının tam olarak gözlendiği bilinmektedir (Brännäs, 1992). Böylece,

$$y_i = \begin{cases} y_i^*, & if\ y_i^* < c_i \\ c_i, & if\ y_i^* \geq c_i \end{cases} \tag{9}$$

Eğer c sabit bir sayı ise sabit sansürleme eşiğine sahip modelimiz vardır (Terza, 1985). Sansürleme noktası c_i her gözlem için değişirse değişken sansürleme eşiğine sahip model söz konusudur (Caudill ve Mixon, 1995). y_i $g$özlemlerine sansür uygulanırsa Eşitlik 10 elde edilmiş olur.

$$\Pr(y_i \geq c_i) = \sum_{j=c_i}^{\infty} \Pr(y_i = j) = \sum_{j=c_i}^{\infty} \mathrm{f(j)} = 1 - \sum_{j=0}^{c_i-1} \mathrm{f(j)} = 1 - F(c_i - 1) \tag{10}$$

Burada d_i bir indikatör değişkeni olarak tanımlanırsa (Eşitlik 11),

$$d_i = \begin{cases} 1, & y_i^* \geq c_i \\ 0, & diğer\ durumlar \end{cases} \tag{11}$$

olarak gösterilebilir.

Sansürlü genelleştirilmiş Poisson regresyon (CGPR) modelinin olabilirlik fonksiyonu Eşitlik 12 ile verilmiştir.

$$L(\alpha, \beta; y_i) = \prod_{i=1}^{n} [p(\lambda_i; y_i)]^{1-d_i} [p(\lambda_i; y_i)]^{d_i} \tag{12}$$

Bu durumda Log-olabilirlik fonksiyonu ise Eşitlik 13'teki gibi yazılır.

$$logL(\alpha,\beta;y_i)=\sum_{i=1}^{n}\{(1-d_i)logp(\lambda_i;y_i)+d_i logp(\lambda_i;y_i)\} \quad (13)$$

Yukarıda verilen Eşitlik 10'da $y_i \geq c_i$ ve $\alpha = 0$ alındığında Eşitlik 13'teki sonuç, sabit bir sansür eşiği ile sansürlü Poisson regresyonuna indirgenir. Bu tür bir sansürleme, anket yoluyla elde edilen verilere uygulanabilir veya bazı teorik veya kuramsal kısıtlamaları yansıtabilir (Terza, 1985). Eşitlik 10'da $\alpha = 0$ ve $x_i \geq c_i$ ve $x_i \leq c_i$ olması durumunda ise Eşitlik 13'teki sonuç, sansürlü Poisson regresyonu değişken sansürleme eşikleriyle azalır (Caudill ve Mixon, 1995).

2.2. Katsayıların Tahmini

Katsayıların tahmini için en çok olabilirlik tahmin eşitliği aşağıda verilmiştir (Raciborski, 2011).

$$L(\beta)=log[\prod_{i=1}^{n}\{f(y_i)^{1-d_i}\}\{1-F(c_i-1)^{d_i}\}]=\sum_{i=1}^{n}[(1-d_i)logf(y_i)+d_i log\{1-F(c_i-1)\}] \quad (14)$$

Eşitlik 14'ün β'ya göre kısmi türevi alınırsa Eşitlik 15 elde edilir.

$$\frac{\partial L}{\partial \beta}=\sum_{i=1}^{n}\{(1-d_i)(y_i-\mu_i)-d_i c_i \emptyset_i\}x_i' \quad (15)$$

Eşitlik 15' in β'ya göre ikinci kısmi türevi alınırsa Eşitlik 16 olarak gösterilen Hessian değeri elde edilir.

$$\frac{\partial^2 L}{\partial \beta^2}=-\sum_{i=1}^{n}[(1-d_i)\lambda_i-d_i c_i\{(c_i-\lambda_i)\emptyset_i-c_i\emptyset_i^2\}]x_i'x_i \quad (16)$$

Yukarıdaki eşitlikler de verilen $\emptyset_i$ Eşitlik 17'deki gibi tanımlanır.

$$\emptyset_i = \frac{f(c_i)}{1} - F(c_i 1) \quad (17)$$

2.3. Modelin Uyum İyiliği

Belirli bir veri seti için birçok regresyon modeli mevcut olduğunda, bazı uygunluk ölçütlerine dayanarak alternatif modellerin performansı karşılaştırılabilir (Husain ve Bagmar, 2015). Literatürde uyumluluk ölçütleri olarak Pearson istatistiği, sapma istatistiği, AIC, BIC gibi çeşitli ölçütler yaygın olarak kullanılmaktadır.

2.3.1. Pearson İstatistiği

Seride aşırı yayılım olup olmadığının belirlenmesinde sıklıkla tercih edilen Pearson istatistiği, en temel uyum iyiliği ölçütlerinden biridir. Genel olarak Eşitlik 18'deki gibi ifade edilir.

$$P = \sum_{i=1}^{n} \frac{(y_i - \widehat{\lambda})^2}{\widehat{\omega}_i} \quad (18)$$

şeklinde ifade edilir. Pearson istatistiği sansürlü Poisson regresyonu için uygulandığında, Poisson dağılımının doğal bir uzantısı olarak $\omega_i = \lambda_i(1 + \alpha\lambda_i)$ olacaktır ve bu durumda formül Eşitlik 19 olarak değişecektir.

$$P_p = \sum_{i=1}^{n} \frac{(y_i - \lambda_i)^2}{\lambda_i(1+\alpha\lambda_i)} \quad (19)$$

Serideki aşırı ya da eksik yayılımın kontrolü ise *(n-k)* serbestlik derecesi olmak üzere aşağıda verildiği gibi kontrol edilir.

$P_p > n - k \Rightarrow seride\ aşırı\ yayılım\ vardır$

$P_p < n - k \Rightarrow seride\ eksik\ yayılım\ vardır$

2.3.2. Akaike Bilgi Kriteri (AIC)

Akaike bilgi kriteri farklı modeller arasından en uygununu seçmek amacıyla kullanılmaktadır. lnL tahmin edilecek modelin log-olabilirlik değeri ve k tahmin edilecek parametre sayısıdır. Buna göre mevcut modeller arasında AIC değeri Eşitlik 20 ile hesaplanır.

$$AIC = -lnL + k \tag{20}$$

AIC değeri küçük bulunan model uygun model olarak seçilir. Parametre sayısının örnek büyüklüğüne göre büyük olduğu durumlarda ise AIC yerine Hurvich ve Tsai tarafından önerilmiş olan AICc'nin kullanılması gerekir. Bu değer ise Eşitlik 21 ile hesaplanır (Akaike, 1973; Sugiuna, 1978; Hurvich ve Tsai, 1989).

$$AICc = AIC + 2k(k + \frac{1}{(n-k-1)}) \tag{21}$$

2.3.3. Bayes Bilgi Kriteri (BIC)

Akaike, doğrusal regresyonda seçilmiş model problemleri için BIC (Bayesian Information Criterion) model seçim kriterini türetmiştir. Bayes bilgi ölçütüne dair eşitlik aşağıdaki gibidir (Eşitlik 22). Farklı modeller içinde hesaplanan AIC değeri küçük bulunan model uygun model olarak seçilir.

$$BIC = -2log(L) + klog(n) \tag{22}$$

2.4. Veri Seti ve Tanımlayıcı İstatistikler

Bir sayma verisine Poisson regresyon modeli uygulandığında bağımlı değişkenin tüm değerleri bilinmektedir. Burada sansür ya da uç noktanın kesilmesi yoluyla sağdan sınırlandırılan bir örnekteki durumlar ele alınmaktadır. Çalışmada 2003 yılında yaşlı hastaların doktor ziyaretlerinin sayısı hakkında bilgi içeren ABD Tıbbi Harcama Paneli Araştırmasından (MEPS) elde edilen bir veri setini kullanarak Poisson ve sansürlü Poisson regresyon modelleri karşılaştırılmıştır. Bağımlı değişken

olarak doktor ziyaret sayıları alınmıştır. Bağımsız değişkenler; özel sigortalı olup olmadığı, sağlık sigortasına sahip olup olmadığı, cinsiyet, kronik durum ve yaş değişkenleri dikkate alınmıştır.

Veri setinin Poisson dağılımına uygunluğunu test etmek için aşağıda verilen hipotezin test edilmesi gerekir:

H_0: *Veriler Poisson dağılımı ile uygunluk gösterir*

H_1: *Veriler Poisson dağılımı ile uygunluk göstermez*

Eşitlik-1 ile verilen Poisson fonksiyonunda yer alan λ, Poisson dağılımının parametresidir. Bu değer bilinmediğinde λ parametresi, $\hat{\lambda} = \frac{\sum O_i y_i}{\sum O_i}$ O_i: *Gözlemlenen değerler* ifadesi ile tahmin edilir. Tahmin edilen $\hat{\lambda}$, Poisson dağılımı fonksiyonu ve gözlenen ve beklenen değerler yardımıyla $\chi^2_{test} = \sum \frac{(O_i - E_i)^2}{E_i}$ istatistiği test edilir (Kılıç, 2016).

Çalışmada kullanılan doktor ziyaretlerinin sayısı belirli bir zaman diliminde kaydedilmiştir. Başka bir ifade ile veri seti ana kütlenin kendisini değil, ana kütleye ait bir örneklemi oluşturmaktadır. Bu nedenle λ değerinin tahminine ulaşılarak yukarıda verilen hipotez test edilmelidir.

Poisson parametresi bilinmediğinden veri seti yardımıyla bu değer $\hat{\lambda}$=6.8226 olarak tahmin edilmiştir. Tahmin edilen bu değerden yola çıkılarak $\chi^2_{hesap} = 1{,}3849$ olarak bulunmuştur. Bu değer tablo değeri ile karşılaştırılırsa H_0 hipotezinin kabul edildiği görülür ($\chi^2_{hesap} = 1{,}3849 < \chi^2_{tablo} = 34{,}76$). Bu sonuçlara göre veri setinin Poisson dağılımı ile uygunluk gösterdiği yorumu yapılabilir.

Şekil 2 bağımlı değişken doktor ziyaret sayısının histogramını göstermektedir. Verilerin Poisson dağılımına uygunluk gösterdiği şekilden de görülmektedir. Bu nedenle önce Poisson regresyon analizi yapılmıştır. Daha sonra frekans dağılımından 12 ve üzeri sayılara sansür uygulanarak veriler yeniden modellenmiştir.

Şekil 2. Doktor Ziyaret Sayısı Frekans Dağılımı

Bağımlı değişken için sansürleme işleminin hangi noktadan itibaren yapıldığını göstermek amacıyla Tablo 2'de frekans tablosu verilmiştir. Tablo 2'de 12'den sonra frekansların giderek azaldığı görülmektedir. Bu nedenle kesme noktası olarak bu sayı alınmıştır.

Tablo 2. Frekans Tablosu

Doktor ziyaret sayısı	Sıklıklar	Oran	Kümülatif	Doktor ziyaret sayısı	Sıklıklar	Oran	Kümülatif
0	401	10.91	10.91	27	11	0.30	98.23
1	314	8.54	19.45	28	4	0.11	98.34
2	358	9.74	29.18	29	6	0.16	98.50
3	334	9.9	38.26	30	8	0.22	98.72
4	339	9.22	47.48	31	2	0.05	98.78
5	266	7.23	54.72	32	6	0.16	98.94
6	231	6.28	61.00	33	3	0.08	99.02
7	202	5.49	66.49	34	3	0.08	99.10
8	179	4.87	71.36	35	5	0.14	99.24
9	154	4.19	75.55	36	1	0.03	99.27
10	108	2.94	78.49	37	2	0.05	99.32
11	127	3.45	81.94	38	2	0.05	99.37
12	89	2.42	84.36	39	2	0.05	99.43
13	85	2.31	86.67	40	4	0.11	99.54
14	81	2.20	88.88	41	2	0.05	99.59
15	70	1.90	90.78	42	1	0.03	99.62
16	51	1.39	92.17	43	2	0.05	99.67
17	43	1.17	93.34	44	2	0.05	99.73
18	33	0.90	94.23	47	2	0.05	99.78
19	27	0.73	94.97	48	2	0.05	99.84
20	26	0.71	95.68	50	1	0.03	99.86
21	19	0.52	96.19	54	1	0.03	99.89
22	21	0.57	96.76	59	1	0.03	99.92
23	17	0.46	97.23	73	1	0.03	99.95
24	15	0.41	97.63	106	1	0.03	99.97
25	6	0.16	97.80	144	1	0.03	100.00
26	5	0.14	97.93				

Doktor ziyaret sayısını etkilediği düşünülen faktörler ve onlara ait tanımlayıcı istatistikler Tablo 3'te verilmiştir.

Tablo 3.a. Tanımlayıcı İstatistikler

Değişkenler	Gözlemler	Ortalama	Standart Sapma	Minimum	Maksimum
Doktor ziyaret sayısı	3677	6.822682	7.394937	0	144
Yaş	3677	7.424476	6.376638	65	90

Tablo 3.b. Tanımlayıcı İstatistikler

Değişkenler	Gözlemler (n)	Yüzde (%)
Özel Sigortalı	1826	49.7
Sağlık Sigortalı	613	16.7
Kadın	2210	60.1
Kronik durum		
0	620	16.9
1	1009	27.4
2	975	26.5
3	643	17.5
4	297	8.1
5	102	2.8
6	26	0.7
7	4	0.1
8	1	0.0

Bağımsız değişkenler için referans düzeyleri özel sigortaya sahip olma, sağlık sigortasına sahip olma, cinsiyetin kadın olması ve kronik durumun 0. kategoride olması değişkenlerine karşılık gelmektedir.

Çalışmadaki analizler için *Stata 14.0* programı kullanılmıştır. Analizin başında, bağımlı değişkene ilişkin ön tanımlayıcı istatistiklerden bazıları verilmiştir. Çalışmanın bulgular bölümünde ise her iki Poisson regresyon modeli tahmin edilmiştir. Bu modeller, uygunluk ölçüleri dikkate alınarak karşılaştırılmıştır.

3. Bulgular

Doktor ziyaret sayıları, Poisson dağılımı ile uygunluk gösterdiğinden bu çalışmada kullanılan modeller için uygun veri setini oluşturmaktadır. Bu amaçla elde edilen veriler üzerinden sağlık sigortalı olma, özel sigortalı olma, cinsiyet, yaş ve kronik durum değişkenleri üzerinden analiz edilerek en uygun Poisson modelinin belirlenmesi ve bu model üzerinden anlamlı değişkenlerin yorumlanması üzerine uygulama çalışması yapılmıştır.

3.1. Sansürsüz Poisson Regresyon Sonuçları

Sansürsüz Poisson regresyon modeline ilişkin tahminler Tablo 4'te verilmiştir.

Tablo 4. Poisson Regresyon Modelinin Parametre Tahminleri

Doktor ziyareti	Katsayılar (β)	Standart Hata	z	P>z	[95% Güven	Aralığı]
1.Özel Sigortalı	.164758	.0141057	11.68	0.000	.1371114	.1924047
1.Sağlık Sigortalı	.0705102	.018266	03.86	0.000	.0347094	.1063109
1.Kadın	-.0505733	.013126	-3.85	0.000	-.0762999	-.0248468
Kronik durum						
1	.5443098	.026792	20.32	0.000	.4917985	.5968212
2	.7956974	.0261698	30.41	0.000	.7444056	.8469892
3	1.105.729	.0265186	41.70	0.000	1.053753	1.157704
4	1.305.748	.0289406	45.12	0.000	1.249025	136.247
5	1.577.381	.0347898	45.34	0.000	1.509194	1.645568
6	1.464.105	.0588422	24.88	0.000	1.348776	1.579433
7	1.466.943	.1419643	10.33	0.000	1.188698	1.745188
8	.7290709	.4091883	1.78	**0.075**	-.0729234	1.531065
Yaş	.0032494	.0010004	3.25	0.001	.0012886	.0052103
Sabit terim	.8120431	.0773512	10.50	0.000	.6604376	.9636486

Poisson regresyon modeli incelendiğinde kronik durumu 8. kategoriye ait olan değişken anlamsız olarak bulunmuştur. Bu değişken dışındaki diğer açıklayıcı değişkenler ise istatistiksel olarak anlamlı bulunmuştur ($p < 0.05$). Doktor ziyaret sayısı *(dzs)* bağımlı değişken olmak üzere, Poisson regresyon modeli aşağıdaki şekilde yazılabilir (Eşitlik 23);

$$log(dzs) = 0.8120431 + 0.164758 \cdot (1.özel) + \ldots + 0.0032494 \cdot (yaş) \qquad (23)$$

Bu modelde katsayılar yarı elastikiyetler gibi yorumlanabilir. Örneğin; özel sağlık sigortası katsayısı için 0.16 değeri, özel sağlık sigortasına sahip olan biri özel sağlık sigortasına sahip olmayan birine göre 0.16 kat daha fazla doktor ziyareti yapması beklendiğini gösterir. Diğer katsayılar da benzer şekilde yorumlanır.

3.2. Sansürlü Poisson Regresyon Sonuçları

Sansürlü Poisson modeli sansür sınırlarının seçimine bağlıdır (Chen, 2016). Başka bir deyişle sansürlemenin yapılacağı noktanın doğru belirlenmesi, modelin anlamlı sonuçlar vermesi yönünde daha güçlü bir model elde edilmesini sağlayacaktır. Sansürlü Poisson regresyon modeline ilişkin tahminler Tablo 5'te verilmiştir. Sansürlü Poisson regresyon modeli incelendiğinde 664 gözlem sağdan sansürlüdür. Cinsiyet ve kronik durumu 8. kategoride olan değişkenler dışındaki diğer değişkenler istatistiksel olarak anlamlı bulunmuştur ($p < 0.05$). Sansürlü Poisson regresyon sonucuna göre cinsiyet değişkeni, sansürlemenin yapıldığı noktadan kaynaklı olarak anlamsız

bulunmuş olabilir. Yani kadın değişkenine ait verilerin bir kısmı sansürleme noktasının sağında kalmış olabilir. Sağlık sigortasına sahip olma değişkeni incelendiğinde $p = 0.045$ değeri ile değişkenin anlamlı olduğu görülmektedir. Ancak bu değer sansürsüz Poisson regresyon sonucuna göre yüksek bulunmuştur. Bu durumun sebebinin cinsiyet değişkeninde olduğu gibi sansürlemeden dolayı bazı verilerde yaşanan kayıplardan olduğu düşünülmektedir.

Tablo 5. Sansürlü Poisson Regresyon Modelinin Parametre Tahminleri

Doktor ziyareti	Katsayılar(β)	Standart Hata	z	P>z	[95% Güven	Aralığı]
1.Özel Sigortalı	.1468167	.0155301	9.45	0.000	.1163782	.1772552
1.Sağlık Sigortalı	.0409751	.0204281	2.1	0.045	.0009368	.0810134
1.Kadın	-.0003294	.0146473	-0.02	**0.982**	-.0290375	.0283787
Kronik durum						
1	.5290146	.0280074	18.89	0.000	.4741212	.583908
2	.743435	.0274959	27.4	0.000	.6895441	.797326
3	.9920289	.0282641	35.10	0.000	.9366323	1.047425
4	1.139355	.0318221	35.80	0.000	1.076.984	1.201725
5	1.301651	.0420727	30.94	0.000	121.919	1.384112
6	1.341425	.070531	19.2	0.000	1.203.187	1.479663
7	1.182095	.1828592	6.46	0.000	.8236981	1.540493
8	.7990987	.4093227	1.95	**0.051**	-.003159	1.601356
Yaş	.0036607	.0011089	3.30	0.001	.0014873	.005834
Sabit terim	.6921083	.0853078	8.11	0.000	.5249081	.8593086
	0 soldan-sansürlü gözlemler					
	3013 sansürsüz gözlemler					
	664 sağdan-sansürlü gözlemler					

Çalışmada yer alan veri setinde sansürleme noktası 12. frekans noktası olarak seçilmiş ve doktor ziyaret sayısı (dzs) bağımlı değişken olmak üzere, sansürlü Poisson regresyon modeli Eşitlik 24 ile ifade edildiği şekilde elde edilmiştir;

$$log(dzs) = 0.6921083 + 0.1468167 \cdot (1.\text{özel}) + \ldots + 0.0036607(\text{yaş}) \qquad (24)$$

Sansürlü Poisson modelindeki parametrelerin yorumlanması, sansürsüz modeldeki ile tamamen aynıdır. Örneğin; özel sağlık sigortası olan kişiler için doktor ziyaret sayısı, özel sigortası olmayanlara göre 0.14 kat daha yüksektir.

Bu çalışmada örnek veri seti üzerinden Poisson ve sansürlü Poisson regresyon modelleri karşılaştırılmıştır. Sansürlü Poisson regresyon modelinde Tablo 6'dan de görüldüğü üzere tüm β katsayıları ve dolaysıyla IRR değerleri daha düşük bulunmuştur. Sansürsüz Poisson modeli,

parametrelerin gerçek değerlerini varsayım bozulmalarından dolayı tutarlı bir şekilde elde edemeyebilir. Çünkü ortalama ve varyans arasındaki ilişki ve yayılım parametresi Poisson regresyonun uygulanabilmesi için önemlidir. AIC ve BIC değerleri incelendiğinde bu değerler sansürlü Poisson regresyon modeli için daha küçük bulunmuştur. Elde edilen sonuçlara göre sansürlü Poisson modelinin bu veri setinde oldukça başarılı olduğu görülmektedir.

Tablo 6. Poisson ve Sansürlü Poisson Regresyon Modellerinin Uyum İyiliği ve IRR Değerleri

	Poisson Regresyon		Sansürlü Poisson Regresyon	
Değişkenler	Katsayılar(β)	IRR	Katsayılar(β)	IRR
1.Özel Sigortalı	.164758	1.179108	.1468167	1.158.142
1.Sağlık Sigortalı	.0705102	1.073.055	.0409751	1.041.826
1.Kadın	-.0505733	.9506842	-.0003294	.9996706
Kronik durum				
1	.5443098	1.723419	.5290146	1.697259
2	.7956974	2.215986	.743435	2.103148
3	1.105.729	3.021426	.9920289	2.6967
4	1.305.748	3.690447	1.139355	3.124751
5	1.577.381	4.842258	1.301651	3.67536
6	1.464.105	4.32367	1.341425	3.824488
7	1.466.943	4.335958	1.182095	3.261201
8	.7290709	2.073154	.7990987	2.223536
Yaş	.0032494	1.003255	.0036607	1.003667
	.8120431	2.252505	.6921083	1.997923
AIC	30275.26		21253.05	
BIC	30355.99		21333.78	
LR chi2	4267.99		2498.37	

4. Sonuçlar ve Öneriler

Literatür incelendiğinde sayma verileri ile ilgili çalışmalarda Poisson regresyon modelinin sıkça kullanıldığı görülmektedir. Ancak veri setinin uygunluğu test edilmediği için sonuçlar bazen yanlı olarak yorumlanmıştır. Bu durumun önüne geçmek için tanımlayıcı istatistiklerin iyi incelenmesi, model uyum testinin yapılması ve sonuçların bilgi kriterleri, IRR değerleri gibi farklı istatistiki analizler ile desteklenmesi gerekmektedir. Bağımlı değişkeninin Poisson dağılışı gösterdiği veri setleriyle çalışılırken belli bir noktadan sonra verinin ya da çalışma alanının doğası gereği veriye sansür uygulanması gerekebilir. Burada önemli olan husus sansür sınırları tanımlandıktan sonra, modelde sadece doğru bilgiler saklanıyor mu veya herhangi bir ölçüm hatası var mı sorularının aranmasının gerekliliğidir.

Bu çalışma ile Poisson dağılımına sahip veri setinin nasıl ele alındığı, analizlerin hangi amaçla yapıldığı açıklanmıştır. Özellikle sağlık alanında sıkça karşılaşılan bir durum olan sansürlü model de çalışma kapsamında ele alınmış ve sansürsüz model ile sonuçları karşılaştırılmıştır. Yapılan bu çalışma ile sansürlü Poisson modelinin sansürsüz modele göre doktor ziyaret sayıları verileri için

daha iyi bir model sunduğu gösterilmiştir. Çalışma kapsamında hesaplanan IRR, model uyum iyiliği ve bilgi kriteri değerleri ile bulunan sonuçlar desteklenmiş ve sansürlü modelin sansürsüz modele göre daha güçlü sonuçlar verdiği ortaya konmuştur. Bu çalışmaya ek olarak, yapılan analizlerde aşırı yayılımın gözlemlendiği durumlarda negatif Binom, genelleştirilmiş Poisson gibi diğer modellerinde kullanılabileceği göz önünde bulundurulmalıdır.

Kaynaklar

Akaike, H., (1973). *Information Theory and an Extension of the Maximum Likelihood Principle*. 2nd International Symposium on Information Theory, 267-281.

Akın F., (2002). *Kalitatif Tercih Modelleri Analizi*. Bursa, Ekin Kitabevi.

Brännäs, K., (1992). *Limited Dependent Poisson Regression*. Statistician, 41, 413-423.

Caudill, S. B. ve Mixon, F. G., (1995). *Modeling Household Fertility Decisions: Estimation and Testing of Censored Regression Models for Count Data*. Empirical Economics, 20(2), 183-196.

Chen, X., (2016). *Censored-Poisson Model Based Approach to The Analysis of RNA-seq Data, Dissertation*. The Faculty of The Columbian, College of Arts and Sciences, The George Washington University.

Çankaya, E., Alpay, O. ve Özer, A., (2017). *Accounting for Zero Inflation of Mussel Parasite Counts Using Discrete Regression Models*. Sakarya University Journal of Science, 21(3), 378-384.

Hilbe, J., (2014). *Modeling Cout Data*. Cambridge University Press, 32 Avenue of the Americas, New York, NY 10013-2473, USA.

Hurvich, C. M. ve Tsai, C., (1989). *Regression and Time Series Model Selection in Small Samples*. Biometrika, 76, 297-307.

Famoye, F., (1993). *Restricted Generalized Poisson Regression Model*. Comm. Statist. Theory Methods, 22, 1335-1354.

Famoye, F., Wulu, J. ve Singh, K. P., (2004). *On The Generalized Poisson Regression Model with an Application to Accident Data*. Journal of Data Science, 2, 287-295.

Husain, M. ve Bagmar, S. H., (2015). *Modeling Under-dispersed Count Data Using Generalized Poisson Regression Approach*. Global Journal of Quantitative Science, 2(4), 22-29.

Karaca, A. G. ve Olmuş, H., (2018). *Sıfır Değer Ağırlıklı Verilerin Analizinde Sıfır Değer Ağırlıklı Regresyon Modellerin İncelenmesi*. Trakya Üniversitesi Sosyal Bilimler Dergisi, 20(2), 105-118.

Kılıç, S., (2016). *Ki Kare Testi*. Journal of Mood Disorders, 6(3), 180-183.

King, G., (1988). *Statistical Models for Political Science Event Counts: Bias in Conventional Procedures and Evidence for the Exponential Poisson Regression Model*. American Journal of Political Science, 32(3), 838-863.

Koç, H., Cengiz, M. A., Koç, T. ve Dünder, E., (2013). *Aşırı Yayılımlı Veriler İçin Genelleştirilmiş Poisson Karma Modellerin Hava Kirliliği Üzerine Bir Uygulaması*. International Anatolia Academic Online Journal, 1(2), 3-7.

Raciborski, R., (2011). *Right-Censored Poisson Regression Model*. The Stata Journal, 11(1), 95–105.

Saffari, S. E., Adnan, R. ve Greene, W., (2012). *Parameter Estimation on Hurdle Poisson Regression Model with Censored Data*. Jurnal Teknologi, 189-198.

Stata Corp LLC., (2019). *Sansürlü Poisson Regresyon Analizi*, https://www.youtube.com/watch?v=6m_SXthPv1U

Sugiuna, N., (1978*). Further Analysis of the Data by Akaike's Information Criterion and the Finite Corrections, Communication in Statistics,* Theory and Methods, 57, 13-26.

Taşkın, A., Karadavut, U. ve Bektaş, S., (2017). *Estimation of the Effect of Water Temperature on the Growth of the Rainbow Trout (Oncorhynchus mykiss) with Poisson Regression Analysis*, Greener Journal of Agricultural Sciences, 3(8), 663-668.

Terza, J. V., (1985). *A Tobit-Type Estimator for the Censored Poisson Regression Model*. Economics Letters, 18, 361-365.

Topaçoğlu, O. ve Göztaşı, S. M., (2019). *Sarıçam Tohum Bahçesinde Yaprak Alanı İndeksi (YAİ) ile Göğüs Çapı, Kozalak Verimi ve Kalıtsallık İlişkisinin Belirlenmesi*. Bartın Orman Fakültesi Dergisi, 21(1), 215-220.

Viwatwongkasem, C., (2016). *EM Algorithm for Truncated and Censored Poisson Likelihoods*, Procedia Computer Science, 86, 240-243.

Wang, W. ve Famoye, F., (1997). *Modeling Household Fertility Decisions with Generalized Poisson Regression*. J. Population Econom, 10, 273-283.

Winkelmann, R. ve Zimmermann, K. F., (1995). *Recent Developments in Count Data Modelling: Theory and Application*. J. Econom, 9, 1-24.