IJEASED

INTERNATIONAL JOURNAL OF EASTERN ANATOLIA SCIENCE ENGINEERING AND DESIGN

Uluslararası Doğu Anadolu Fen Mühendislik ve Tasarım Dergisi
ISSN: 2667-8764 , 2(1), 48-66, 2020
https://dergipark.org.tr/tr/pub/ijeased

Araştırma Makalesi / *Research Article*

Aşırı ya da Eksik Yayılım Durumunda Poisson ve Negatif Binom Regresyon Modellerinin Karşılaştırılması

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

[1] Muğla Sıtkı Koçman Üniversitesi, Fen Fakültesi, İstatistik Bölümü, Muğla, 48000, Türkiye.

Yazar Kimliği / *Author ID (ORCID Number)*	**Makale Süreci / *Article Process***	
[*]**Sorumlu Yazar** / *Corresponding author* : burcudurmus@mu.edu.tr	**Geliş Tarihi / *Received Date* :**	14.03.2020
https://orcid.org/0000-0003-3677-7121, Ö. İşçi Güneri	**Revizyon Tarihi / *Revision Date* :**	29.03.2020
https://orcid.org/0000-0002-0298-0802, B. Durmuş	**Kabul Tarihi / *Accepted Date* :**	04.04.2020
	Yayım Tarihi / *Published Date* :	15.07.2020

Alıntı / *Cite* : Güneri İşçi, Ö., Durmuş, B. (2020). Aşırı ya da Eksik Yayılım Durumunda Poisson ve Negatif Binom Regresyon Modellerinin Karşılaştırılması, Uluslararası Doğu Anadolu Fen Mühendislik ve Tasarım Dergisi, 2(1),48-66.

Özet

Bağımlı değişkenin sürekli olduğu durumlarda değişkenler arasındaki ilişki incelenirken En Küçük Kareler Yöntemi (EKKY) kullanılarak doğrusal regresyon analizi yapılmaktadır. Ancak bağımlı değişkenin kesikli ya da sayma verisi olması durumunda doğrusal regresyon modelleri kullanılarak yapılacak analizler etkisiz, tutarsız ve çelişkili sonuçlar verecektir. Bu nedenle sayma verileri için farklı regresyon modelleri geliştirilmiştir. Bunlar arasında en bilinen regresyon modelleri Poisson ve negatif binom regresyon modelleridir. Poisson regresyon modeli uygulamada, eşit yayılım durumunda kullanılmaktadır. Aşırı yayılım durumunda genelleştirilmiş Poisson regresyon modeli ya da negatif binom regresyon modeli tercih edilmektedir. Bu çalışma, aşırı yayılım durumunda Poisson ve negatif binom regresyon modellerinin analiz edilerek karşılaştırılmasını araştırmaktadır. Ampirik sonuçlar bağımlı değişkeninin aşırı yayılım göstermesi durumunda negatif binom regresyon modelinin daha iyi sonuçlar verdiğini göstermektedir. Bunu doğrulamak için her iki model AIC, BIC ve G^2 bilgi kriterleri ile karşılaştırılmıştır. Ayrıca modellerin katsayılarının yorumlanması için marjinal etkiler ve insidans oranı (IRR: Incidence Ratio Rate) değerleri hesaplanmıştır. Sonuç olarak, Poisson regresyon ile analiz yapılacak durumlarda aşırı yayılımın varlığı kontrol edilmeli, var olduğu durumlarda negatif binom regresyonu ile analize devam edilebileceği göz önünde bulundurulmalıdır.

Anahtar Kelimeler: Sayma verileri, Poisson regresyon, Negatif binom regresyon, Marjinal etki, IRR.

ISSN: 2667-8764
https://dergipark.org.tr/tr/pub/ijeased

Comparison of Poisson and Negative Binomial Regression Models in Case of Over-Dispersed or Under-Dispersion

Abstract

When the dependent variable is continuous, the linear regression analysis is performed by using Least Squares Method (OLS). However, if the dependent variable is discrete or count data, analysis using linear regression models will yield ineffective, inconsistent and contradictory results. Therefore, different regression models have been developed for count data. Among these, the best known regression models are Poisson and negative binomial regression models. Poisson regression model is used in case of equal dispersed in the application. In case of over-dispersed, generalized Poisson regression model or negative binomial regression model is preferred. This study investigates the comparison of Poisson and negative binomial regression models in case of over-dispersed. Empirical results show that the negative binomial regression model gives better results if the dependent variable shows over-dispersed. To confirm this, both models were compared with the AIC, BIC and G^2 information criteria. In addition, marginal effects and incidence ratio (IRR: Incidence Ratio Rate) values were calculated to interpret the coefficients of the models. As a result, the presence of over-dispersed should be checked in cases to be analysed by Poisson regression and it should be taken into consideration that the analysis can be continued with negative binomial regression when it exists.

Keywords: *Count data, Poisson regression, Negative binomial regression, IRR, Marginal effect.*

1. Giriş

Bağımlı değişken sürekli olduğunda iki ya da daha çok değişken arasındaki ilişkiyi ölçmek için kullanılan regresyon analizi, en temel istatistiksel analizlerden birisidir. Her bir değişkenin bağımlı değişkeni nasıl etkilediği regresyon katsayısı ile ifade edilir. Burada amaç, bağımlı değişken ile bağımsız değişkenler arasında sebep sonuç ilişkisi bulmaktır.

Bağımlı değişkenin kesikli değer aldığı fakat kategorik olmadığı durumlar vardır. Bu tür durumlara sayma verileri denilmektedir. Sayma verileri, uygulamada genelleştirilmiş doğrusal modeller arasında yer almaktadır. Sayma sonuçlarının özelliklerini kesin olarak veren birçok model vardır. Ancak Poisson regresyon birçok analizin başlangıç noktası olarak düşünülür. Poisson regresyon modeli sayma verileri için en sık kullanılan ve en basit olan yöntemdir. Poisson regresyon modelinde, bağımsız değişkenlerin doğrusal yapısını bağımlı değişkenin beklenen değerine bağlayan link fonksiyonu logaritmiktir. Bu model ile sayımın olasılığı, Poisson dağılımı ile belirlenir. Modelin en belirgin özelliği, sonucun koşullu ortalamasının koşullu varyansına eşit olmasıdır (Deniz, 2005). Ancak uygulamada bazen koşullu varyans, koşullu ortalama değerini aşabilir.

Poisson dağılımında, varyansın ortalamadan büyük olması haline aşırı yayılım (overdispersion) ve varyansın ortalamadan küçük olması haline az yayılım (underdispersion) denilmektedir (Cox, 1983). Bağımlı değişkende aşırı yayılım olması durumunda genellikle iki yol izlenmektedir. Bunlardan birincisi, bir yayılım parametresi tahmin ederek (α) bununla test

istatistikleri ve artıkların düzeltilmesidir. İkincisi ise aşırı yayılımın etkisini gideren yöntemlerden negatif binom regresyon modelinin uygulanmasıdır (Hilbe 2007). Uygulamada negatif binom regresyon modelinin yaygın kullanıldığını, bunun dışında genelleştirilmiş Poisson regresyon modeli ve Poisson quasi- Lindley regresyon modelinin de kullanıldığını da görmekteyiz. Veri setinde, aşırı yayılım olup olmadığını belirlemek için sapma (deviance) uyum iyiliği istatistiği yaygın olarak kullanılmaktadır.

Poisson regresyon modeli bağımlı değişkenin sayma verilerinden oluştuğu durumlarda doğrusal regresyon analizine alternatif olabilen bir modeldir. Bu sebeple aktüeryal bilimler, biyoistatistik, demografi, iktisat, politik bilimler ve sosyoloji gibi pek çok alanda kullanım imkânı bulabilmektedir.

Sayma regresyon modelleri geçmişten günümüze birçok alanda kullanım imkânı bulmuştur. King (1988) Amerika Birleşik Devletleri 'de temsilciler meclisi üyelerinin 1802-1876 tarihleri arasında parti değiştirme davranışlarını analiz etmiştir. Bağımsız değişken olarak bir yılda parti değiştiren temsilciler meclisi üye sayısı kullanılmıştır. Michener ve Tighe (1992) Amerika Birleşik Devletleri'nde otobanda meydana gelen ölümcül kazaları incelemiştir. Poisson regresyon modelini kullanarak Khalat vd. (1997) savaş döneminde Beyrut'ta doğurganlık düzeyleri farklılıklarını, Burg vd. (1998) akademik işgücü piyasasında erkek ve kadın akademisyenlerin yükselmelerini incelemiştir. Şahin (2002), 1964-1998 dönemi ve Arısoy ve Yaprak (2016), 1984-2015 dönemi Türkiye'deki grevlerin belirleyicileri için Poisson regresyon uygulaması yapmıştır. Memiş ve Önder (2018) yapay veri ile Poisson regresyon tahmin yöntemlerini karşılaştırmışlardır.

Bu çalışmada, bağımlı değişkenin sayma verilerinden oluştuğu durumlarda sıkça tercih edilen Poisson regresyonu ve Poisson modellerinin bir genellemesi olan negatif binom regresyon modelleri örnek bir seti ile detaylı açıklanmaya çalışılmıştır. Bu amaçla *Stata 13.0* programı ile model analizleri yapılmıştır (Stata, 2019). Her iki regresyon modeli için elde edilen katsayılar ile modellerin marjinal etkileri ve IRR değerleri açıklanmıştır.

2. Materyal ve Metot

2.1. Poisson Regresyon Analizi

Poisson regresyonu, genelleştirilmiş doğrusal model ailesine ait regresyon analizlerinden biridir. Bu yöntem, çoklu regresyon yöntemine benzemektedir. Ancak bağımlı değişken (Y) sadece Poisson dağılımı gösteren, negatif ve kategorik olmayan ve 0, 1, 2, 3 gibi sayma sayıları değerlerini

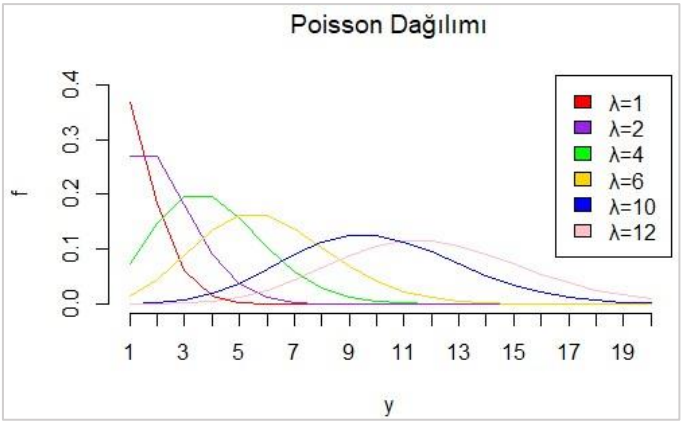
alır. Bundan dolayı Poisson regresyon, lojistik regresyona da benzer. Fakat lojistik regresyonda bağımlı değişkenler belirli değerlerle sınırlıdır.

Sayma veri modelleri regresyon modellerinin özel bir türüdür. Bu nedenle bu modeller için farklı çözüm yöntemleri gerekmektedir. Poisson ve negatif binom regresyon modelleri sıklıkla kullanılan yöntemlerdir. Her iki yöntemde de bilinen regresyon modellerinde olduğu gibi bağımsız değişkenler ile sayım veri niteliğindeki bağımlı değişken arasındaki ilişkiyi araştırmak amaçlanmaktadır (Cameron ve Trivedi, 2013).

Poisson regresyon analizi, bağımlı değişken Y_i'nin Poisson dağılımı gösterdiğini varsaymaktadır. λ parametreli Poisson dağılımı için olasılık yoğunluk fonksiyonu aşağıdaki formülde verildiği gibidir (Denklem 1):

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

Bu ifade de Y_i, olayların meydana gelme sayısı, λ ise olayların tekrarlanmasının zaman birimi başına oranıdır. Başka bir deyişle λ, dağılımın ortalamasını verir. Buradaki olasılık, λ değerinin bir fonksiyonu olarak değişir. Poisson olasılık dağılımı sağa eğiktir. Fakat λ_i büyüdükçe dağılım normal dağılıma yaklaşır. Şekil 1'de farklı λ_i değerlerini alarak çizdiğimiz dağılımın değişimi görülmektedir.



Şekil 1. Poisson olasılık dağılımı

EKKY yönteminde olduğu gibi Poisson regresyon modelinin de bazı varsayımları vardır; bağımlı değişkenin sayma verisi olması, gözlemlerin birbirinden bağımsız olması, ortalama ile varyansın birbirine eşit olması (Denklem 2) ve $log(\lambda)$'nın x'in doğrusal bir fonksiyonu olması

(Legler ve Roback, 2019). Fazla veya az dağılmış veri setleri Poisson dağılımı ile modellenemez. Çünkü koşullu beklenen değerin varyansa eşit olduğu varsayımında bozulmalar görülür ve varsayım sağlanmaz. Bu durumda veri setinin güncellenmesi veya farklı yöntemler ile analize geçilmesi bir çözüm olabilir.

$$\lambda_i = E(Y_i|x_i) = Var(Y_i|x_i) \quad (2)$$

Uygulamalarda sayma değişkenler genellikle ortalamadan daha büyük varyansa sahip olduklarından aşırı yayılım gösterirler. Verinin aşırı yayılım göstermesi; gözlemlenen sıfır değerlerin sayısının Poisson modeli ile ortaya konulan sıfır değerlerini aşması ve gözlenmemiş heterojenlik gibi durumlara neden olmaktadır (Kibar, 2008). Modeldeki aşırı yayılım katsayı tahminini etkilemez, ancak tahminin standart hatasının etkisi altında olmaya sebep verir, böylece modelin güvenirliğini yükseltir (Al-Ghirbal ve Al-Ghamdi, 2006). β'lar modelde katsayıları göstermek üzere Poisson regresyon modeli Denklem 3 ile verilmektedir.

$$log(\lambda_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \dots, \beta_m x_m \quad (3)$$

Yukarıdaki eşitlikten λ_i değerini Denklem 4 ile yazabiliriz.

$$\lambda_i = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 \dots, \beta_m x_m) \quad (4)$$

Poisson regresyon analizinde β tahmincilerini hesaplamak için pek çok yöntem vardır. En çok olabilirlik yöntemi (MLE: Maximum Likelihood Estimation Method), yapay en çok olabilirlik yöntemi (AMLE: Artificial Maximum Likelihood Method) ve genelleştirilmiş doğrusal modeller (GLM: Generalized Linear Models) bu yöntemlerin en bilinenleridir.

2.2. Negatif Binom Regresyon Analizi

Negatif binom regresyonu, varyansın Poisson modeli tarafından hesaplanan ortalamaya eşit olduğu ve kısıtlayıcı varsayımı gevşeten Poisson regresyonunun bir genellemesidir. Bu model Poisson-Gama karışımı bir dağılıma dayanmaktadır.

Poisson dağılımı, ortalaması *1* ve ölçek parametresi *v* olan bir gama gürültü değişkeni dahil edilerek genelleştirilebilir. α yayılım parametresi olmak üzere elde edilen Poisson-Gama karışımı (negatif binom) dağılımı Denklem 10 ile ifade edilir.

$$P(Y_i|\lambda_i,\alpha) = \frac{\Gamma(Y_i+\alpha^{-1})}{\Gamma(Y_i+1)\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda_i}\right)^{\alpha^{-1}}\left(\frac{\lambda_i}{\alpha^{-1}+\lambda_i}\right)^{Y_i} \qquad \lambda_i = t_i\lambda\,,\ \ \alpha = \frac{1}{v} \tag{10}$$

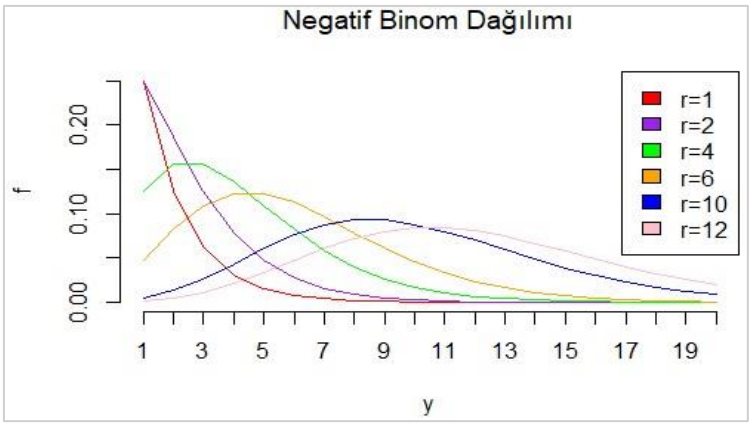
Negatif binom modeli için varyans Denklem 11 biçimindedir.

$$Var(Y_i|x_i) = \lambda_i + \alpha{\lambda_i}^2 \tag{11}$$

Bu modele göre negatif binom regresyon modeli t_i, maruz kalma süresi ve $\beta_1, \beta_2, \ldots, \beta_k$, bilinmeyen parametreler olmak üzere Denklem 12 ile gösterilir.

$$\lambda_i = \exp(\ln(t_i)\,\beta_{1i}x_{1i} + \beta_{2i}x_{2i}, \ldots, \beta_{ki}x_{ki}) \tag{12}$$

Regresyon katsayıları, en çok olabilirlik yöntemi kullanılarak tahmin edilebilir (Cameron ve Trivedi, 2013). Negatif binom dağılımının Poisson dağılımdan farklı olarak bir parametresi daha vardır. Bu nedenle ikinci parametre, varyansı ortalamadan bağımsız olarak ayarlamak için kullanılabilir. Şekil 2'de farklı değerlere göre çizdirdiğimiz negatif binom olasılık dağılımları görülmektedir.



Şekil 2. Negatif binom olasılık dağılımı

Negatif binom regresyon modellerinin katsayılarını tahmin etmek için farklı yöntemler geliştirilmiştir. Bu yöntemler arasında en çok olabilirlik tahminleri ve Monte Carlo Markov Zinciri en yaygın kullanılan yöntemlerdir.

2.3. Modelin Uyum İyiliğinin Sınanması

Regresyon modellerinin uyum iyiliğinin sınanmasında Pearson istatistiği, sapma istatistiği (Deviance), Akaike Bilgi Ölçütü (AIC: Akaike Information Criterion) ve Bayes Bilgi Ölçütü (BIC: Bayesian Information Criteria) yaygın olarak kullanılan ölçütlerdir.

2.3.1. Pearson İstatistiği

Pearson istatistiği, en temel uyum iyiliği ölçütlerinden biridir. Bu istatistiğe Pearson ki-kare test istatistiği de denilmektedir. Genel olarak Denklem 17' deki eşitlik ile ifade edilir.

$$P = \sum_{i=1}^{n} \frac{(y_i - \hat{\lambda})^2}{\widehat{\omega}_i} \tag{17}$$

Pearson istatistiği Poisson regresyonu için uygulandığında, Poisson dağılımının doğal bir uzantısı olarak $\omega_i = \lambda_i$ olacaktır ve bu durumda formül değişecektir (Denklem 18).

$$P_p = \sum_{i=1}^{n} \frac{(y_i - \hat{\lambda}_i)^2}{\hat{\lambda}_i} \tag{18}$$

Serideki aşırı ya da eksik yayılımın kontrolü ise *(n-k)* serbestlik derecesi olmak üzere aşağıda verildiği gibi kontrol edilir. Burada k, parametre sayısını ve n, gözlem sayısını gösterir (Denklem 19).

$$\begin{aligned} P_p > n - k &\Rightarrow seride\ aşırı\ yayılım \\ P_p < n - k &\Rightarrow seride\ eksik\ yayılım \end{aligned} \tag{19}$$

2.3.2. Sapma İstatistiği (Deviance)

Uyum iyiliğinin ölçülmesinde yaygın kullanılan tekniklerden biri de sapma istatistiğidir. Bu istatistik değerine aynı zamanda "G^2 istatistiği" de denilmektedir. Sapma istatistiği 1 serbestlik dereceli ki-kare dağılımı gösterir. Bu istatistik Denklem 20 ile ifade edilir.

$$G^2 = 2\sum_{i=1}^{n} y_i ln\left(\frac{y_i}{\lambda_i}\right) \tag{20}$$

Bu istatistik değerinin *0*'a yakınsaması model uyumunun arttığının göstergesidir. Eğer istatistik değeri *0*'a eşit ise uyumun çok iyi olduğu söylenir.

2.3.3. Akaike Bilgi Ölçütü (AIC)

Akaike tarafından önerilen ve farklı modellerin karşılaştırılmasında yaygın olarak kullanılan ölçüt, Akaike bilgi ölçütü olarak tanımlanır. Akaike bilgi ölçütü;

$$AIC = -2log(\mathcal{L}) + 2k \tag{21}$$

şeklinde ifade edilir. Bu eşitlikte *L* log olabilirlik fonksiyonunun maksimum değerini; *k* açıklayıcı değişken sayısını göstermektedir. Mevcut modeller arasında Denklem 21 ile hesaplanan AIC değerinin en küçük olduğu 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 Denklem 22 ifadesine eşittir (Akaike, 1973; Hurvich ve Tsai, 1989).

$$AICc = AIC + 2k(k + \frac{1}{(n-k-1)} \tag{22}$$

2.3.4. Bayes Bilgi Ölçütü (BIC)

Akaike, doğrusal regresyonda seçilmiş model problemleri için BIC (Bayesian Information Criterion) model seçim kriterini türetmiştir (McQuarrie ve Tsai, 1998). Bayes bilgi ölçütüne dair eşitlik aşağıdaki gibidir (Denklem 23).

$$BIC = -2log(\mathcal{L}) + klog(n) \tag{23}$$

Akaike bilgi ölçütünde olduğu gibi mevcut modeller arasında en küçük değerli BIC değerine sahip model, uygun model olarak seçilir.

2.4. Marjinal Etkiler ve İnsidans Oranı

2.4.1. Marjinal Etkiler

Marjinal etkileri hesaplamak için delta metodu ya da standart hatalar için bootstrap metodu kullanır. Poisson ve Negatif binom regresyonun koşullu ortalama fonksiyonu Denklem 24 ile gösterilebilir.

$$E(Y/X) = e^{X\beta} \tag{24}$$

Buradan örneğin x_1 değişkenin marjinal etkisi şu şekildedir (Denklem 25) (Denny, 2009):

$$\frac{\partial Y(E(Y/X))}{\partial x_1} = \beta_1 e^{X\beta} \tag{25}$$

Bunun için başka bir seçenek ortalamaları kullanmaktır; çünkü yukarıda elde edilen bu sonuç Denklem 26 formülüne yaklaşır.

$$\frac{\partial Y(E(Y/X))}{\partial x_1} \cong \beta_1 \overline{Y} \tag{26}$$

Marjinal etkinin yorumu şu şekildedir: x' teki bir 1 birim artış, marjinal etkiye bağlı olarak bağımlı değişkenin değerini ortalama olarak artıracak ya da azaltacaktır (Katchova, 2013).

2.4.2. İnsidans Oranı (IRR: Incidence Ratio Rate)

Model katsayılarını yorumlamanın bir başka yolu da insidans oranı (IRR)'dır. Başka bir deyişle olayların meydana gelme oranına IRR oranı denir. Bu oran iki bağımsız olayın oranlanması ile bulunur. Bu orana bazen Göreli Risk (RR: Relative Risk)'de denilmektedir. Herhangi bir zamanda meydana gelen olayları karşılaştırmak için kullanılan IRR değerleri göreceli bir fark ölçüsüdür. Uygulamalarda belirli bir risk faktörü ile sonuç arasında nedensel bir ilişki araştırmasında kullanılmaktadır. Sayma regresyon uygulamasından elde edilen tahminler herhangi

bir olayın belirli bir zaman aralığındaki IRR değerine çevrilebilir. Örneğin; belirli bir hastalığa maruz kalanlar ve kalmayanlar şeklinde iki grup varsa, IRR oranları Tablo-1' deki gibi hesaplanır (Kanchanaraksa, 2008);

Tablo 1. Hastalığa maruz kalma durumu için karışıklık matrisi

Hastalığa maruz kalma durumu	Hastalık durumu		Toplam	Hastalık Sıklığı
	Hastalık gelişir	Hastalık gelişmez		
Maruz kaldı	a	b	a+b	a/a+b
Maruz kalmadı	c	d	c+d	c/c+d

IRR (ya da RR) oranı aşağıdaki gibi yazılabilir (Denklem 27);

$$IRR = \frac{a/(a+b)}{c/(c+d)} \tag{27}$$

IRR=1 ise hastalığa maruz kalan ve kalmayan gruplar arasındaki oran eşittir.

IRR> 1 ise hastalığa maruz kalanlar, maruz kalmayanlara göre daha yüksektir.

IRR<1 ise hastalığa maruz kalanlar, maruz kalmayanlara göre daha düşüktür.

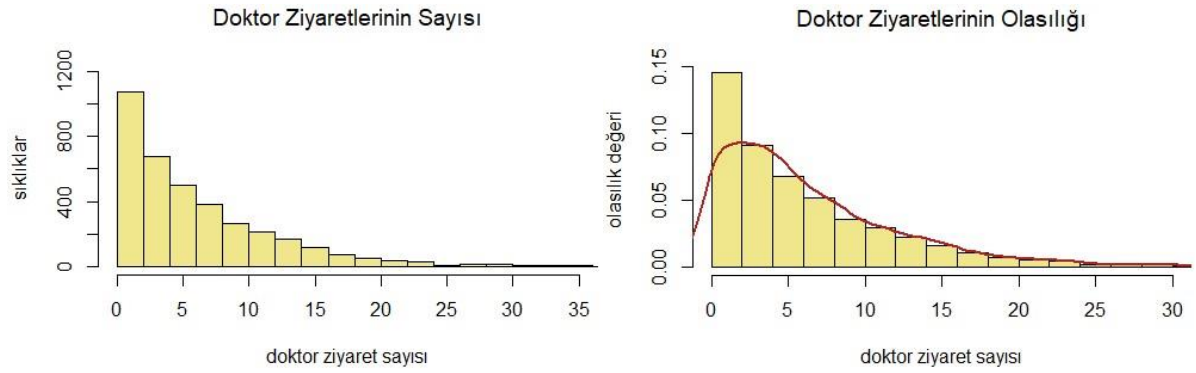
Örneğin; IRR=2 değeri, hastalığa maruz kalanların kalmayanlara oranla hasta olma oranın 2 katı olduğu anlamındadır.

3. Bulgular ve Tartışma

Bu çalışmada doktor ziyaret sayısını etkileyen faktörlerin sayma modelleri aracılığı ile incelenmesi hedeflenmiştir. Bu amaçla 2003 yılında 3677 yaşlı hastanın (65-90) doktor ziyaretlerinin sayısı hakkında bilgi içeren ABD Tıbbi Harcama Paneli Araştırmasından (MEPS) elde edilen bir veri seti kullanılarak Poisson ve negatif binom regresyon modelleri incelenmiştir (Katchova, 2013).

Bağımlı değişken olarak doktor ziyaret sayıları alınmıştır. Bağımsız değişkenler; özel sigortalı olup olmadığı, sağlık sigortasına sahip olup olmadığı, yaş, eğitim düzeyi, cinsiyet ve kronik durum değişkenleridir. Şekil 3'te bağımlı değişken olan doktor ziyaretlerinin frekans dağılımı ve olasılıkları verilmiştir. Şekilden de görüldüğü üzere veriler sağa doğru eğilimlidir. Materyal ve

metot kısmında Şekil 1 ve 2 ile verilen dağılıma uygunluk gösterilmektedir. Bu nedenle EKK regresyonu bu verilere uygun değildir.



Şekil 3. Doktor ziyaret sayısı frekans ve olasılık dağılımı

Hastalara ilişkin doktor ziyaret sayısı frekans değerleri Tablo 2'de verilmiştir.

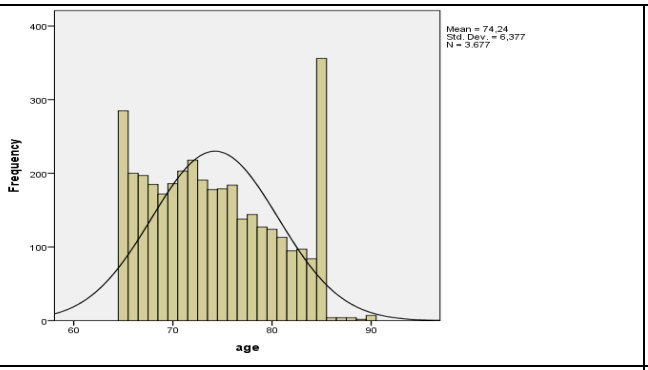
Tablo 2. Doktor Ziyaret Sayısı Frekans Tablosu

Doktor ziyaret sayısı	Frekans	%	Kümülatif %	Doktor ziyaret sayısı	Frekans	%	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				

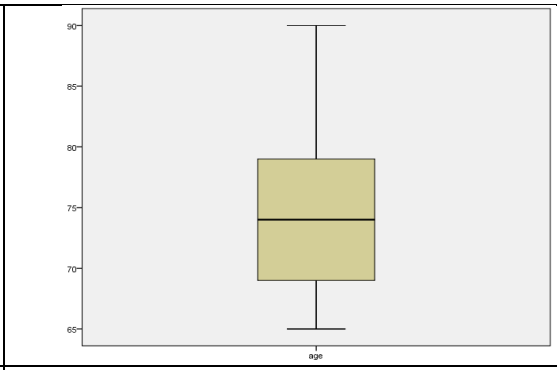
Doktor ziyaret sayısını etkileyen değişkenlerden biri yaş değişkenidir. Tablo 3'de 65-90 yaş arası hastaların frekansları ve yüzdelik değerleri verilmektedir. Yaş değişkenin dağılımını göstermek amacıyla histogram ve kutu grafiği de Şekil 4'de verilmiştir.

Tablo 3. Yaş Değişkeni Frekans Tablosu

Yaş	Frekans	%	Yaş	Frekans	%
65	285	7.8	78	144	3.9
66	200	5.4	79	127	3.5
67	197	5.4	80	124	3.4
68	185	5.0	81	113	3.1
69	172	4.7	82	95	2.6
70	186	5.1	83	97	2.6
71	203	5.5	84	84	2.3
72	218	5.9	85	356	9.7
73	191	5.2	86	4	0.1
74	178	4.8	87	4	0.1
75	179	4.9	88	4	0.1
76	184	5.0	89	1	0.1
77	138	3.8	90	2	0.2



Yaş değişkeni histogram



Yaş Değişkeni kutu grafiği

Şekil 4. Yaş Değişkenin Histogramı ve Kutu Grafiği

Doktor ziyaret sayısını etkileyen diğer değişkenler kategorik değişkenlerdir. Bu değişkenler özel sigortalı olup olmadığı, sağlık sigortası sahibi olup olmadığı, cinsiyet, eğitim düzeyi ve kronik durum değişkenleridir. Tablo 4-6'da bu değişkelere ilişkin frekans değerleri ve yüzdelik değerleri verilmiştir.

Tablo 4. Binary (0/1) Değişkenler Frekans Tablosu

Kategorik Değişken	Değişkenin açıklaması	Kategorisi	Frekans	%
Özel sigorta	Özel sigortası yok	0	1851	50.3
	Özel sigortası var	1	1826	49.7
Sağlık Sigortası	Sağlık sigortası yok	0	3036	83.3
	Sağlık sigortası var	1	613	49.7
Cinsiyet	Erkek	0	1467	39.9
	Kadın	1	2210	60.1

Tablo 5. Eğitim Düzeyi Değişkeni Frekans Tablosu

Eğitim Düzeyi	Frekans	%	Eğitim Düzeyi	Frekans	%
0	56	1.5	9	156	4.2
1	25	0.7	10	215	5.8
2	42	1.1	11	216	5.9
3	77	2.1	12	1151	31.3
4	72	2.0	13	188	5.1
5	57	1.6	14	260	7.1
6	139	3.8	15	94	2.6
7	78	2.1	16	312	8.5
8	286	7.8	17	253	6.9

Tablo 6. Kronik Koşul Değişkeni Frekans Tablosu

Kronik Koşul	Frekans	%
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

Poisson regresyonu, sayım verilerinin modellenmesinde sıklıkla kullanılır. Bununla birlikte bu modelin kullanılabilmesi için Poisson dağılımına uyması gerekir. Yani bağımlı değişkenin ortalaması varyansına eşit olmalıdır. Doktor ziyaret sayısını etkilediği düşünülen faktörler ve onlara ait tanımlayıcı istatistikler Tablo 7'de verilmiştir.

Tablo 7. Tanımlayıcı istatistikler

Değişkenler	Ortalama	Std. Sapma	Minimum	Maksimum	Değişken Türü
Doktor ziyaret sayısı	6.82	7.39	0	144	Kesikli
Özel sigorta	-	-	0	1	Kategorik
Sağlık sigortası	-	-	0	1	Kategorik
Yaş	74.24	6.37	65	90	Sürekli
Eğitim düzeyi	-	-	0	17	Kategorik
Cinsiyet	-	-	0	1	Kategorik
Kronik durum	-	-	0	8	Kategorik

Çalışmada kullanılan veriler için bağımlı değişkene ait ortalama varyanstan küçüktür ($6.82 < 7.39^2$). Bu nedenle eksik yayılım olduğunu söyleyebiliriz. Yüksek ki-kare değeri Poisson dağılımının iyi bir seçim olmadığını gösteren bir başka göstergedir. Bu nedenle negatif binom regresyon kullanılarak analiz tekrar edilmiştir. Çünkü negatif binom regresyonu, aşırı ya da eksik yayılım durumlarında genellikle daha uygundur. Poisson ve negatif binom regresyon için elde edilen sonuçlar Tablo 8'de verilmiştir.

Tablo 8. Modellerin parametre tahminleri ve uyum iyiliği istatistikleri

Değişkenler	Poisson regresyon			Negatif binom regresyon		
	β	Std. Hata	p	β	Std. Hata	p
Özel sigorta	0.123	0.014	0.000	0.152	0.033	0.000
Sağlık sigortası	0.128	0.019	0.000	0.138	0.045	0.002
Yaş	0.005	0.001	0.000	0.006	0.002	0.006
Eğitim düzeyi	0.027	0.002	0.000	0.027	0.004	0.000
Cinsiyet	-0.055	0.013	0.000	-0.151	0.030	0.623
Kronik durum	0.272	0.004	0.000	0.299	0.011	0.000
Sabit terim	0.554	0.080	0.000	0.397	0.192	0.040
Yayılım parametresi(α)	-----	-----	----	0.652	0.019	
AIC	30304			21242.49		
BIC	30348			21292.17		
LR chi2(6)	4226.40			725.63		
Sapma(G^2)	18646.72			3135.445		
Pearson(P)	23709.72			3295.150		
Pseudo R^2	0.1224			0.0331		
Log likelihood	-15145.43			-10613.245		

Poisson regresyon modeli incelendiğinde tüm değişkenler istatistiksel olarak anlamlı bulunmuştur ($p \leq 0.05$). Doktor ziyaret sayısı (dzs) bağımlı değişken olmak üzere, Poisson regresyon modeli Denklem 28 ile yazılabilir;

$$log(dzs) = 0.554 + 0.123 \cdot \text{özel sigorta} + \ldots + 0.272 \cdot kronik\ durum \tag{28}$$

Bu modelde katsayılar yarı elastikiyetler gibi yorumlanabilir. Örneğin; Eğitim düzeyi katsayısı için (0.027), bir yıl daha fazla eğitim görmüş bir hastanın %2.7 daha fazla doktor ziyareti yapması beklendiğini gösterir. Özel sigortalı bireylerin doktor ziyareti sayısının yaklaşık %13 oranında artması beklenmektedir. Diğer katsayılar benzer şekilde yorumlanır.

Negatif binom regresyon modeli incelendiğinde cinsiyet değişkeni dışında kalan diğer değişkenler istatistiksel olarak anlamlı bulunmuştur. Negatif binom regresyon modeli Denklem 29 ile ifade edilir.

$$log(dzs) = 0.397 + 0.151 \cdot \text{özel sigorta} + \ldots + 0.299 \cdot kronik\ durum \tag{29}$$

Bu modele göre; özel sigorta sahibi bir hastanın %15 daha fazla doktor ziyareti yapması beklenir.

Uyum iyiliğini incelemek için hipotezler;

H_0: Veriler Poisson modele uygunluk göstermektedir $[E(y_i|x_i) = Var(y_i|x_i)]$

H_1: Veriler Poisson modele uygunluk göstermemektedir$[E(y_i|x_i) > Var(y_i|x_i)]$

şeklinde kurulabilir. Tablo 8'de yayılım parametresi alfa (α) verilmiştir. Yayılım parametresi sıfır olduğunda negatif binom dağılımı Poisson dağılımına eşdeğerdir. Burada, α sıfırdan önemli ölçüde farklıdır$(\alpha = 0.652)$ ve bu nedenle poisson dağılımının uygun olmadığını bir kez daha pekiştirir. Böylece yokluk (sıfır) hipotezi red edilebilir.

Bunun yanı sıra en iyi modele karar vermek için AIC, BIC ve G^2 değerleri hesaplanmıştır. Veri setine göre Poisson regresyon modeline göre negatif binom regresyon modeli için daha küçük değerler bulunmuştur. Bu sonuçlara göre en uygun modelin negatif binom regresyon modeli olduğu söylenebilir.

Katsayıları yorumlamak için değişkenlerin marjinal etkileri de hesaplanabilir. Bu amaçla Poisson ve negatif binom dağılımlar için ayrı ayrı marjinal etkilerde hesaplanmıştır.

Tablo 9. Poisson ve negatif binom regresyon modellerin marjinal etkileri

Değişkenler	Poisson regresyon			Negatif binom regresyon		
	dy/dx	Std. Hata	p	dy/dx	Std. Hata	p
Özel sigorta	0.841	0.097	0.000	1.046	0.232	0.000
Sağlık sigortası	0.879	0.128	0.000	0.953	0.312	0.002
Yaş	0.039	0.006	0.000	0.045	0.016	0.006
Eğitim düzeyi	0.187	0.012	0.000	0.186	0.029	0.000
Cinsiyet	-0.375	0.089	0.000	-0.104	0.212	0.623
Kronik durum	1.858	0.032	0.000	2.065	0.097	0.000

Tablo 9'da dy/dx sütunları marjinal etkileri gösterir. Bir bireyin özel sigortası varsa Poisson regresyona göre 0.841; negatif binom regresyona göre 1.046 ek doktor ziyareti yapacağı şeklinde yorumlanır. Aynı zamanda doktor ziyaret sayısı, Poisson regresyona göre her yaş için 0.039; negatif binom regresyona göre 0.045 artmaktadır. Marjinal etkilere göre doktor ziyaret sayısını en fazla arttıran değişken kronik durum değişkenidir. Hastanın kronik bir rahatsızlığı varsa Poisson regresyona göre yaklaşık 1.858; negatif binom regresyona göre 2.065 daha fazla doktora gitmek zorundadır.

Tablo 10. Poisson ve negatif binom regresyon modellerin IRR değerleri

Değişkenler	Poisson regresyon			Negatif binom regresyon		
	IRR	Std. Hata	p	IRR	Std. Hata	p
Özel sigorta	1.131	0.016	0.000	1.164	0.038	0.000
Sağlık sigortası	1.137	0.021	0.000	1.148	0.051	0.002
Yaş	1.005	0.001	0.000	1.006	0.002	0.006
Eğitim düzeyi	1.027	0.001	0.000	1.027	0.004	0.000
Cinsiyet	0.946	0.012	0.000	0.984	0.030	0.623
Kronik durum	1.310	0.005	0.000	1.349	0.015	0.000
Sabit terim	1.741	0.139	0.000	148.755	0.285	0.039
/lnalpha				-0.427	0.030	
alpha(α)				0.652	0.019	

Regresyon modelleri için katsayıları yorumlamanın bir başka yolu da IRR değerleridir. Poisson regresyon ve negatif binom regresyon için IRR oranları Tablo 9'da verilmiştir. Poisson regresyona göre hastanın özel sigortası varsa diğer değişkenlere göre doktor ziyaret sayısını 1.131 kat artmaktadır ($e^{0.123}$~1.131). Kronik bir hastalığa sahip olmak ise doktor ziyaret sayısını yaklaşık 1.31 kat artırır. Diğer katsayılar için benzer yorum yapılabilir. Negatif binom regresyonuna göre ise; hastanın özel sigortasının olması doktor ziyaret sayısını 1.164 kat artmaktadır ($e^{0.152}$~1. 164). Benzer şekilde kronik bir hastalığı varsa ziyaret sayısı yaklaşık 1.35 kat daha fazladır.

4. Sonuçlar ve Öneriler

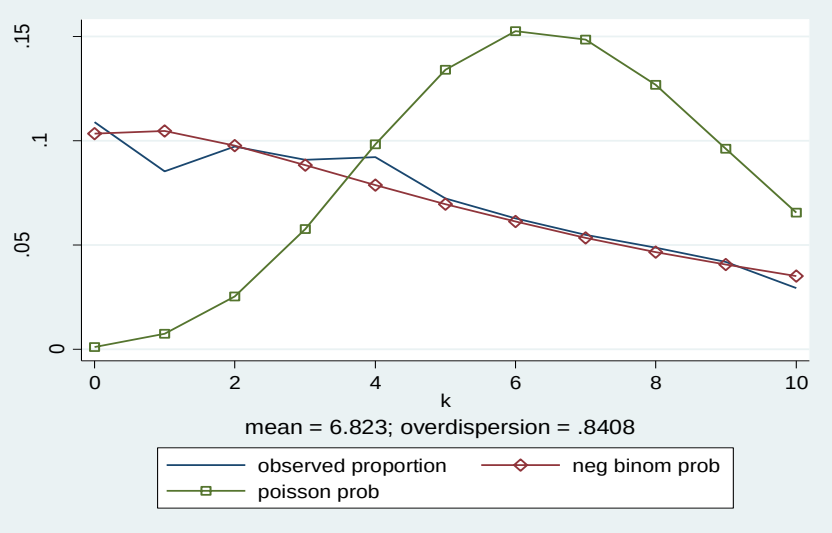
Bağımlı değişkenin kesikli ya da sayma verisi olması durumunda doğrusal regresyon modelleri kullanılarak yapılacak analizler tutarsız sonuçlar verecektir. Bu nedenle sayma verileri için farklı regresyon modelleri kullanılmaktadır. Bunlar arasında en bilinen modeller Poisson ve negatif binom regresyon modelleridir. Poisson modelin kullanılabilmesi için dikkat edilmesi gereken en önemli koşul, koşullu varyans değerinin koşullu ortalama değerine eşit olmasıdır. Bu ise uygulamada nadir görülebilen bir durumdur. Birçok uygulamada koşullu varyans değeri koşullu ortalama değerini aşar. Böyle durumlarda Poisson regresyonun kullanılması doğru değildir. Bunun gibi aşırı yayılım (ya da eksik yayılım) olduğunda, negatif binom regresyon uygulanmalı ya da yayılım parametresi ile test istatistikleri ve artıklar düzeltilmelidir. Negatif binom dağılımda varyansın, ortalamanın karesel fonksiyonu olduğu varsayılır. Bu ifade aşırı yayılımın ortadan kalkması için en önemli varsayımlardan biridir.

Poisson regresyon ve negatif binom regresyon analizleri neticesinde ele alınan veri seti için aşağıdaki sonuçlara ulaşılmıştır:

- Çalışmada kullanılan veriler için bağımlı değişkene ait ortalama, varyanstan küçüktür ($6.82 < 7.39^2$). Bu nedenle eksik yayılım olduğunu söyleyebiliriz.
- Yüksek ki-kare değeri (χ^2=4226.40) Poisson regresyonun iyi bir seçim olmadığını gösteren bir başka göstergedir. Bu nedenle negatif binom regresyon kullanılarak analiz tekrar edilmiştir. Çünkü aşırı ya da eksik yayılım durumlarında negatif binom regresyonu genellikle daha uygundur.
- Yayılım parametresi α sıfırdan önemli ölçüde farklıdır ($\alpha = 0.65$) ve bu nedenle Poisson regresyonun uygun olmadığı bir kez daha pekiştirilir.
- En iyi modele karar vermek için AIC, BIC ve G^2değerleri hesaplanmıştır. Veri setine göre negatif binom regresyon modeli için daha düşük değerler bulunmuştur. Bu sonuçlara göre en uygun modelin negatif binom regresyon modeli olduğu söylenebilir.
- Doktor ziyaret sayıları ile yapılan bu çalışmada Poisson regresyon modelinde tüm değişkenler anlamlı bulunmuştur. Fakat negatif binom regresyon analizi ile cinsiyetin etkili olmadığı bulunmuştur.
- Marjinal etkiler ve IRR oranları hesaplanarak da katsayı yorumları yapılabilir.

Poisson regresyon analizi pek çok konuda uygulanabilen regresyon analiz yöntemidir. Özellikle bağımlı değişkenin sayma verisi olduğu durumlarda ilk akla gelen yöntemlerden biridir. Ancak Poisson regresyonu, eşit ortalama ve varyans varsayımını her zaman sağlayamamaktadır (aşırı yayılım ya da eksik yayılım durumları).

Gerçek şu ki, aşırı yayılım gerçek verilerde çok yaygındır. Teoride iyi çalışan Poisson dağılımı pratikte bu kadar iyi performans göstermez (Ender, 2020). Negatif binom olasılık eğrisinin Poisson olasılık eğrisine göre verilere daha iyi uyduğu Şekil 5'de görülmektedir. Düz çizgiler gözlem değerlerini, yeşil çizgiler Poisson olasılık değerlerini ve kırmızı olan çizgiler ise negatif binom olasılık değerlerini göstermektedir. Buradan eksik yayılımdan dolayı negatif binom regresyon modelinin çok daha uygun olduğu görülmektedir.



Şekil 5. Poisson ve negatif binom olasılık değerleri

Sonuç olarak bu çalışmada örnek bir uygulama ile varsayımın sağlanmadığı durumlarda Poisson regresyon analizine alternatif kullanılabilecek olan negatif binom regresyon modeli açıklanmıştır. Poisson regresyon ile analiz yapılacağı durumlarda aşırı ya da eksik yayılım olup olmadığının kontrol edilmesi, var olduğu durumlarda genelleştirilmiş Poisson ve negatif binom regresyonu ile analize devam edilebileceğinin göz önünde bulundurulması vurgulanmıştır.

Kaynaklar

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

Al-Ghirbal A.S. and Al-Ghamdi A.S., (2006). Predecting Severe Accidents Rates at Roundabouts Using Poisson Distribution, *TRB Annual Meeting*, 06-1684.

Arısoy, İ. ve Yaprak, Ş., (2016). 1984-2015 Türkiye'de Grevlerin Belirleyicileri, *Ekonomi Bilimleri Dergisi*, 8(2). 130-116.

Burg, B.V.D., Siegers. J., and Ebmer, R.W., (1998). Gender and Promotion in the Academic Labour Market Labour, 12(4), 701-713.

Cameron, A.C. and Trivedi, P.K., (2013). Regression Analysis of Count Data. Cambridge University Press. New York.

Cox, R., (1983). Some Remarks on Overdispersion, *Biometrika*, 70: 269-274.

Denny, K.J., (2009). Very simple marginal effects in some discrete choice models. UCD Geary Institute Discussion Paper Series.

Deniz, Ö., (2005). Poisson Regresyon Analizi, *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*, 4(7), 59-72.

Ender, P.B., (2020). Applied Categorical & Nonnormal Data Analysis, Poisson Models, http://www.philender.com/courses/categorical/notes1/pois1.html (Erişim Tarihi: 10 Mart 2020).

Hilbe, J.M., (2007). Negative Binomial Regression. Cambridge, U.K.

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

Kanchanaraksa, S. (2008). Estimating Risk, http://docplayer.net/25612689-Estimating-risk-sukon-kanchanaraksa-phd-johns-hopkins-university.html (Erişim Tarihi: 10 Mart 2020).

Katchova, A., (2013). Count Data Models.
https://sites.google.com/site/econometricsacademy/econometrics-models/count-data-models. (Erişim Tarihi: 22 Şubat 2020).

Khalat, M., Deep, M. and Courbage, Y., (1997). Fertility Levels and Differentials in Beirut during Wartime: An Indirect Estimation Based on Maternity Registers, *Population Studies*, 51(1), 85-92.

Kibar, F.T., (2008). *Trafik Kazaları ve Trabzon Bölünmüş Sahil Yolu Örneğinde Kaza Tahmin Modelinin Oluşturulması,* Karadeniz Teknik Üniversitesi. Fen Bilimleri Enstitüsü. Yüksek Lisans Tezi. Trabzon.

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*, 3(3). 838-863.

Legler, J. and Roback, P., (2019). Broadening Your Statistical Horizons: Generalized Linear Models and Multilevel Models, https://bookdown.org/roback/bookdown-bysh/ (Erişim Tarihi: 10 Mart 2020).

McQuarrie, A.D. and Tsai, C. L., (1998). Regression and Time Series Model Selection. World Sciencetific.

Memiş, M. and Önder, H., (2018). Poisson Regresyon Tahmin Yöntemlerinin Karşılaştırılması, *Black Sea Journal of Engineering and Science*, 1(4), 140-146.

Michener, R. and Tighe, C., (1992). Gender and Promotion in the Academic Labour Market. *American Economic Review*, 82(2). 452-56.

Stata, (2019). StataCorp. Statistical software package, Stata v.14. https://www.stata.com/

Şahin, H., (2002). Poisson Regresyon Uygulaması: Türkiye'deki Grevlerin Belirleyicileri 1964-1998, *Doğuş Üniversitesi Dergisi*, 5, 173-180.