BŞEÜ Fen Bilimleri Dergisi
9(1), 286-303, 2022

BSEU Journal of Science
https://doi.org/10.35193/bseufbd.1018339

e-ISSN: 2458-7575 (https://dergipark.org.tr/tr/pub/bseufbd)

Araştırma Makalesi- Research Article

Sayma Verileri ile Kantil Regresyon: Aşırı Yayılım Veri Örneği

Quantile Regression with Count Data: Example of Overdispersion Data

Burcu Durmuş[1*], Öznur İşçi Güneri[2], Aynur İncekırık[3]

Geliş / Received: 03/11/2021 ***Revize / Revised: 25/02/2022*** ***Kabul / Accepted: 21/03/2022***

ÖZ

Sayma modellerinde klasik regresyon varsayımları sağlanamamaktadır. Bu nedenle sayma verileri için Poisson ve negatif binom dağılım en bilinen yöntemlerdir. Poisson model eşit yayılım durumunda, negatif binom dağılım aşırı yayılım durumunda kullanılabilir. Uygulamada veriler genellikle aşırı yayılım göstermektedir. Eğer sayma verilerinde fazla sıfır değerli varsa eşit yayılım durumunda zero-inflated Poisson, aşırı yayılım durumunda zero-inflated negatif binom modelleri, Poisson Hurdle ve negatif binom Hurdle modelleri veya bunların genelleştirilmiş modelleri tercih edilebilir. Bu modeller genel olarak bağımlı değişkenin koşullu ortalamasını modellemeye odaklanır. Ancak koşullu ortalama regresyon modelleri, bağımlı değişkenin aykırı değerlerine duyarlı olabilir ya da diğer koşullu dağılım özellikleri hakkında hiçbir bilgi sağlamayabilir. Bu durumda sayma verileri için sağlam yöntemlerden olan kantil regresyon kullanılabilir. Kantil regresyon aykırı değerlerin varlığında sağlam tahmin avantajlarına sahiptir. Bu makalede bağımlı değişken sayma verilerinden oluşan makale sayısıdır. Bağımsız değişkenler cinsiyet, evli olup olmadığı, 5 yaşının altında çocuk sayısı, doktora prestiji ve danışmanın son 3 yıldaki makale sayısı değişkenlerinden oluşmaktadır. Çalışmada Poisson ve negatif binom dağılım uygulandıktan sonra %25, %50, %75 ve %90 kantil regresyon tahminleri elde edilmiştir.

Anahtar Kelimeler- Sayma Verisi, Kantil Regresyon, Poisson Regresyon, Negatif Binom Regresyon

ABSTRACT

Classical regression assumptions are not valid in count models. Therefore, Poisson and negative binom distribution are the most common methods for count data. The Poisson model can be used in case of equal spread, while negative binom distributions in case of overdispersion. In practice, data is often over dispersed. If there are too many zero values in the count data, zero-inflated Poisson models in case of equal spread, zero-inflated negative binom models, Poisson Hurdle and negative binom Hurdle models or their generalized models can be preferred in case of overdispersion. These models generally focus on modeling the conditional average of the dependent variable. However, conditional average regression models may be sensitive to outliers of the dependent variable or provide no information about other conditional distribution properties. In this case, quantile regression, which is one of the robust methods for count data, can be used. The quantile regression has the advantages of robust prediction in the presence of outliers. In this study, count data was taken to show the dependent variable number of articles. Independent variables include of gender, marital status, number of children under the age of 5, prestige

[1]*Sorumlu yazar iletişim: burcudurmus@mu.edu.tr (https://orcid.org/0000-0002-0298-0802)
İstatistik Bölümü, Fen Fakültesi, Muğla Sıtkı Koçman Üniversitesi, Muğla, Türkiye
[2]İletişim: oznur.isci@example.com (https://orcid.org/0000-0003-3677-7121)
İstatistik Bölümü, Fen Fakültesi, Muğla Sıtkı Koçman Üniversitesi, Muğla, Türkiye
[3]İletişim: aynur.incekirik@bayar.edu.tr (https://orcid.org/0000-0002-5029-6036)
Ekonometri Bölümü, İktisadi ve İdari Bilimler Fakültesi, Manisa Celal Bayar Üniversitesi, Manisa, Türkiye

BŞEÜ Fen Bilimleri Dergisi
9(1), 286-303, 2022

BSEU Journal of Science
https://doi.org/10.35193/bseufbd.1018339

e-ISSN:2458-7575 (https://dergipark.org.tr/tr/pub/bseufbd)

of the doctorate, and the number of articles by the consultant in the last 3 years. After applying Poisson and negative binom distribution in the study, 25%, 50%, 75% and 90% quantile regression estimates were obtained.

Keywords- Count Data, Quantile Regression, Poisson Regression, Negative Binom Regression

I. GİRİŞ

Bağımlı değişkenin kesikli değer (0, 1, 2,…,k) aldığı fakat kategorik olmadığı durumlara sayma verileri denilmektedir. Sayma verilerine klasik En küçük kareler yöntemi (EKKY) uygulandığında hataların dağılımı normal dağılım göstermediğinden iyi sonuçlar verememektedirler. Bağımlı değişken sayma verilerinden oluştuğunda Poisson regresyon modeli, doğrusal regresyon analizine alternatif olarak kullanılmaktadır. Bu nedenle Poisson analizi, pek çok alanda kullanılmaktadır. Poisson modelinin en temel özelliği, koşullu ortalamanın koşullu varyansa eşit olmasıdır. Ancak uygulamalarda koşullu varyans, koşullu ortalamayı aşabilir ya da koşullu ortalamanın altında kalabilir. Aşırı (eksik) yayılım olarak adlandırılan bu durum da farklı modeller kullanılması önerilmektedir. Bu modeller arasında en bilinen yöntem negatif binom regresyon (NBR) modelidir. Aşırı (eksik) yayılım olduğu durumlarda Poisson regresyon analizi, standart hatalarının ve parametre tahminlerinin sapmalı olmasına neden olmaktadır [1]. Aşırı yayılım durumunda negatif binom dağılımının dışında, genelleştirilmiş Poisson regresyon modeli, genelleştirilmiş negatif binom regresyon modeli, quasi model gibi farklı modellerde kullanılmaktadır.

Sayma verileri doğası gereği sıfır değerlerini alır. Veri setinde sıfır sayıları çok olduğunda Poisson ve negatif binom dağılımları yeterince iyi tahmin yapamazlar. Bu durum sıfır değer yayılımı (zero-inflation) olarak tanımlanmaktadır [2,3]. Böyle bir durumda aşırı dağılım veya yetersiz dağılım (over-dispersion or under-dispersion) hesaba katılmalıdır. İlk olarak Lambert (1992) tarafından zero-inflation Poisson (ZIP) modelini önerilmiştir [4]. Daha sonra, Green 1994'te Poisson ve negatif binom regresyon modellerinde aşırı sıfırların ve örneklem seçiminin hesaba katılması ile ilgili bir çalışma yapmıştır [5]. Fazla sıfırlar ve aşırı yayılım durumunda zero-inflation negatif binom (ZINB) modeli tercih edilebilir. Bu modellerin genelleştirilmiş olanları için de uygulamalar vardır. Sayma verilerindeki fazla sıfırları modellemede kullanılan bir başka popüler yaklaşım eşit yayılım için Poisson Hurdle ve aşırı yayılım için Negatif binom Hurdle modelleridir.

Sayma verileri için geliştirilmiş birçok model vardır. Bu modellerin özellikle sağlık bilimleri ve sosyal bilimler olmak üzere birçok alanda uygulamaları mevcuttur. Kantil regresyon uygulamaları özellikle son yıllarda karşımıza çıkmaktadır. Sayma verileri için Poisson ve benzeri dağılımlar normal dağılım göstermezler sağa eğik bir dağılım gösterirler. Ayrıca aykırı gözlemler olduğunda EKK tahmin edicileri etkinlik özelliklerini kaybederler. Klasik regresyon modelleri, bağımlı değişkenin aykırı değerlerine duyarlı olabilir ve bağımlı değişkenin diğer dağılım noktalarını (örneğin, üst ve alt %5'lik kantiller) etkileyen faktörler hakkında hiçbir bilgi sağlamayabilir. Bu durumda sağlam (robust) regresyon yöntemleri önerilmektedir. En küçük mutlak sapmalar (LAD), en küçük medyan kareler (LMedS) ve kantil regresyon (QR) yöntemi bunlar arasında yer alır. LAD, kantil regresyonun özel bir durumudur. Kantil regresyonda kantil değeri 0.50 olduğunda tahmin ediciler LAD analizi ile elde edilir [6].

Tipik QR uygulamaları, sürekli popülasyonlardan rastgele örnekleme varsaymasına rağmen, farklı veri setlerine ait çalışmalar da vardır. Bunların temel çıkış noktası Manski' nin (1975, 1985) ikili ve çok terimli modeller için medyan regresyon üzerine çalışması olmuştur [7, 8]. Daha sonra Horowitz (1992, 1998) tarafından genişletilmiştir [9-10]. Powell (1984, 1986) sansürlenmiş veriler için QR' yi incelemiş [11, 12] ve Lee (1992) sıralı kesikli bağımlı değişken için medyan tahminini analiz etmiştir [13]. Daha yakın zamanlarda Koenker ve Bilias (2001), Koenker ve Geling (2001), Machado ve Portugal (2002), tarafından QR uygulamaları yapılmıştır [14-16].Machado ve Silva (2005), sayma verileri için koşullu kantillerin tahminini incelemişlerdir. Yaptıkları çalışmada verilerdeki ayrıklığı göz önüne alarak, standart kantil regresyon teknikleri kullanılarak çıkarım yapılmasına izin verecek şekilde verileri düzeltmenin mümkün olabileceğini göstermişlerdir [17]. Wu ve ark. (2014), kaza sıklığını analiz etmede sayma modellerine alternatif olarak kantil regresyon kullanımını keşfetmeyi amaçlamışlardır [18]. Congdon (2017) aykırı değerlerin varlığında ve sağlık uygulamalarında, kantil tahminlerinin risk faktörlerini yansıtabileceği etkin aykırı değer tespiti ve sağlam tahmin avantajlarına sahip olduğunu ifade etmektedir [19]. Chernozhukov ve ark. (2020), kesikli yanıt değişkenleri için eşzamanlı güven bantları sunan kantil fonksiyonlarını araştırmışlardır [20]. Frumento ve Salvati (2021) sayma verilerine kantil regresyon

uygulayarak, kesikli yanıt değişkenini yapay olarak düzleştiren bir yöntem sunmuşlardır [21]. Lamarche ve ark. (2021), boylamsal veri ortamında sıfır şişirme ile sayma yanıtları için koşullu kantil fonksiyonlarının tanımlanmasını ve tahmin edilmesini incelemişlerdir [22].

QR yöntemi, dağılımdaki fonksiyonel ilişkileri farklı noktalarda ortaya koyar. Başka bir anlatımla kantil regresyon yöntemi, bağımlı değişkeni yüzdelik dilimlere (%25, %50, %75 gibi) ayırır ve her bir dilim için ayrı ayrı tahmin edici sunar. Ayrıca bu yöntem aykırı değerlere karşı esnektir. Bu nedenle, yanlış fonksiyonel ilişkilerin oluşturduğu hatayı önler [23].

EKKY varsayımlarından biri de eşit varyans varsayımıdır. Modelde değişen varyans olması halinde tahminciler, yüksek varyans değerine sahip olurlar. Bunun bir sonucu olarak da hatalar artar. Hata teriminin dağılımı x'e bağlı değilse, tüm kantiller paralel olacaktır. Başka bir ifadeyle; hata terimleri sabit varyanslı dağılıyorsa, kantiller her zaman medyana eşit yani regresyon doğrusuna paralel olacaktır [24]. Kantil regresyonda sabit varyans olması durumunda, açıklayıcı değişkenlerle elde edilecek katsayılar tüm kantil regresyonlarda aynı, sabit terimler farklı olacaktır. Bu durumda EKK ile bulunan regresyon modeli ve medyan regresyon modeli aynı olacaktır [25]. QR, sayma verisi analizi için daha yapılandırılmış ve iyi kanıtlanmış modellerin yerini alamaz. Bununla birlikte, daha karmaşık modellerin oluşturulmasına yardımcı olan ve regresyonların yalnızca koşullu dağılımın konumunu değil, aynı zamanda tüm şeklini nasıl etkilediğine dair iç görü sağlayan değerli bir ek araç olabilir [17].

Bu çalışmanın amacı, sayma verilerinde EKKY tahmin yönteminin dayandığı model varsayımlarının sağlanamaması durumunda kullanılabilecek Poisson model, negatif binom model ve kantil regresyon yöntemini tanıtmak ve bir uygulama üzerinde tahmin edicileri karşılaştırmaktır. Bu nedenle çalışmanın ikinci bölümünde Poisson regresyon, negatif binom regresyon ve kantil regresyon yöntemleri tanıtılmıştır. Üçüncü bölümünde sayma modellerinde sık kullanılan model seçim kriterlerinden bahsedilmiş, dördüncü bölümde bir uygulama verilmiştir. Son bölümde elde edilen sonuçlar, farklı sağlam regresyon yöntemleri ile karşılaştırmalı olarak tartışılmıştır.

II. POİSSON REGRESYON (PR)

Sayma verileri için genel olarak ilk yapılan analiz Poisson regresyon yöntemidir. Poisson regresyon analizinde, bağımlı değişken y_i' nin Poisson dağılımı gösterdiğini varsaymaktadır. Poisson rasgele değişkeni bir sayı olduğundan, minimum değeri sıfırdır ve teorik olarak maksimum değeri sınırsızdır. Bir veya daha fazla ortak değişkenin bir fonksiyonu olarak, zaman birimi başına ortalama oluşum sayısı olan λ parametresi modellenmek istenir. Poisson regresyonu için belli başlı varsayımlar şunlardır:

- Bağımlı değişken y_i sayma verisidir. Sayılar pozitif tam sayılar 0 veya daha büyük olmalıdır. Eğer sayma verisi değilse, Poisson regresyonu iyi bir yöntem değildir. Poisson dağılımı kesikli bir dağılım olduğundan, bu yöntem kesirlerle veya negatif sayılarla çalışmayacaktır.
- Sayımlar bir Poisson dağılımını takip etmelidir. Bu nedenle, ortalama ve varyans aynı olmalıdır.
- Bağımsız değişkenler sürekli, ikili veya sıralı olmalıdır.
- Gözlemler bağımsız olmalıdır.

λ parametresi ile Poisson dağılımının olasılık yoğunluk fonksiyonu [26],

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

şeklindedir. Bu ifade de y_i, olayların meydana gelme sayısı, λ ise olayların zaman birimi başına tekrarlanmasının oranını ifade etmektedir. Yani λ, dağılımın ortalamasını vermektedir. Buradaki olasılık, λ değerinin bir fonksiyonu olarak değişmektedir. Poisson olasılık dağılımı sağa eğiktir. Fakat λ_i büyüdükçe dağılım normal dağılıma yaklaşır.

Poisson dağılımı çoğunlukla nadir olayların (belli bir zaman aralığında bir kavşaktan geçen arabaların sayısı, belli bir zaman aralığında bir hastalığa yakalananların sayısı, belli bir yılda meydana gelen doğal afetler

BŞEÜ Fen Bilimleri Dergisi
9(1), 286-303, 2022

BSEU Journal of Science
https://doi.org/10.35193/bseufbd.1018339

e-ISSN:2458-7575 (https://dergipark.org.tr/tr/pub/bseufbd)

gibi) oluş sayısını modellemek için kullanılmaktadır. Poisson regresyon modelinin en belirgin özelliği $\lambda_i = Var(Y|X) = E(Y|X)$ ortalama ile varyansın birbirine eşit olmasıdır. Aşırı ya da eksik dağılmış bir veri seti Poisson dağılımı ile modellenemez. Bunun nedeni koşullu beklenen değerin varyansa eşit olduğu varsayımının bozulmasıdır. Poisson regresyon analizinde genellikle log-doğrusal modelden yararlanılmaktadır. Poisson dağılımının ortalaması olan λ, bağımsız değişkenlerin x_i doğrusal bir fonksiyonu olduğu varsayılır. Poisson regresyon modeli;

$$log(\lambda_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \dots, \beta_m x_m = x_i'\beta \tag{2}$$

olarak verilebilir. Burada *x*, bağımsız değişken vektörünü ve β da tahmin edilecek parametre vektörünü ifade eder. Bu eşitlikte λ_i bağımsız değişkenlerin üstel bir fonksiyonu olmaktadır.λ_i değeri aşağıdaki gibi yazılabilir.

$$\lambda_i = exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 \dots, \beta_m x_m) = exp(x_i'\beta) \tag{3}$$

Poisson regresyonu, maksimum olasılık tahminiyle tahmin edilir. Poisson modelinin log-likelihood fonksiyonu;

$$LL = \sum_{i=1}^{n} [-\lambda_i - y_i ln(\lambda_i) - ln(y_i!)] \tag{4}$$

olarak yazılabilir. Eşitlik (4) daha sonra maksimum değere yakınsamaya kadar k kez yinelenir [27]. Poisson regresyon modeli genellikle büyük örneklem gerektirir.

Verinin aşırı ya da eksik yayılımlı olup olmadığını anlamak için Poisson modelinin varyansı ortalamaya göre hesaplanır. Hesaplanan bu değer 1'den fazla ise aşırı yayılım, 1'den küçük ise eksik yayılım olarak ifade edilir. Aşırı yayılım durumunu belirleyebilmek için Wald testi, olabilirlik oranı testi (LRT), skor testi gibi test istatistiklerinden yararlanılmaktadır. Poisson regresyon modelinde aşırı yayılım göz ardı edilirse, regresyon parametrelerine ait tahminler tutarlı olmasına karşın varyanslar olması gerekenden düşük tahmin edilir [28].

A. Poisson Regresyonda Aşırı Yayılımı Doğrulama Testi

Poisson dağılımının en önemli varsayımlarından biri varyans değeri ve ortalama değerin eşit olması (eş dağılım) olmasıdır. Ancak varsayım karşılanmazsa, aşırı dağılım oluşur ($Var(Y|X)>E(Y|X)$). Uygulamalarda genellikle sayma değişkenler ortalamadan daha büyük varyansa sahiptirler ve aşırı yayılım gösterirler. Poisson regresyonundaki aşırı dağılımı test etmek için farklı yöntemler kullanılır. Cameron ve Trivedi Poisson regresyon modelinde aşırı yayılımın mevcudiyetini doğrulamak amacıyla $Var(Y|X)=E(Y|X)+\phi[E(Y|X)]^2$ eşitliğini test etmişlerdir [29]. Burada, H_0:ϕ=0 ve H_1:ϕ>0 olmak üzere ϕ parametresinin önemi doğrulanmalıdır. Sayım verilerindeki aşırı dağılımın tespiti için, belirli bir önem düzeyinde, bir Poisson regresyon modelinin önceden tahmin edilmesi gerektiği varsayılır.y_i^* değişkeni aşağıdaki gibidir;

$$y_i^* = \frac{[(y_i - \lambda_i)^2 - y_i]}{y_i} \tag{5}$$

Bu eşitlikte λ_i, Poisson regresyon modeli tahmin edildikten sonra her bir gözlem için beklenen olayların geliş sayısını ve $(y_i - \lambda_i)^2$ her bir gözlem için gerçek sayı ile tahmin edilen sayı arasındaki farkı gösterir. Denklem 5'deki yardımcı model, Denklem 6'yı izleyerek λ'yı tek tahmin değişkeni olarak ayarlar,

$$y_i^* = \beta\lambda_i \tag{6}$$

Eğer elde edilen sonuç $Z_{\alpha/2}$ değerinden daha düşük bir değerle sonuçlanırsa aşırı dağılım meydana gelir [27]. Aşırı yayılımın sonuçlarından biri, parametre tahminlerinin standart sapması aşağı yönlü ve tahmin edici değişkenlerin önemi yukarı yönlüdür, bu nedenle geçersiz sonuçlara yol açar [30]. Gözlemlenen sıfır değerlerinin

BŞEÜ Fen Bilimleri Dergisi
9(1), 286-303, 2022

BSEU Journal of Science
https://doi.org/10.35193/bseufbd.1018339

e-ISSN:2458-7575 (https://dergipark.org.tr/tr/pub/bseufbd)

sayısının Poisson modeli ile ortaya konulan sıfır değerlerinin sayısını aşması ve gözlemlenememiş heterojenlik gibi durumlar verinin aşırı yayılım göstermesine neden olmaktadır [31]. Bu nedenle aşırı yayılım durumunda farklı yöntemler tercih edilmelidir. Aşırı yayılım da en çok kullanılan yöntem NB regresyon modelidir.

III. NEGATİF-BİNOM REGRESYON (NBR)

Negatif binom regresyon (NBR), Poisson regresyonunun özel bir durumu olarak kabul edilmektedir. Sıfır değerlerinin aşırı ya da eksik yayılım gösterdiği durumlarda alternatif bir yöntem olarak karşımıza çıkmaktadır. NBR modeli, kısıtlayıcı varsayımları gevşeten Poisson regresyonunun bir genellemesidir. Bu model Poisson-Gama karışımı bir dağılıma dayanır. Poisson dağılımı, ortalaması 1 ve ölçek parametresi v olan bir gama gürültü değişkeni dahil edilerek genelleştirilebilir. α yayılım parametresi olmak üzere Poisson-Gama karışımı negatif binom dağılımı aşağıdaki gibi elde edilir [32]:

$$P(\lambda_i,\alpha)=\frac{\Gamma(y_i+\alpha^{-1})}{\Gamma(y_i+1)\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda_i}\right)^{\alpha^{-1}}\left(\frac{\lambda_i}{\alpha^{-1}+\lambda_i}\right)^{y_i},\ i=1,2,\dots,n \tag{7}$$

$$\lambda_i=t_i\lambda,\ \ \alpha=\frac{1}{v}$$

NBR modelinin beklenen değeri $E(y_i)=\lambda$ ve varyansı aşağıdaki biçimindedir.

$$Var(x_i)=\lambda_i+\alpha{\lambda_i}^2 \tag{8}$$

NBR modeli t_i maruz kalma süresi ve $\beta_1,\beta_2,\dots,\beta_k$ bilinmeyen parametreler olmak üzere,

$$\lambda_i=exp(ln\ ln(t_i)\beta_{1i}x_{1i}+\beta_{2i}x_{2i},\dots,\beta_{ki}x_{ki}) \tag{9}$$

şeklinde gösterilir. Regresyon katsayıları tahmin etmek için, maksimum likelihood (MLE) yöntemi [33] ve Monte Carlo Markov Zinciri en yaygın kullanılan yöntemlerdir. Negatif binom modelinin log-likelihood fonksiyonu,

$$LL=\sum_{i=1}^{n}(y_i ln\alpha+y_i(\alpha_i\beta_i)-(y_i+\frac{1}{\alpha})\ ln\ ln\left(1+\alpha e^{\alpha_i\beta_i}\right)+\left(y_i+\frac{1}{\alpha}\right)-ln\ ln\Gamma(y_i+1)-ln\Gamma(\frac{1}{\alpha})) \tag{10}$$

olarak verilebilir [34]. NBR modelinde aşırı yayılımdan kaynaklanan etkiyi ortaya koymak için modele yeni bir parametre eklenir [35].Bu nedenle Poisson dağılımdan farklı olarak bir parametresi daha vardır. Bu ikinci parametre, varyansı ortalamadan bağımsızlaştırmak için kullanılabilir.

IV.KANTİL REGRESYON

EKKY değişkenler arasındaki ilişkiyi belirlerken en yaygın kullanılan modeldir. Fakat EKKY gözlemlenen verilerin normallik bozulumu, değişen varyans durumu, aykırı değere sahip olma gibi varsayımların sağlanmadığı durumlarda güvenilir tahminler vermez [36,37]. Bu varsayımlar sağlanırsa tahminciler en iyi tahminciler olabilmektedir aksi halde tahminciler etkin değildir. Ayrıca bağımlı değişken sürekli olduğu zaman EKKY kullanılabilir. Kantil regresyon yönteminde EKKY bulunan hataların dağılımının normal olması ve varyansın homojen olması varsayımı gerekli değildir. Bu nedenle bu yöntem EKKY göre daha esnek bir yöntemdir.

Kantil Regresyon yöntemi ilk olarak Koenker ve Bassett (1978) tarafından iklim çalışmaları için önerilmiş bir regresyon modelidir [6]. Yöntemin ilk uygulamalarında bağımlı değişkeni sürekli olan veriler için analiz yapılmıştır. Kesikli verilerin koşullu fonksiyonlarını modellemek literatürde daha az yaygındır. Bununla birlikte, sürekli durumda olduğu gibi kesikli durumda da anlamlı bilgiler sağlayabileceği son yapılan çalışmalarla gösterilmektedir. Gerçekten de sürekli dağılımların koşullu kantillerine yönelik yöntemler, kesikli bağımlı değişkenli yöntemlere uyarlanabilir ve uyarlanmıştır. Kesikli bağımlı değişkenli yani kategorik ve sayma modeller

BŞEÜ Fen Bilimleri Dergisi
9(1), 286-303, 2022

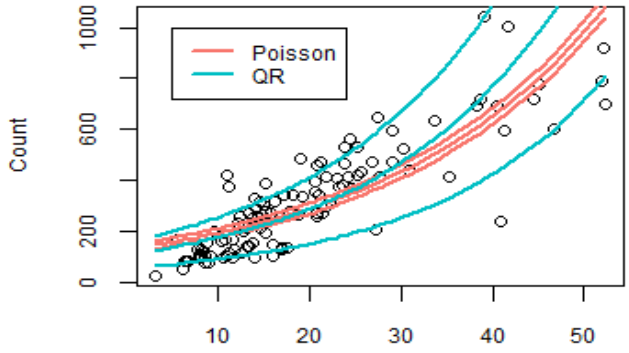
BSEU Journal of Science
https://doi.org/10.35193/bseufbd.1018339

e-ISSN:2458-7575 (https://dergipark.org.tr/tr/pub/bseufbd)

için genel olarak MLE yaklaşımı kullanılır. Bu yöntemin hesaplama avantajları ve asimptotik özellikleri nedeniyle sayma verilerinde Poisson ve binom regresyon birçok uygulamalı bilimde iyi bilinmektedir. Kantil regresyon ise Bootstrap yöntemi ve doğrusal programlama yönteminde genel olarak tahmin için kullanılmaktadır.

Aykırı değer, veri setindeki gözlemlerin büyük bir kısmının sahip olduğu dağılıma (veya modele) uymayan gözlemdir [38]. Veri kümesindeki bazı gözlemler, diğer verilere göre aşırı büyük ya da küçük olduğunda bu gözlemler veri setindeki çoğunluk verilerle benzer dağılmaz. Bu tür gözlemler, örnekleme ilişkin bilgiyi özetleyen tahmin edicileri etkileyebilir. Tahmin edici, veri setinde bulunan aykırı gözlemin varlığından etkilenmiyorsa o tahmin ediciye dayanıklı, etkileniyorsa dayanıklı olmayan tahmin edici denir [39]. Aykırı değere sahip veri setinde varsayım bozulmalarından dolayı EKKY regresyon modelinden elde edilen sonuçlar yanıltıcı olabilir. Aykırı değerlerin veri setinden çıkartılması, regresyon denklemini tamamen veya kısmen etkileyebilir. Bu nedenle aykırı değer varlığında tahminlerin güvenirliliği için sağlam regresyon yöntemlerini tercih etmek daha uygun olmaktadır [40]. Bu durumda kullanılacak alternatif regresyon modellerinden biri kantil regresyon yöntemidir.

Kantil regresyon tek değişkenlide bilinen kantil kavramının bir veya daha fazla değişken için genelleştirilmesidir. Kantil regresyon yöntemi farklı kantillere göre (örneğin 0.05, 0.10,…, 0.95) bağımlı değişkenin koşullu dağılımının farklı bölgelerdeki ortak değişkenlerin etkilerini tahmin eder. Uygulamada kantiller genellikle 0.25, 0.50 ve 0.75 olarak seçilir. Çoklu doğrusal regresyon doğrusu aşırı değerleri tespit etmede başarısız olurken, kantil regresyon doğruları farklı kantillere sahip olduğundan aşırı değerleri daha kolay tespit edebilir [41]. Şekil 1'de sayma verilerinde en bilinen yöntem olan Poisson regresyonu ve farklı kantil değerlerine göre (0.10; 0.50; 0.90) regresyon grafiği görülmektedir [42].



Şekil 1. Farklı Kantil Değerlerine Göre Poisson ve QR Grafiği

Sayma modellerinde en çok kullanılan Poisson regresyon Şekil 1'de görüldüğü gibi orta kısımda yer almaktadır. Farklı kantiller ile analiz yapıldığında diğer uç gözlemlerde analize dahil olmaktadır. Ayrıca EKKY ile tahmin edilen regresyon doğrusu dağılımın orta bölgesinden geçmekte ve değerleri dikkate almamaktadır. Ancak kantil regresyonda örneğin 0.25'lik ve 0.75'lik dilimde yer alan tahmin değerleri farklıdır. Böylece her bir bağımsız değişkenin ilgili değişkeni nasıl etkilediği konusunda daha eksiksiz bilgi edinebilir. Klasik EKK yöntemi, koşullu ortalamaya bağlı iken kantil regresyon yöntemi ise koşullu kantil fonksiyonuna bağlıdır [43]. Bu nedenle kantil regresyon, özellikle koşullu kantillerin değişkenlik gösterdiği durumlarda oldukça kullanışlıdır. Doğrusal kantil regresyon modeli şu şekilde yazılır;

$$Q_{y_t}(\tau) = \sum_{i=1}^{k} \beta_{\tau,i} x_{ti} \tag{11}$$

BŞEÜ Fen Bilimleri Dergisi
9(1), 286-303, 2022

BSEU Journal of Science
https://doi.org/10.35193/bseufbd.1018339

e-ISSN:2458-7575 (https://dergipark.org.tr/tr/pub/bseufbd)

$\beta_{\tau,i}$ bilinmeyen parametreler ve τ kantil değerini gösterir. 0<τ<1 olmak üzere $Q_{y_t}(\tau)$ ise y_t'nin τ.'nin koşullu kantilini gösterir. Örneğin, τ=0.50 alınırsa $Q_{y_t}(0.50)$ dağılımın medyanını ifade eder. τ=0.90 değeri bağımlı değişkenin en yüksek %90'lık kantil içerisinde yer aldığını, τ=0.10 değeri ise bağımlı değişkenin en düşük %10'luk kantil içerisinde yer aldığını gösterir. Regresyon katsayıları, asimetrik bir mutlak kayıp fonksiyonu kullanılarak tahmin edilir. Ayrıca medyan regresyon (LAD), τ =0.5 kantil değeri için genişletilmiş kantil regresyon durumudur.

Katsayı tahminleri doğrusal regresyona benzer olarak görülmektedir. Fakat kantil regresyonda bağımlı değişkenin koşullu dağılımının farklı noktaları için tahmin yapılmaktadır. Bu modeldeτ değerine göre kantil regresyon olmak üzere kantil regresyon tahmincileri doğrusal programlama şeklinde ifade edilip simpleks algoritması ile çözülebilir. F dağılım fonksiyonuna sahip bağımlı değişken Y için τ. regresyon kantili;

$$min_{\beta \epsilon R^K} \frac{1}{n}\left[\sum_{i\epsilon\{i:i\geq x_i\beta\}} \tau|y_i - x_i\beta| + \sum_{t\epsilon\{t:y_t<x_i\beta\}} (1-\tau)|y_i - x_i\beta|\right] \quad (12)$$

ifadesinin minimize edilmesi ile elde edilir [6].

Kantil regresyonunun bu şekilde gösterimi doğrusal programlama gösterimidir. Kantil regresyondaki parametre tahminleri, açıklayıcı değişkendeki bir birim değişiminbağımlı değişken y'nin belli bir kantilindeki değişimini gösterir [44].

V. MODEL SEÇİMİ

Sayma modelleri için model seçiminde genellikle Log Likelihood, Pearson istatistiği, sapma istatistiği (Deviance), Akaike Bilgi Kriteri (AIC) ve Bayes Bilgi Kriteri (BIC) yaygın olarak kullanılmaktadır. Literatürde en çok kullanılan kriterler AIC ve BIC olduğu içinaşağıda bu değerlerinin nasıl hesaplandığına ilişkin formüller verilmiştir.

A. Akaike Bilgi Kriteri (AIC)

Faklı modellerin karşılaştırılmasında yaygın olarak kullanılan bu ölçüt,

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

şeklinde ifade edilir [45]. Bu eşitlikte *L*, log olabilirlik fonksiyonunun maksimum değerini ve *k*, açıklayıcı değişken sayısını gösterir. Elde edilen modeller arasında AIC değerinin en küçük olduğu model en uygun modeldir. Parametre sayısı örnek büyüklüğüne göre büyük ise AIC yerine Hurvich ve Tsai tarafından önerilmiş olan AICc'nin kullanılması gerekir. Bu değer ise,

$$AICc = AIC + \frac{2k(k+1)}{n-k-1} = \frac{2kn}{n-k-1} - 2l(L) \quad (14)$$

olarak yazılabilir [46-48].

B. Bayes Bilgi Kriteri (BIC)

Akaike, doğrusal regresyonda seçilmiş model problemleri için BIC (Bayesian Information Criterion) model seçim kriterini ortaya koymuştur [47].Bayes bilgi ölçütü aşağıdaki gibi ifade edilir:

$$BIC = -2log(L) + klog(n) \quad (15)$$

Akaike bilgi ölçütüne benzer şekilde, modeller arasında en küçük BIC değerine sahip model uygun model olarak seçilir.

BŞEÜ Fen Bilimleri Dergisi
9(1), 286-303, 2022

BSEU Journal of Science
https://doi.org/10.35193/bseufbd.1018339

e-ISSN:2458-7575 (https://dergipark.org.tr/tr/pub/bseufbd)

VI. UYGULAMA

Uygulama için doktoradaki biyokimyacılar tarafından üretilen yayınların sayısı ile ilgili Long (1990) verileri kullanılmıştır. Bu veriler ayrıca Long ve Freese (2001) tarafından bazı sayma modelleri için analiz edilmiş [49] ve Stata web sitesinde mevcuttur [50]. Bu çalışmada bağımlı değişken doktoranın son 3 yılındaki makale sayısı ve bağımsız değişkenler cinsiyet, evli olup olmadığı, 5 yaşının altında çocuk sayısı, doktora prestiji ve danışman son 3 yıldaki makale sayısı dikkate alınmıştır. Analizler için SPSS 22 ve STATA 13 paket programı kullanılmıştır. Veri kümesindeki değişkenler ve tanımlayıcı istatistikler Tablo 1 ve Tablo 2'deverilmiştir.

Tablo 1. Bağımlı ve Bağımsız Değişkenler

Değişken Adı	**Açıklama**	**Kategori**
art	doktoranın son üç yılındaki makale sayısı	0,1,..,19
fem	cinsiyet	0:erkek 1:kadın
mar	evli olup olmadığı	0:hayır 1:evet
kid5	altı yaşın altındaki çocuk sayısı	0, 1, 3
phd	doktora programının prestiji	-
ment	son üç yılda danışman tarafından yazılan makaleler	0, 1, 2,..,77

Tablo 2. Tanımlayıcı İstatistikler

Değişkenler	**n**	**Ort.**	**Std. Sap.**	**Min.**	**Maks.**
art	915	1.693	1.926	0	19
fem	915	.46	.499	0	1
mar	915	.662	.473	0	1
kid5	915	.495	.765	0	3
phd	915	3.103	.984	.755	4.62
ment	915	8.767	9.484	0	77

Ortalama makale sayısı 1.69'dur ve standart sapma 1.926 (varyans 3.71)'dır. Buna göre varyans ortalamanın iki katından biraz fazladır. Şekil 2'de art değişkeni için histogram verilmiştir.

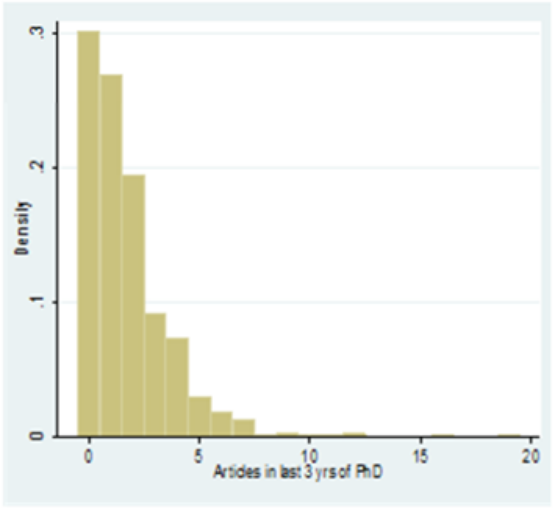
BŞEÜ Fen Bilimleri Dergisi
9(1), 286-303, 2022

BSEU Journal of Science
https://doi.org/10.35193/bseufbd.1018339

e-ISSN:2458-7575 (https://dergipark.org.tr/tr/pub/bseufbd)

art	Freq.	Percent	Cum.
0	275	30.05	30.05
1	246	26.89	56.94
2	178	19.45	76.39
3	84	9.18	85.57
4	67	7.32	92.90
5	27	2.95	95.85
6	17	1.86	97.70
7	12	1.31	99.02
8	1	0.11	99.13
9	2	0.22	99.34
10	1	0.11	99.45
11	1	0.11	99.56
12	2	0.22	99.78
16	1	0.11	99.89
19	1	0.11	100.00
Total	915	100.00	

art değişkenin frekans dağılımı

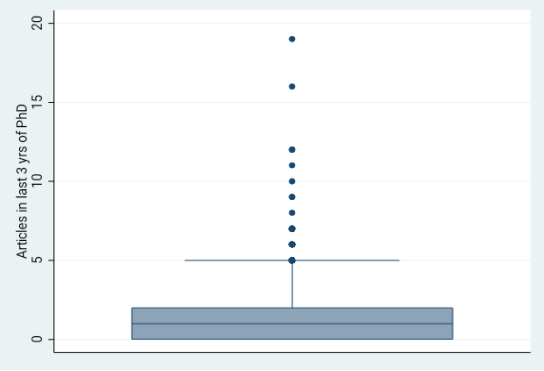


Histogram

Şekil 2. Makale Sayılarına İlişkin Histogram

Şekil 2 incelendiğinde veriler kuvvetli bir şekilde sağa çarpıktır. Bu nedenle EKKY bu verilere uygun olmayacaktır. Sayma verileri genellikle bir Poisson dağılımını takip eder, bu nedenle bir tür Poisson analizi uygun olabilir. Değişkenler arasındaki ilişkilerin incelenmesinde yaygın olarak kullanılan EKKY verilerin normal dağılım göstermesine ve aykırı değerlerin bulunmamasına bağlı olarak etkin sonuçlar vermektedir. Ancak birçok veri seti normal dağılım göstermemekte ve içinde aykırı değerlerde bulundurmaktadır. Aykırı değerler diğer gözlem değerlerinden uzakta bulunan ve onlarla tutarlılık göstermeyen değerlerdir. Veri setinde aykırı değer olması durumda bu sapmalardan etkilenmeyen ve daha esnek olan sağlam yöntemlerin kullanılması gerekmektedir. Aykırı değerler görsel olarak histogram ve kutu grafiği (Box Plot) yardımıyla belirlenebilir. Bunun dışında farklı test yöntemleri de mevcuttur.

Bu çalışmada veri setinde aykırı gözlemlerin olup olmadığını göstermek amacıyla kutu grafiği kullanılmıştır. Şekil 3'de Box Plot grafiğinde görüldüğü üzere aykırı gözlem değerleri mevcuttur. Aykırı gözlemler olduğunda sağlam regresyon yöntemlerinin kullanılması uygundur. Bu çalışma da kantil regresyon yöntemi kullanılarak analiz yapılmıştır.



Şekil 3. Makale Sayılarına İlişkin Box Plot Kutu Grafiği

BŞEÜ Fen Bilimleri Dergisi
9(1), 286-303, 2022

BSEU Journal of Science
https://doi.org/10.35193/bseufbd.1018339

e-ISSN:2458-7575 (https://dergipark.org.tr/tr/pub/bseufbd)

A. Poisson Regresyon (PR)

Poisson regresyonu, sayma verilerinin modellenmesinde kullanılmaktadır. Bunun yanı sıra Poisson modelinin kullanılabilmesi için verilerin Poisson dağılımına uyması ve eşit yayılım göstermesi gerekmektedir. Yani bağımlı değişkenin ortalaması ve varyansı eşit olmalıdır. Poisson regresyon analizi için elde edilen sonuçlar Tablo 3'te verilmiştir.

Tablo 3. Poisson Regresyon Analizi

Poisson regresyon	Gözlem Sayısı	=	915
	LR chi2(5)	=	183.03
	Prob > chi2	=	0.0000
Log likelihood = -1651.0563	Pseudo R2	=	0.0525

art	Katsayılar	Std. Hata	z	P>z	% 95 Güven	Aralığı
fem	-.2245942	.0546138	-4.11	0.000	-.3316352	-.1175532
mar	.1552434	.0613747	2.53	0.011	.0349512	.2755356
kid5	-.1848827	.0401272	-4.61	0.000	-.2635305	-.1062349
phd	.0128226	.0263972	0.49	0.627	-.038915	.0645601
ment	.0255427	.0020061	12.73	0.000	.0216109	.0294746
sabit	.3046168	.1029822	2.96	0.003	.1027755	.5064581

Regresyon katsayıları hakkındaki hipotezleri test etmek için, tam dağılım varsayımları yapıldığı için mümkün olan Wald testleri veya olabilirlik oranı testleri kullanabilir. Poisson modeli bütün olarak test edildiğinde istatistiksel olarak anlamlı olduğunu görebiliriz (Prob > chi2=0.000). Bu model incelendiğinde phd değişkeni dışındaki tüm değişkenler istatistiksel olarak anlamlı bulunmuştur ($p<0.05$). Aşırı yayılım testi sonuçları Tablo 4'te verilmiştir.

Tablo 4. Aşırı Yayılım Testi

Kaynak	KT	sd	KO			
				Gözlem sayısı	=	915
				F(1, 914)	=	42.64
Model	785.612431	1	785.612431	Prob > F	=	0.0000
Hata	16840.9983	914	18.4255998	R-squared	=	0.0446
				Adj R-squared	=	0.0435
Toplam	17626.6107	915	19.2640554	Root MSE	=	4.2925

Yasterisk (y*)	Katsayı	Std. Hata	t	P>t	%95 Güven Aralığı	
lambda	.5091216	.0779701	6.53	0.000	.3561003	.6621428

Tablo 4'te lambda değişkeninin β katsayısı 0.05 den küçük bu nedenle aşırı yayılım vardır. Modelin uyum iyiliği istatistikleri şu şekildedir;Deviance goodness-of-fit = 1634.371, Prob > chi2(909) = 0.0000 ve Pearson goodness-of-fit = 1662.547, Prob > chi2(909) = 0.0000.Uyum iyiliğinden elde edilen anlamlı ($p<0.05$) bir test istatistiği, Poisson modelinin uygun olmadığını gösterir. Aşırı yayılım nedeniyle negatif binom dağılım uygulanmıştır.

B. Negatif-Binom Regresyon (NBR)

NBR modeli Poisson regresyonu ile aynı ortalama yapıya sahip olduğu ve aşırı dağılımı modellemek için fazladan bir parametreye sahip olduğu için Poisson regresyonunun bir genellemesi olarak düşünülebilir. Sonuç değişkeninin koşullu dağılımı aşırı dağılmışsa, Negatif binom regresyon için güven aralıklarının Poisson regresyonuna kıyasla daha dar olması muhtemeldir [50]. Negatif binom regresyon analizi için elde edilen sonuçlar Tablo 5'te verilmiştir.

BŞEÜ Fen Bilimleri Dergisi
9(1), 286-303, 2022

BSEU Journal of Science
https://doi.org/10.35193/bseufbd.1018339

e-ISSN:2458-7575 (https://dergipark.org.tr/tr/pub/bseufbd)

Tablo 5. Negatif Binom Regresyon Analizi

	Negatif binom regresyon	Gözlem sayısı	=	915	
		LR chi2(5)	=	97.96	
	Dispersion = mean	Prob > chi2	=	0.0000	
	Log likelihood = -1560.9583	Pseudo R2	=	0.0304	
art	**Katsayılar**	**Std. Hata**	**z**	**P>z**	**%95 Güven**	**Aralığı**
fem	-.2164184	.0726724	-2.98	0.003	-.3588537	-.0739832
mar	.1504895	.0821063	1.83	0.067	-.0104359	.3114148
kid5	-.1764152	.0530598	-3.32	0.001	-.2804105	-.07242
phd	.0152712	.0360396	0.42	0.672	-.0553652	.0859075
ment	.0290823	.0034701	9.38	0.000	.0222811	.0358836
_cons	.256144	.1385604	1.85	0.065	-.0154294	.5277174
/lnalpha	-.8173044	.1199372			-1.052377	-.5822318
alpha	.4416205	.0529667			.3491069	.5586502
Likelihood-ratio test of alpha=0: chibar2(01) = 180.20 Prob>=chibar2 = 0.000						

NBR modeli bütün olarak test edildiğinde istatistiksel olarak anlamlı olduğunu görebiliriz (Prob > chi2=0.000). Bu model incelendiğinde mar ve phd değişkeni dışındaki değişkenler istatistiksel olarak anlamlı bulunmuştur ($p<0.05$).

NBR aşırı dağılmış sayım verileri için, yani koşullu varyans koşullu ortalamayı aştığında kullanılabilir. Bu modelde α (alpha=0.44) yayılım parametresini gösterir. Poisson modeli, bu alfa değerinin sıfırla sınırlandırıldığı modeldir. Başka bir deyişle; yayılım parametresi sıfır ise negatif binom dağılımı ve Poisson dağılımı eşdeğerdir. Bu analizde α sıfırdan önemli derecede farklı bulunmuştur. Bu durumda Poisson dağılımının uygun olmadığı pekiştirilir.

Aşırı dağılmanın yaygın bir nedeni de ek bir veri oluşturma işleminin neden olduğu aşırı sıfırlardır. Bazı durumlarda yanlış belirlenmiş bir model, aşırı dağılma problemi gibi bir belirti sunabilir. Örneğin, bazı değişkenlerin modelden çıkarılması ya da modelin fonksiyonel biçiminin yanlış seçilmesi gibi. Aşırı dağılmanın yaygın bir nedeni fazladan sıfırlar olup, bunların da ek bir veri oluşturma işlemi tarafından üretilmesidir.

Bu veri setinde 915 gözlem arasında 275 gözlem sıfırlardan oluşmaktadır. Yani verilerin yaklaşık %30'u sıfırlardan oluşmaktadır. Bu durumda zero-inflation (ZI) modeller de yapılabilir.

C. Kantil Regresyon (QR)

Kantil regresyon, doğrusal regresyondaki klasik varsayımlardan hata terimlerinin normal dağılmasını ihmal eden sağlam bir regresyon tekniği olarak ortaya çıkmıştır. Çoklu doğrusal regresyon doğrusu aşırı değerleri yakalayamazken, farklı kantillerdeki regresyon doğruları aşırı değerleri rahat bir şekilde yakalayabilir. Çalışmada normallik varsayımı için Shapiro-Wilktesti kullanılmıştır. Shapiro-Wilk testi en güçlü normallik testlerinden biridir. Tablo 6 ile elde edilen sonuçlar verilmiştir.

Tablo 6. Shapiro-Wilk Normallik Testi

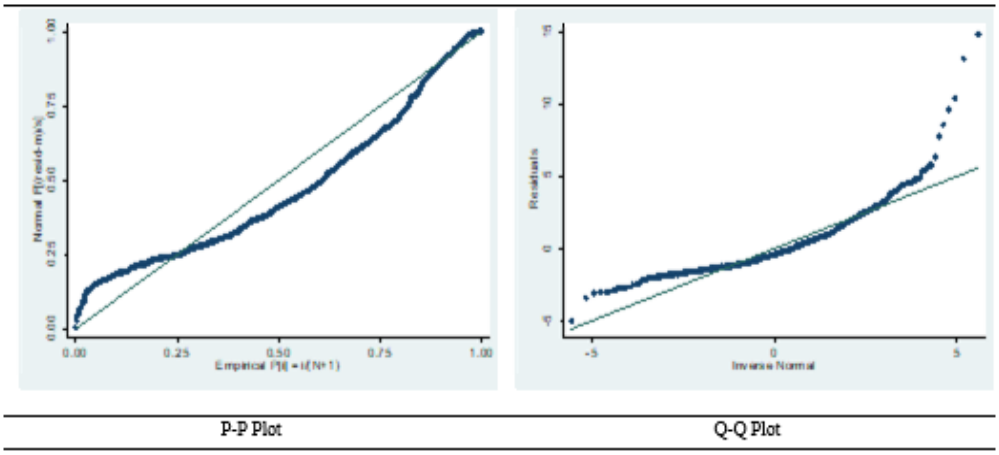
Değişken	**n**	**W**	**V**	**z**	**Prob>z**
Artıklar	915	0.85042	87.008	11.019	0.00000

Tablo 6'ya göre p değeri 0,05'ten küçüktür. Buna göre artıklar normal dağılımı normal göstermemektedir. Böylece hata terimlerinin normal dağılım varsayımının ihlali söz konusudur. Bu durum P-P plot ve Q-Q Plot ile de görülebilir. Şekil 4' de grafikler verilmiştir. Bu grafikler bir istatistiksel test olmasa da artıkların normal dağılıp dağılmadığını görsel olarak kontrol etmenin kolay bir yolunu sunar.

BŞEÜ Fen Bilimleri Dergisi
9(1), 286-303, 2022

BSEU Journal of Science
https://doi.org/10.35193/bseufbd.1018339

e-ISSN:2458-7575 (https://dergipark.org.tr/tr/pub/bseufbd)



Şekil 4. Artıklar için P-P Plot ve Q-Q Plot Grafikleri

Değişen varyans(heteroscedasticity), hata teriminin varyansının veri setindeki tüm gözlemler için eşit olmaması durumudur. Değişen varyans problemi, önemli bir değişkenin modelin dışında kaldığı durumlarda, aykırı değerlerin olduğu durumlarda ya da model kurma hatasının olduğu durumlarda oluşur. Değişen varyans durumunda hipotezler şu şekilde kurulur:

H_0: $\sigma_1^2, \sigma_2^2, \dots, \sigma_k^2 = \sigma^2$ (Değişen varyans yoktur yani sabit varyanslıdır)

H_1: $\sigma_1^2, \sigma_2^2, \dots, \sigma_k^2 \neq \sigma^2$ (Değişen varyans vardır)

Değişen varyansın tespiti için farklı testler geliştirilmiştir. Bu çalışmada değişen varyans problemi, Breusch-Pagan testi ile test edilmiştir. Chi2(1) = 244.04 (Prob > chi2 = 0.0000). Test sonucuna göre $p < 0.05$ olarak bulunmuştur. Buna göre, sıfır hipotezi reddedilir (alternatif hipotezi reddedecek kadar yeterli kanıt yoktur) ve veride değişen varyans olduğu görülür. Böylece aykırı değerlerin veri setinde bağımlı ve bağımsız değişkenin dağılımını ve varyansının değiştirdiği yorumu yapılır. Yukarıda elde edilen sonuçlar bize EKKY yerine sayma verileri için başka yöntemlerin kullanılması gerektiğini ispatlar. Genel olarak bu gibi durumlarda sağlam yöntemler kullanılmaktadır. Yapılan çalışmalarda özellikle aykırı gözlemler varsa kantil regresyonun kullanımın uygun olduğu görüşü vardır. Bu nedenle en yaygın kullanılan kantil değerleri için hesaplamalar yapılmıştır. Tablo 7'de %25, %50, %75, %90 kantil değerleri için bulunan tahminler verilmiştir.

Tablo 7. Kantil Regresyon

Simultaneous quantile regression	Gözlem sayısı			=	915	
bootstrap(100) SEs	.25 Pseudo R2			=	0.0380	
		.50 Pseudo R2		=	0.0319	
		.75 Pseudo R2		=	0.0630	
		.90 Pseudo R2		=	0.0724	

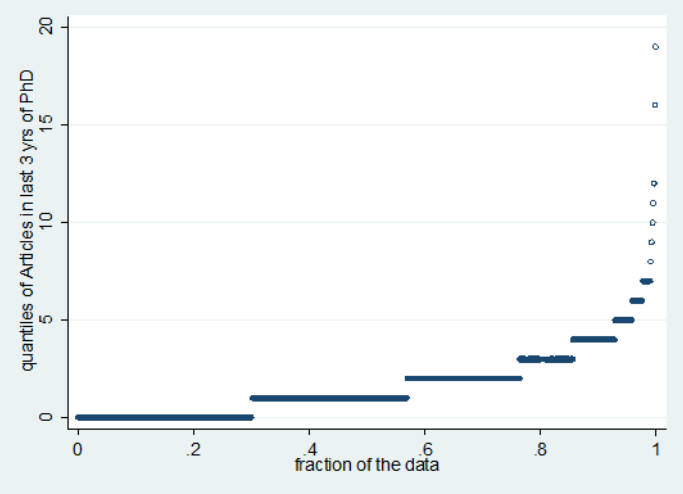
			Bootstrap				
	art	**Katsayılar**	**Std. Hata**	**t**	**P>t**	**%95 Güven**	**Aralığı**
q25							
	fem	-.0898876	.0707628	-1.27	0.204	-.2287651	.0489898
	mar	.0674157	.1007282	0.67	0.503	-.1302712	.2651026
	kid5	-.0674157	.0590475	-1.14	0.254	-.1833011	.0484696
	phd	1.00	.0386999	0.00	1.000	-.0759516	.0759516
	ment	.0449438	.0060887	14.062	0.000	.0329944	.0568933
	sabit	1.00	.1093207	0.00	1.000	-.2145504	.2145504
q50							
	fem	-.1274233	.0972173	-1.31	0.190	-.3182198	.0633732
	mar	.1991342	.1151135	1.73	0.084	-.0267849	.4250533
	kid5	-.1633729	.0710861	-2.30	0.022	-.3028847	-.023861
	phd	.0376435	.0593335	0.63	0.526	-.0788031	.1540902
	ment	.0479955	.0086205	5.57	0.000	.031077	.064914
	sabit	.706757	.2110244	3.35	0.001	.2926053	1.120.909
q75							
	fem	-.1542744	.2071772	-0.74	0.457	-.5608755	.2523268
	mar	.2576541	.1799044	1.43	0.152	-.0954222	.6107303
	kid5	-.2934394	.13209	-2.22	0.027	-.5526761	-.0342026
	phd	-.0397614	.1012572	-0.39	0.695	-.2384866	.1589637
	ment	.07833	.0224941	3.48	0.001	.0341835	.1224765
	sabit	1.933.201	.2760592	7.00	0.000	1.391.413	2.474.988
q90							
	fem	-.5092639	.2771373	-1.84	0.066	-1.053.167	.0346394
	mar	.0916375	.3516539	0.26	0.794	-.5985104	.7817854
	kid5	-.3885829	.1753955	-2.22	0.027	-.7328101	-.0443557
	phd	.2003004	.1573264	1.27	0.203	-.1084647	.5090656
	ment	.116675	.0258961	4.51	0.000	.065852	.1674981
	sabit	2.758.137	.6215158	4.44	0.000	1.538.365	397.791

Kantil regresyon sonuçları doğrusal regresyon sonuçları ile aynı şekilde yorumlanır. Fakat anlamlı olan değişkenler farklılık gösterebilir. Tablo 7'deki sonuçlara baktığımızda %25'de ment değişkeni anlamlı diğer değişkenler anlamsız bulunmuştur. %50, %75 ve %90'da kid5 ve ment, değişkeni anlamlı fem, mar ve phd değişkenleri anlamsızdır. Buna göre ment değişkeni tüm kantillerde yüksek öneme sahiptir.

BŞEÜ Fen Bilimleri Dergisi
9(1), 286-303, 2022

BSEU Journal of Science
https://doi.org/10.35193/bseufbd.1018339

e-ISSN:2458-7575 (https://dergipark.org.tr/tr/pub/bseufbd)



Şekil 5. art Değişkenin Kantillere Göre Serpilme Diyagramı

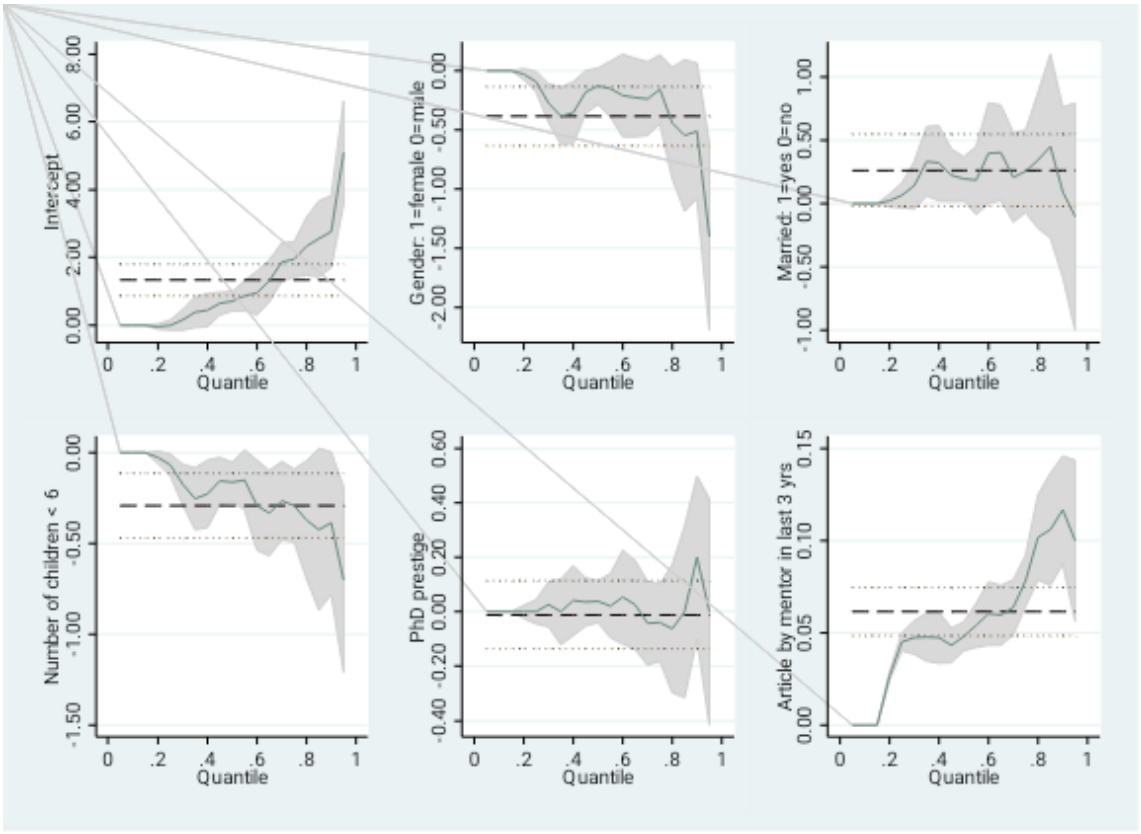
Şekil 5'de kantillere göre serpilme diyagramı verilmiştir. Buna göre bağımlı değişken çok düşük değerlerle başlıyor sonra yükseliyor. Kantil regresyon analizinden seçilen değişkenlerin tahmini katsayıları 0,10–0,90 yüzdelik dilimler olarak gösterilir (Şekil 6).

Panelli kantil işlem grafikleri, bağımlı değişkenin dağılımının farklı bölümleriyle hangi tahmincilerin ilişkili olduğunu kolayca belirlemenize yardımcı olur. Böylece bağımsız değişkenlerin her birinin etkilerine her bir kantilde bakabiliriz. Bunu yapmanın en iyi yolu grafiksel gösterimdir. Şekil 6 ile her bir değişen için kantiller gösterilmektedir. Bu kantil grafiği, kantil düzeyinin bir fonksiyonu olarak parametre tahminlerini ve %95 güven sınırlarını göstermektedir.

BŞEÜ Fen Bilimleri Dergisi
9(1), 286-303, 2022

BSEU Journal of Science
https://doi.org/10.35193/bseufbd.1018339

e-ISSN:2458-7575 (https://dergipark.org.tr/tr/pub/bseufbd)



Şekil 6. Kantil Regresyon Katsayıları

Bağımlı değişkenin kantilleri yatay eksende, katsayı büyüklükleri dikey eksendedir. EKKY katsayılarının (- - - -) kantillere göre değişmediği görülmektedir. Kantil regresyon katsayıları (kalın düz çizgiler), etraflarında güven aralıklarıyla kantiller arasında değişen çizgiler olarak çizilir. Kantil katsayısı EKKY güven aralığının dışındaysa, kantil ve EKK katsayıları arasında önemli farklar vardır. Buna göre en etkili değişken mentor değişkenidir. %50 kantilden sonra kid5 değişkenin de anlamlı olduğunu görmekteyiz. NBR modelinde cinsiyet (fem) değişkeni önemli bulunmasına rağmen QR'da da bu değişken tüm kantillerde anlamsız bulunmuştur.

VII. SONUÇLAR

Bağımlı değişken ile bağımsız değişkenler arasındaki ilişki araştırılırken en yaygın kullanılan yöntem regresyon analizidir. Regresyon analizinde tercih edilen EKKY'de bağımlı değişken sürekli olmalıdır. Sayıma dayalı olarak elde edilen sayma verileri ile analiz yapılacak olursa sapmalı sonuçlar elde edildiği bilinmektedir. Ayrıca bu analiz için uygulanan EKKY ile elde edilecek sonuçlar, varsayımlar (hata teriminin dağılımın normal olması, otokorelasyon olmaması, eşit varyans olması vb.) sağlanmadığında, güvenilir olmaz. Ayrıca gözlem değerlerinde aykırı gözlem olup olmaması da tahminlerin güvenirliliğini etkileyecektir. EKK tahmin edicisi aykırı değerlere karşı hassasiyet gösterir ve güvenilir sonuçlar vermekten uzaktır.

Sayma verileri eşit yayılım, aşırı yayılım ve eksik yayılım durumuna göre farklı modeller ile analiz edilmektedir. Sayma verilerinde en yaygın kullanılan modeller PR ve NBR modelleridir. PR modelinin uygulanabilmesi için veri kümesinin ortalama ile varyansının birbirine eşit olması gerekmektedir. Varyansın ortalamadan büyük ya da küçük olması veri kümesinde aşırı (ya da eksik) yayılım olduğunu gösterir. Bu durumda Genelleştirilmiş PR, NBR, Quasi Poisson gibi farklı sayma verisi modelleri kullanılmaktadır. Bunlar arasında NBR literatürde en çok tercih edilen modeldir. Veri kümesinde fazladan sıfırlar olursa PR ve NBR modellerinin sıfır yayılımlı (Zero Inflated-ZI) olanları ve Hurdle modeller tercih edilmelidir.

Bu çalışmada kullanılan örnek veri setinde bağımlı değişken sayma verilerinden oluşmaktadır. Bu nedenle, önce sayma verilerinde çok kullanılan PR ile analiz yapılmıştır. Fakat veri setinde aşırı yayılım tespit edildiğinden daha sonra NBR analiz uygulanmıştır. Bu sayma modelleri karşılaştırmak için log-likelihood, AIC ve BIC değerleri kullanılmıştır. NBR modelinden elde edilen değerler Poisson modelinde elde edilen sonuçlara göre daha küçük bulunmuştur (Tablo 8). Bu sonuç NBR modelin PR modelinden daha iyi bir model olduğunu göstermektedir.

Tablo 8. PR ve NBR İçin Log Likelihood, AIC, BIC Değerleri

	PR	**NBR**
Log likelihood	-1651.0563	-1560.9583,
AIC	3314.113	3135.917
BIC	3343.026	3169.649

Sayma modelleri genel olarak bağımlı değişkenin koşullu ortalamasını modellemeye dayanır. Ancak koşullu ortalama modelleri bağımlı değişkenin aykırı değerlerine duyarlı olabilir. Sayma modelleri için Poisson ve bunun uzantısı olan modeller yaygın kullanılsa da son yapılan çalışmalarda sayıma dayalı elde edilen veriler için sağlam regresyon yöntemlerinden biri olan kantil regresyonun da kullanıldığını görmekteyiz.

Kantil regresyon özellikle aykırı gözlem değerleri olduğunda EKKY gibi klasik yöntemlerden daha iyi sonuç verdiği bilinmektedir. Bu nedenle çalışmada önce aykırı gözlemlerin varlığı araştırılmıştır. Aykırı gözlemler olduğu tespit edildiğinden PR ve NBR sayma modellerinin yanında kantil regresyon analizi uygulanarak farklı kantillerdeki (%25, %50, %75 ve %90) aşırı değerler tespit edilmiştir.

KAYNAKLAR

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

[2] Cui, Y. & Yang, W. (2009). Zero-İnflated Generalized Poisson Regression Mixture Model for Mapping Quantitative Trait Loci Underlying Count Trait With Many Zeros. *Journal of Theoretical Biology*, 256, 276-285.

[3] Martin, S.W., Rose, C.E, Wannemuehler, K.A. & Plikaytis, B.D. (2006). On the of Zero-inflated and Hurdle Models for Medelling Vaccine Adverse event Count Data. *Journal of Biopharmaceutical Statistics*, 16, 463-481.

[4] Lambert, D. (1992). Zero-Inflated Poisson Regression, with An Application to Defects in Manufacturing, *Technometrics*, 34(1), 1-14.

[5] Green, W.H. (1994). Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binom Regression Models, NYU Working Paper No. EC-94-10, 1-32.

[6] Koenker, R. & Basett, G. (1978). Regression Quantiles, *Econometrica*, 46(1): 33-50.

[7] Manski, C. F. (1975). Maximum Score Estimation of the Stochastic Utility Model of Choice. *Journal of Econometrics*, 3, 205–228.

[8] Manski, C.F. (1985). Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator. *Journal of Econometrics*, 27, 313-333.

[9] Horowitz, J. L. (1992). A Smooth Maximum Score Estimator for the Binary Response Model, *Econometrica*, 60, 505-531.

[10] Horowitz, J.L. (1998). *Semiparametric Methods in Econometrics*, New York: Springer-Verlag, 100.

[11] Powell, J.L. (1984). Least Absolute Deviation Estimation for the Censored Regression Model. *Journal of Econometrics*, 25, 303-325.

[12] Powell, J.L. (1986). Censored Regression Quantiles, *Journal of Econometrics*, 32, 143-155.

[13] Lee, M.J. (1992). Median Regression for Ordered Discrete Response, *Journal of Econometrics*, 51, 59-77.

[14] Koenker, R., & Bilias, Y. (2001). Quantile Regression for Duration Data: A Reappraisal of the Pennsylvania Reemployment Bonus Experiments. *Empirical Economics*, 26, 199-220.

[15] Koenker, R., & Geling, O. (2001). Reappraising Medfly Longevity: A Quantile Regression Survival Analysis. *Journal of the American Statistical Association*, 96, 458-468.

[16] Machado, J.A.F. & Portugal, P. (2002). Exploring Transition Data through Quantile Regression Methods: An Application to U.S. Unemployment Duration. *Statistical Data Analysis Based on the L1-Norm and Related Methods*, 77-94.

[17] Machado, J.A.F & Santos Silva, J.M.C. (2005). Quantiles for Counts. *Journal of the American Statistical Association*, 100(472), 1226-1237.

[18] Wu, H., Gao, L. & Zhang, Z. (2014). Analysis of Crash Data Using Quantile Regression for Counts, *Journal of Transportation Engineering*,140(4).

[19] Congdon, P. (2017). Quantile Regression for Overdispersed Count Data: A Hierarchical Method. *Journal of Statistical Distributions and Applications*. 4(18), 1-19.

[20] Chernozhukov, V., Fernández-Val,I., Blaise Melly, B. & Kaspar Wüthrich, K. (2020). Generic Inference on Quantile and Quantile Effect Functions for Discrete Outcomes. *Journal of The American Statistical Association*, 115(5299, 123-137.

[21] Frumento, P., & Salvati, N. (2021). Parametric modeling of quantile regression coefficient functions with count data. *Statistical Methods & Applications*, 30:1237–1258.

[22] Lamarche, C., Shib, X., & Young, D.S. (2021). Conditional Quantile Functions for Zero-Inflated Longitudinal Count Data. Econometrics and Statistics (basımda) https://gattonweb.uky.edu/faculty/lamarche/ZIPQR.pdf

[23] Çınar, U.K. (2019). En Küçük Kareler Regresyonuna Alternatif Bir Yöntem: Kantil Regresyon. *Avrasya Uluslararası Araştırmalar Dergisi*, 7(18), 57-71.

[24] Yu, K., Lu, Z. & Stander, J. (2003). Quantile Regression: Applications and Current Research Areas. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52, 331-350.

[25] Saçaklı, İ., (2005). *Kantil Regresyon ve Alternatif Regresyon Modelleri ile Karşılaştırılması*. Yayınlanmış Yüksek Lisans Tezi, Marmara Üniversitesi, Sosyal Bilimler Enstitüsü, Ekonometri Anabilim Dalı, İstanbul.

[26] Sinharay, S. (2010). *Discrete Probability Distributions*. International Encyclopedia of Education (Third Edition), 1-11.

[27] Favero, L.P., Souza, R.F., Belfiore, P., Corrêa, H.L. & Haddad, M.F.C. (2021). Count Data Regression Analysis: Concepts, Overdispersion Detection, Zero-inflation Identification, and Applications with R, *Practical Assessment, Research, and Evaluation*, 26, 1-22.

[28] Yip, K.C.H. & Yau, K.K.W. (2005). On Modeling Claim Frequency Data in General Insurance With Extra Zeros. *Insurance: Mathematics and Economics*, 36(2), 153-163.

[29] Cameron, A.C. & Trivedi, P.K. (1990). Regression-based Tests for Overdispersion in the Poisson Model. *Journal of Econometrics*, 46**,** 347-364.

[30] Ismail, N. & Jemain, A.A. (2007). Handling Overdispersion with Negative Binom and Generalized Poisson Regression Models. *Virginia: Casualty Actuarial Society Forum*, 103-158.

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

[32] NNCS. (2020). *Negative Binom Regression.* NCSS Statistical Software, Chapter 326.

[33] Cameron, A.C. & Trivedi, P.K. (2013). *Regression Analysis of Count Data*, Cambridge University Press, 566.

[34] Zwilling, M.L. (2013). Negative Binom Regression, *The Mathematica Journal*, 15, 1-18.

[35] Boucher, J.P., Denuit, M. & Guillen, M. (2007). Risk Classification for Claim Counts: Mixed Poisson, Zero-Inflated Mixed Poisson and Hurdle Models. *North American Actuarial Journal*, 11(4), 110- 131.

[36] Fox J. (1997). *Applied Regression Analysis: Linear Models And Related Methods*. Sage Publication, USA, 123-240.

[37] Neter, J., Kutner, M., Nachtsheim, C. & Wasserman,W. (1996). *Applied Lineear Regression Models*, Irwin, USA, 561.

[38] Barnett. V., & Lewis, T. (1994). *Outliers in Statistical Data*. John Wiley Sons, Canada, 7–25.

[39] Çamurlu, S. & Erilli, N.A. (2019). Kantil Regresyon Analizinde Bootstrap Tahmini. *Erciyes Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 35(2), 16-25.

[40] Rousseeuw, P. & Leroy, A. (1987). *Robust Regression and Outlier Detection*. John Wiley Sons, Canada, 84-143.

[41] Wang, H. (2007). Quantile Regression: Overview and Applications to Risk Assessment. *North Caroline State University*, 1-26.

[42] Geraci, M. (2021). *Qtools: A Collection of Models and Tools for Quantile Inference*.

[43] Koenker, R. (2005). *Quantile Regression*, London: Cambridge University Press, 349.

[44] Elmalı, K. (2014). *Kantil Regresyon ve Negatif Binom Regresyon İle İllerde Kullanılan İlaç Sayısına Etki Eden Faktörlerin İncelenmesi*. Yüksek Lisans Tezi, Atatürk Üniversitesi, Ekonometri Anabilim Dalı, Erzurum.

[45] Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In Petrov B.N., & Csaki F. (Eds.), *Proceedings of the 2nd International Symposium on Information Theory*, 267-281.

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

[47] McQuarrie, A.D.R. & Tsai, C. (1998) Regression and Time Series Model Selection. World Scientific Publishing Company, Singapore, 480.

[48] 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.

[49] Long, S.J. & Freese, J. (2001). Predicted Probabilities for Count Models. *The Stata Journal*, 1, 51–57.

[50] Ucla Statistical Consulting (2021), *Poisson Regression-Stata Data Analysis Examples.* https://stats.idre.ucla.edu/stata/dae/poisson-regression