



**CONWAY MAXWELL POISSON REGRESYON MODELİNDE KANTİL
ARTIKLARIN İNCELENMESİ**

Şima ATALAY

**YÜKSEK LİSANS TEZİ
İSTATİSTİK ANA BİLİM DALI**

**GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

OCAK 2025

ETİK BEYAN

Gazi Üniversitesi Fen Bilimleri Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmada;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,
- Tez çalışmada yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- Bu tezde sunduğum çalışmanın özgün olduğunu,

bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

Şima ATALAY

10/01/2025

CONWAY MAXWELL POISSON REGRESYON MODELİNDE KANTİL ARTIKLARIN İNCELENMESİ

(Yüksek Lisans Tezi)

Şima ATALAY

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Ocak 2025

ÖZET

Regresyon analizi, bağımlı ve bağımsız değişkenler arasındaki ilişkiyi belirlemek için kullanılan bir yöntemdir. Ancak bağımlı değişkenin sayma verisi olarak kesikli bir yapıya sahip olduğu durumlarda, klasik regresyon modelleri yerine genelleştirilmiş lineer modellerin kullanımı daha uygundur. Sayma verilerinde yaygın olarak kullanılan Poisson regresyon modeli, ortalama ile varyans eşitliği varsayımına dayanır. Ancak, gözlemsel verilerin büyük bir kısmında bu varsayım sağlanamamakta ve aşırı yayılım veya az yayılım durumlarıyla karşılaşmaktadır. Modeller ile veriler arasındaki genel uyumu değerlendirmek ve veriler arasındaki tutarsızlıkları belirlemek için artık analizi yapılmaktadır. Doğrusal regresyon modellerinde Pearson ve sapma artıkları sıklıkla kullanılmaktadır. Ancak bağımlı değişken kesikli olduğunda bu artıklar normallikten uzaklaşır ve görsel inceleme için büyük zorluklar yaratır. Bu sorunları aşmak için sayma verisi modellerinde kantil artıkları kullanmak daha doğru sonuç vermektedir. Ancak, bu yaklaşım diğer yöntemlere göre popülerlik kazanmamıştır. Bu nedenle tez çalışmasında hem aşırı hem de az yayılım durumlarına uygun analiz yapma imkânı sağlayan Conway-Maxwell-Poisson (COM-Poisson) regresyon modeli ayrıntılı bir şekilde ele alınmıştır. COM-Poisson dağılımı, parametre tahmini süreçleri, anlamlılık testleri ve bu modelin yorumlanması üzerinde durulmuş ve simülasyon çalışmasıyla kantil artıkların sayma regresyon modellerini teşhis etmekteki performansı geleneksel artıklarla karşılaştırılmıştır. Ayrıca, az yayılım ve aşırı yayılım özellikleri gösteren iki farklı veri setine dayalı veri analizi çalışmaları da sunulmuştur. Veri analizi, iki yayılım durumunda da en uygun modelin COM-Poisson regresyon modeli olduğunu göstermiştir. Simülasyon çalışmaları ayrıca kantil artıkların diğer artık türlerine kıyasla sayma regresyon modellerinde birçok model uyumsuzluğunu tespit etmede etkili olduğunu, model performansını değerlendirme ve model seçiminde daha güvenilir bir yöntem olarak kullanılabileceğini göstermiştir.

Bilim Kodu : 20513

Anahtar Kelimeler : COM-Poisson Regresyon, sayma verisi, aşırı yayılım, az yayılım, artık analizi

Sayfa Adedi : 80

Danışman : Doç. Dr. Hatice Tül Kübra AKDUR

EXAMINING QUANTILE RESIDUALS
IN CONWAY MAXWELL POISSON REGRESSION MODEL

(M. Sc. Thesis)

Şima ATALAY

GAZİ UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

January 2025

ABSTRACT

Regression analysis is a method used to determine the relationship between dependent and independent variables. However, in cases where the dependent variable has a discrete structure as count data, it is more appropriate to use generalized linear models instead of classical regression models. The Poisson regression model, which is widely used in count data, is based on the assumption of equality of mean and variance. However, this assumption cannot be provided in a large part of observational data and overdispersion or underdispersion situations are encountered. Residual analysis is performed to evaluate the general fit between models and data and to determine inconsistencies between data. Pearson and deviation residuals are frequently used in linear regression models. However, when the dependent variable is discrete, these residuals are distributed far from normality and create great difficulties for visual inspection. In order to overcome these problems, it is more accurate to use quantile residuals in count data models. However, this approach has not gained popularity compared to other methods. For these reasons, the Conway-Maxwell-Poisson (COM-Poisson) Regression Model, which provides the opportunity to perform analysis suitable for both over- and under-dispersion cases, was discussed in detail in the thesis study. The COM-Poisson distribution, parameter estimation processes, significance tests and the interpretation of this model were emphasized and the performance of quantile residuals in diagnosing count regression models was compared with traditional residuals with a simulation study. In addition, data analysis studies based on two different data sets showing underdispersion and overdispersion properties are also presented. Data analysis showed that the most suitable model in both dispersion cases is the COM-Poisson regression model. Simulation studies also showed that quantile residuals are effective in detecting many model incompatibilities in count regression models compared to other types of residuals and can be used as a more reliable method in evaluating model performance and selecting models.

Science Code : 20513

Key Words : COM-Poisson Regression, count data, overdispersion, underdispersion, residual analysis

Page Number : 80

Supervisor : Assoc. Prof. Dr. Hatice Tül Kübra AKDUR

TEŞEKKÜR

Tez çalışmam boyunca bana rehberlik eden, bilgi ve deneyimlerini cömertçe paylaşarak akademik yolculuğuma ışık tutan değerli danışman hocam Doç. Dr. Hatice Tül Kübra AKDUR'a sonsuz teşekkürlerimi sunarım. Sabırla verdiğiniz destek, yol göstericiliğiniz ve çalışmama olan inancınız sayesinde bu süreci başarıyla tamamlayabildim. Hayatımın her anında olduğu gibi bu süreçte de yanımda olan sevgili eşime bana gösterdiği sabır, anlayış ve motivasyon için teşekkürlerimi sunuyorum. Ayrıca canım aileme; ablama, anne ve babama, her zaman arkamda oldukları ve beni cesaretlendirdikleri için en içten teşekkürlerimi sunuyorum. Sunduğunuz sevgi, güven ve destek, hayatım boyunca benim için en büyük güç kaynağı oldu.

İÇİNDEKİLER

	Sayfa
ÖZET	iv
ABSTRACT.....	v
TEŞEKKÜR.....	vi
İÇİNDEKİLER	vii
ÇİZELGELERİN LİSTESİ.....	ix
ŞEKİLLERİN LİSTESİ.....	xv
SİMGELER VE KISALTMALAR.....	xvi
1. GİRİŞ.....	1
2. LİTERATÜR ARAŞTIRMASI	5
3. GENELLEŞTİRİLMİŞ LİNEER MODELLER	7
4. SAYMA VERİLERİ	13
4.1. Aşırı Yayılım	14
4.2. Az Yayılım	15
4.3. Poisson Regresyon Modeli	16
4.4. Negatif Binom Regresyon Modeli	17
5. COM-POISSON DAĞILIMI, DAĞILIMSAL VE İSTATİSTİKSEL ÖZELLİKLER	19
5.1. Olasılık Dağılımı	20
5.2. Parametre Tahmini ve İstatistiksel Çıkarım	24
5.2.1. COM-Poissonness grafiğinin ağırlıklandırılmış en küçük kareler ile birleştirilmesi	24
5.2.2. En çok olabilirlik tahmini	25

Sayfa

6. COM-POISSON DAĞILIMININ PARAMETRELENDİRİLMESİ.....	29
6.1. Model Formülasyonları.....	29
6.2. COM-Poisson Parametrelelendirmesi.....	32
6.2.1. ACMP parametrelelendirmesi (Approximate COM-Poisson parametrization).....	33
6.2.2. MCMP1 parametrelelendirmesi (Mean-parametrized COM-Poisson).....	35
6.2.3. MCMP2 parametrelelendirmesi (Alternative mean-parametrized COM-Poisson)	38
7. ARTIKLARIN ANALİZİ.....	43
7.1. Pearson Artıklar.....	43
7.2. Sapma (Deviance) Artıklar	44
7.3. Kantil (Quantile) Artıklar	44
8. SİMÜLASYON ÇALIŞMASI.....	47
9. UYGULAMA.....	55
9.1. Teklif Sayısı Veri Seti.....	55
9.2. Devamsızlık Veri Seti	62
10. SONUÇ VE ÖNERİLER	69
KAYNAKLAR	71
EKLER.....	75
EK-1. Teklif sayısı veri seti R programlama kodu	76
ÖZGEÇMİŞ	80

ÇİZELGELERİN LİSTESİ

Çizelge	Sayfa
Çizelge 3.1. Üstel ailenin bazı dağılımları için kanonik bağlantı fonksiyonları ve elde edilen modeller	8
Çizelge 5.1. COM-Poisson parametrelendirme dağılımlarının özel durumları	22
Çizelge 8.1. Kullanılan parametreler	48
Çizelge 8.2. COM-Poisson, Poisson ve Negatif Binom modellerine ait AIC değerleri .	53
Çizelge 9.1. Modelde kullanılan bağımsız değişkenler ile ortalama ve standart sapmaları	56
Çizelge 9.2. Teklif sayısı verileri	57
Çizelge 9.3. COM-Poisson ve Poisson modellerine ait AIC değerleri	59
Çizelge 9.4. Artıkların Anderson-Darling normallik testi p değerleri	60
Çizelge 9.5. COM-Poisson model için parametre tahmini	61
Çizelge 9.6. COM-Poisson ve Poisson modellerinin bağımsız değişkenler üzerindeki etkileri	62

ŞEKİLLERİN LİSTESİ

Şekil	Sayfa
Şekil 5.1. COM-Poissonness grafikleri	25
Şekil 8.1. Az yayılım ($\nu = 1,5$) COM-Poisson dağılımı altında veri seti üretildiğinde sayma modelleri için artıkların Anderson-Darling testinin p-değerlerinin histogramı	48
Şekil 8.2. Az yayılım ($\nu = 2$) COM-Poisson dağılımı altında veri seti üretildiğinde sayma modelleri için artıkların Anderson-Darling testinin p-değerlerinin histogramı	49
Şekil 8.3. Aşırı yayılım ($\nu = 0.5$) COM-Poisson dağılımı altında veri seti üretildiğinde sayma modelleri için artıkların Anderson-Darling testinin p-değerlerinin histogramı.....	50
Şekil 8.4. Aşırı yayılım ($\nu = 0.8$) COM-Poisson dağılımı altında veri seti üretildiğinde sayma modelleri için artıkların Anderson-Darling testinin p-değerlerinin histogramı.....	50
Şekil 8.5 Poisson dağılımı altında veri seti üretildiğinde sayma modelleri için artıkların Anderson-Darling testinin p-değerlerinin histogramı	51
Şekil 8.6. Negatif Binom dağılımı altında veri seti üretildiğinde sayma modelleri için artıkların Anderson-Darling testinin p-değerlerinin histogramı	52
Şekil 9.1. COM-Poisson modeli artık değerlerinin QQ grafikleri	58
Şekil 9.2. Poisson modeli artık değerlerinin QQ grafikleri	59
Şekil 9.3. COM-Poisson ve Poisson modellerinin tahmin, gerçek değer ve artık analizi karşılaştırması.....	60
Şekil 9.4. Devamsızlık gün sayısı, akademik program türlerine göre gruplandırılmış olarak gösteren histogram	63
Şekil 9.5. COM-Poisson, Poisson ve NB modellerinin tahmin değerleri ve kantil artık analizi grafikleri.....	65
Şekil 9.6. COM-Poisson, Poisson ve NB modellerinin kantil artık değerlerinin QQ grafikleri.....	66

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış simgeler ve kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Simgeler

Açıklamalar

η	Doğrusal tahminleyici
θ	Kanonik parametre
μ	Ortalama
ν	Dağılım parametresi
λ	Konum parametresi
$Z(\lambda, \nu)$	Normalleştirme sabiti
$\kappa(\theta)$	Kümülant üreten fonksiyon
β	Model parametre vektörü

Kısaltmalar

Açıklamalar

ACMP	Yaklaşık COM-Poisson
AIC	Akaike Bilgi Kriteri
CMP(λ, ν)	Conway-Maxwell-Poisson dağılımı
COM-POISSON	Conway-Maxwell-Poisson
FİFO	İlk Gelen İlk Hizmet Görür
GLM	Genelleştirilmiş Lineer Modeller
GP	Genelleştirilmiş Poisson
MCMP1	Ortalama Parametrelili COM-Poisson
MCMP2	Alternatif Ortalama Parametrelili COM-Poisson
NB	Negatif Binom Dağılımı

1. GİRİŞ

Regresyon analizi, bir bağımlı değişken ve bir veya birden fazla bağımsız değişken arasındaki ilişkiyi matematiksel olarak belirlemek için kullanılan en temel yöntemidir. Bağımlı değişken sürekli ya da kesikli yapıda olabilir. Bağımlı değişkenin sürekli olduğu durumlarda klasik regresyon modelleri kullanılabilir. Ancak bağımlı değişkenin sayma verisi olduğu durumlarda, klasik regresyon modelleri yerine genelleştirilmiş lineer modellerin kullanımı daha uygundur.

Sayma verileri günlük hayatta ve işlemlerde karşımıza çıkan verilerdir. Bir markete gelen müşteri sayısı, bir hastaneye gün içinde gelen hasta sayısı, bir yoldan geçen araba sayısı sayma verilerine örnek olarak verilebilir.

Sayma verileriyle yapılan analizler, birçok araştırma alanında önemli bir yer tutmaktadır. Bağımlı değişkenin sayma verisi olduğu durumlarda, klasik regresyon modellerinin varsayımlarının yetersiz kaldığı gözlemlenmektedir.

Poisson, Negatif Binom (NB) ve Conway-Maxwell-Poisson (COM-Poisson) gibi genelleştirilmiş lineer modellerin kullanımı, verinin doğasına uygun analizler yapılmasını sağlamaktadır. Sayma verilerinde yaygın olarak kullanılan Poisson regresyon modeli, ortalama ile varyans eşitliği varsayımına dayanır. Ancak, gözlemsel verilerin büyük bir kısmında bu varsayım sağlanamamakta ve aşırı yayılım veya az yayılım durumlarıyla karşılaşmaktadır. Bu durumlarda modellerin etkinlikleri ve doğrulukları farklılık göstermektedir.

Modeller ile veriler arasındaki genel uyumu değerlendirmek ve veriler arasındaki tutarsızlıkları belirlemek için ise artık analizi yapılmaktadır. Normal doğrusal regresyon modellerinde Pearson ve sapma artıklar sıklıkla kullanılmaktadır. Ancak bağımlı değişken kesikli olduğunda artıklar normal dağılıma uymaz ve görsel inceleme için büyük zorluklar yaratır. Bu sorunları aşmak için sayma verisi modellerinde kantil artıkları kullanmak daha doğru sonuç vermektedir. Ancak, bu yaklaşım diğer yöntemlere göre popülerlik kazanmamıştır.

Bu tez çalışmasında sayma regresyon modeli olan, hem aşırı hem de az yayılım durumlarına uygun analiz yapma imkânı sağlayan COM-Poisson regresyon modeli ayrıntılı bir şekilde ele alındı. COM-Poisson dağılımı, parametre tahmini süreçleri, anlamlılık testleri ve bu modelin yorumlanması üzerinde duruldu.

Simülasyon çalışmasında, üç dağılımdan (COM-Poisson, Poisson ve negatif binom) üretilen dört farklı sayma verisi yapılarında, Pearson, sapma ve kantil artıkların performansları karşılaştırılmıştır. Simülasyon çalışması, kantil artıkların diğer artık türlerine kıyasla sayma regresyon modellerinde birçok model uyumsuzluğunu tespit etmede yüksek istatistiksel güce sahip olduğunu, model performansını değerlendirme ve model seçiminde daha güvenilir bir yöntem olarak kullanılabileceğini gösterdi.

Ayrıca uygulama bölümünde az ve aşırı yayılım gösteren iki farklı veri seti kullanılarak sayma regresyon modellerinden hangisinin veriye daha iyi uyum sağladığını belirlemek amaçlanmıştır. Artık analizi ile uygun modelin belirlenmesi sağlanmıştır. Modellerin performanslarının karşılaştırılmasında Akaike Bilgi Kriteri (AIC) kullanılmıştır.

Bu tez çalışmasının ikinci bölümünde literatür araştırmasına, üçüncü bölümünde genelleştirilmiş lineer modeller hakkında ayrıntılı bilgilere yer verildi. Dördüncü bölümünde sayma verileri tanıtıldı, aşırı yayılım, az yayılım hakkında ve Poisson regresyon ile negatif binom regresyon modeli ile ilgili bilgiler verildi. Beşinci bölümde COM-Poisson dağılımın olasılık dağılımı, parametre tahminleri, dağılımsal ve istatistiksel özellikleri üzerinde duruldu. Altıncı bölümde COM-Poisson dağılımının parametrelendirilmesi (ACMP, MCMP1, MCMP2) konularına yer verildi. Yedinci bölümde kullanılan artık türleri tanıtıldı. Sekizinci bölümde simülasyon çalışmaları detaylı bir şekilde açıklandı, COM-Poisson, Poisson ve negatif binom dağılımlarından üretilen sayma verilerine ait Pearson, sapma ve kantil artıkların performansları karşılaştırıldı ve bu çalışmalardan elde edilen sonuçlar sunuldu. Dokuzuncu bölümde iki farklı veri seti kullanılarak, sayma modellerinin az yayılıma ve aşırı yayılıma sahip veriler üzerindeki uyumları değerlendirildi. Artık analizleri uygulanarak elde edilen sonuçlar yorumlandı. Onuncu ve son bölümde tez çalışması sonucu elde edilen bulgular incelendi ve yorumlandı.

Tez çalışmasının simülasyon ve uygulama bölümlerinde R programlama dili kullanılmış olup, Huang (2017) tarafından geliştirilen ortalama parametrelili COM-Poisson (MCMP1)

parametrelendirmesine dayalı regresyon modeli ele alınmıştır. Bu parametrelendirmeye dayalı analizler için ise Fung, Alwan, Wishart ve Huang (2019) tarafından geliştirilen mpcmp paketi kullanılmıştır.

Ancak bu tez çalışmasında, ortalama parametrelili COM-Poisson (MCMP1) regresyon modeline yönelik önemli bir yenilik olarak, kantil artıkların kullanımı ilk kez önerilmektedir. Bu bağlamda, kantil artık hesaplamaları için gerekli olan kodlar tamamen özgün bir şekilde tarafımızdan yazılarak mpcmp paketi ile kullanılmıştır. Yazılan bu kodlar mevcut mpcmp paketi içeriğinde ya da herhangi bir R kütüphanesinde yer almamaktadır. Çalışmada kullanılan mpcmp paketine ek olarak, kantil artıkların hesaplanmasına ve analizine olanak sağlayan özel fonksiyonlar geliştirilmiş, analiz ve simülasyon sürecine entegre edilmiştir.

2. LİTERATÜR ARAŞTIRMASI

Sayma verilerinde yaygın olarak kullanılan Poisson regresyon modeli, ortalama ile varyans eşitliği varsayımına dayanır. Ancak, gözlemsel verilerin büyük bir kısmında bu varsayım sağlanamamakta ve aşırı yayılım veya az yayılım durumlarıyla karşılaşilmektedir. COM-Poisson regresyon modeli ise hem aşırı hem de az yayılım durumlarına uygun analiz yapma imkânı sağlayan bir model olması nedeniyle oldukça önemlidir.

Shmueli, Minka, Kadane, Borle ve Boatwright (2005) COM-Poisson dağılımının istatistiksel ve olasılıksal özelliklerini ele almış, dağılımın parametre tahminlerini elde etmek için üç yöntem açıklamıştır.

Sellers ve Shmueli (2010) COM-Poisson modelini farklı modellerle karşılaştırmış ve farklı yayılım seviyelerine sahip üç veri setini kullanarak dağılımın avantajlarını ve kullanılabilirliğini göstermiştir.

Sellers, Borle ve Shmueli (2011) COM-Poisson modelinin pazarlama, ulaşım, biyoloji gibi alanlardaki uygulamaları incelemiştir.

Hilbe'nin (2014) "Modeling Count Data" adlı kitabında sayma verilerinin nasıl seçileceği, oluşturulacağı, yorumlanacağı ve değerlendirileceği konuları ele alınmıştır.

Huang (2017) COM-Poisson dağılımının ortalama parametrelendirmesi ile nasıl daha basit ve kolay yorumlanabileceğini göstermiştir.

Ribeiro, Zeviani, Bonat, Demetrio ve Hinde (2019) COM-Poisson dağılımının yeniden parametrelendirilmesini önermiştir.

Sellers'ın (2023) "The Conway-Maxwell-Poisson Distribution" adlı kitabı, COM-Poisson modelini kapsamlı bir şekilde ele almaktadır. Ayrıca Cran R ortamında hesaplamalı örneklerle yer verilmiştir.



3. GENELLEŞTİRİLMİŞ LINEER MODELLER

Lineer ve lineer olmayan regresyon modellerinde normal dağılım önemli bir rol oynamaktadır. Hem lineer hem de lineer olmayan modellerde sonuç çıkarımı, yani model parametrelerinin tahmini, model parametreleri üzerindeki istatistiksel testler ve güven aralıkları için y bağımlı değişkeninin normal dağılıma sahip olduğu varsayılmaktadır. Ancak bu varsayımın gerçekçi olmadığı pek çok durum vardır. Bağımlı değişkenin normal dağılıma sahip olmadığı durumlarda sabit olmayan varyans ortaya çıkmaktadır. Birçok araştırmacı varyansı sabitleştirmek için bağımlı değişken üzerinde dönüşüm yapmaktadır. Sabit olmayan varyans normal olmayan bağımlı değişkenlere özgüdür. Bu durumda bağımlı değişkenin modellenmesi için bağımlı değişken dönüşüm teknikleri kullanılarak, bağımlı değişkenin normal dağılım özelliklerini kazanması sağlanabilir. Ancak, bazen normallik, sabit varyans ve basit model formu gibi istenilen özelliklerin tümü tek bir dönüşüm ile elde edilememektedir (Korucu, 2010).

Genelleştirilmiş Lineer Modeller (GLM), veri dönüşümüne güçlü bir alternatiftir. McCullagh ve Nelder (1989), Myers ve Montgomery (1997), Hamada ve Nelder (1989) deneysel çalışmalarda bağımlı değişkenin normal olmadığı durumlar için veri dönüşümüne alternatif olarak genelleştirilmiş lineer modellerin kullanılmasını önermişlerdir (Korucu, 2010).

GLM, bağımlı değişkenin normal olmadığı durumlar için kullanılan lineer modellerin bir genişlemesidir. GLM kesikli ve sürekli bağımlı değişkenler için modellerin büyük bir sınıfı olarak birleştirilmiş bir yapıyı ifade eder. Bu modeller, bağımlı değişken üzerinde normallik ve sabit varyans varsayımına ihtiyaç duymadıkları için pek çok uygulama alanında etkili olmaktadır (Khuri, 2010; Korucu, 2010). Lineer modeller;

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon$$

şeklinde verilebilir. Modelde y , bağımlı değişken, $x_1, x_2, \dots, x_k$ açıklayıcı değişken, $\beta_0 + \beta_1 + \dots + \beta_k$ bilinmeyen parametrelerin bir kümesi ve ε hata terimidir.

Bir GLM üç bileşenden meydana gelmektedir. Bu bileşenler; bağımlı değişkenin dağılımı, sistematik yapı ve bağlantı (link) fonksiyonudur.

Bağımlı değişkenin dağılımı, üstel aileye ait bir dağılımdır. $x_1, x_2, \dots, x_p$ açıklayıcı değişkenler ve β parametresi için η doğrusal tahminleyicisi, $\eta = \sum_{j=1}^p x_j \beta_j$ şeklinde tanımlanır. Bu ilişki, bağımlı değişkenin uygun aralıkta kalmasını sağlar. Ayrıca farklı türdeki bağımlı değişkenler (ikili, sayma, sürekli) için farklı bağlantı fonksiyonları kullanılarak modelin esnekliği artırılır.

GLM, bağımlı değişken ile tahmin edici değişkenler arasında esnek bir ilişki kurar ve bu ilişkiyi kurarken uygun bir bağlantı fonksiyonu kullanır. GLM'lerde gözlemlerin ortalaması ile açıklayıcı değişkenler arasındaki ilişkiyi kurmak için bağlantı fonksiyonları kullanılır.

İki tip bağlantı fonksiyonu vardır. Bunlar kanonik ve kanonik olmayan bağlantı fonksiyonlarıdır. Çizelge 3.1'de bazı dağılımların kanonik bağlantı fonksiyonları ile birlikte modeller gösterilmiştir.

Çizelge 3.1. Üstel ailenin bazı dağılımları için kanonik bağlantı fonksiyonları ve elde edilen modeller

Dağılım	Konum Parametresi	Bağlantı Fonksiyonu	Model
Normal	$\theta_i = \mu_i$	$\mu_i = x_i' \beta_i$	$\mu = x' \beta$
Binom	$\theta_i = \ln \left(\frac{\pi_i}{1 - \pi_i} \right)$	$\ln \left(\frac{\pi_i}{1 - \pi_i} \right) = x_i' \beta_i$	$\pi = \frac{\exp \{x' \beta\}}{1 + \exp \{x' \beta\}}$
Poisson	$\theta_i = \ln(\mu_i)$	$\ln(\mu_i) = x_i' \beta_i$	$\mu = \exp(x' \beta)$
Gamma ve Üstel	$\theta_i = 1 / \mu_i$	$(\mu_i)^{-1} = x_i' \beta_i$	$\mu = 1 / x' \beta$

Genelleştirilmiş lineer modellerin yapısı daha açık şekilde aşağıdaki gibi verilebilir.

1. $y_1, y_2, \dots, y_n$ sırasıyla $\mu_1, \mu_2, \dots, \mu_n$ ortalamalı, bağımsız gözlemlerdir.

2. y_i gözlemleri üstel aile üyesi olan bir dağılıma sahiptir.
3. Modelin sistematik kısmı $x_1, x_2, \dots, x_k$ açıklayıcı değişkenleri içermektedir.
4. Model $\eta = x\beta = \beta_0 + \sum_{i=1}^k \beta_i x_i$ doğrusal tahminleyici çerçevesinde inşa edilir.
5. Model bir bağlantı fonksiyonu aracılığı ile bulunur.

$\eta_i = g(\mu_i)$, $i = 1, 2, \dots, n$ şeklinde tanımlanan bağlantı fonksiyonu, ortalama ve tahmin edicilerin doğrusal kombinasyonu arasında köprü kuran bir fonksiyondur (Korucu, 2010).

Genelleştirilmiş lineer modellerde bağlantı fonksiyonu bağımlı değişkenin doğal dağılım avantajını üstlenmektedir. Burada bahsedilen doğal dağılım avantajı, bağımlı değişkenin olasılık dağılımının özelliklerine uygun bir bağlantı fonksiyonunun seçilmesiyle, modelin daha iyi bir uyum sağlamasına ve yorumlanabilirliğin artmasına yönelik bir avantajdır. Bu fonksiyonlar, bağımlı değişkenin aralığını ve modelin doğrusallığını koruyarak, GLM'nin esnek ve uyumlu olmasını sağlar. Bağlantı fonksiyonunun yanlış seçimi, yapılan analiz nedeniyle elde edilen sonuçların iyi olmamasına neden olabilir. Doğru bağlantı seçimi, modelin veriyle daha iyi uyum sağlamasını ve gerçek veri yapısını daha doğru bir şekilde yakalamasını sağlar. Ancak, genel olarak, bağlantı fonksiyonunun belirlenmesi için güç aile bağlantı fonksiyonları (power family of link) kullanılmaktadır (Korucu, 2010).

Güç aile bağlantı fonksiyonu, bağımlı değişkenin normallik varsayımlarını gerçeklememesi ile yapılan Box-Cox dönüşümü ile aynıdır. Ancak burada dönüşüm yanıt üzerinde değil ortalama üzerinde yapılmaktadır (Korucu, 2010; McCullagh, 1989; Agresti, 2002; Dobson ve Barnett, 2018).

Kanonik bağlantı fonksiyonu, belirli bir dağılım ailesi için doğal olarak türetilen bağlantı fonksiyonudur. Eğer $\eta_i = \theta_i$ olacak şekilde seçilirse, bu durumda η_i 'ye kanonik bağlantı denir. Bu fonksiyon, GLM'nin bir parçası olarak seçilen bağımlı değişkenin dağılımına bağlı olarak belirlenir. Kanonik bağlantı fonksiyonu, bağımlı değişkenin dağılımı ile doğal bir bağlantı kurar. Örneğin, Poisson dağılımı için kanonik bağlantı fonksiyonu logaritmik (log) fonksiyondur. Genel bağlantı fonksiyonları ise bağımlı değişkenin aralığını ve doğrusallığını korumak için belirli bir esneklik sağlar, ancak her zaman dağılımla doğal bir bağlantı

kurmaz. Kanonik bağlantı fonksiyonları, matematiksel olarak daha basit ve yorumlanması daha kolay sonuçlar üretir. Bu, modelin parametrik formlarının daha doğrudan yorumlanmasına olanak tanır. Genel bağlantı fonksiyonları, matematiksel olarak daha karmaşık olabilir ve bazen modelin yorumlanmasını zorlaştırabilir. Kanonik bağlantı fonksiyonları, model tahminlerinde genellikle daha verimli sonuçlar üretir ve algoritmik olarak daha kolay hesaplanır. Genel bağlantı fonksiyonları, bazı durumlarda tahminlerin daha az verimli olmasına neden olabilir (McCullagh, 1989; Agresti, 2002; Dobson ve Barnett, 2018).

$$f_Y(y; \theta) = c(y, \lambda) \exp(\lambda \{\theta y - \kappa(\theta)\}) \quad (3.1)$$

Bir olasılık yoğunluk fonksiyonunun Eş. 3.1’de belirtildiği gibi yazılması üstel dağılım ailesi olarak adlandırılır. $\lambda > 0$ dağılım parametresine kesinlik ya da indeks parametresi denir. Dağılım ailesinin temel fikri, kümülanant üreten fonksiyon tarafından tanımlanan dağılımsal özelliklerle ilgili ortalama değeri, örneklem büyüklüğü, ortak varyans veya ortak aşırı yayılım gibi ortalama değerle ilgili olmayan dağılım özelliklerinden ayırmaktır. Bu formülasyonlar Bernoulli, Poisson, normal dağılım gibi birçok dağılımın bir çerçevede ifade edilmesini sağlar (Madsen ve Thyregod, 2011: 90, 91).

$$f_Y(y; \theta) = c(y) \exp(\theta y - \kappa(\theta)), \quad \theta \in \Omega \quad (3.2)$$

Bir olasılık yoğunluk fonksiyonu Eş. 3.2’de belirtildiği gibi yazılabiliyorsa bu, doğal üstel aile olarak adlandırılır. $c(y)$ gözlemlere bağlı bir ağırlık fonksiyonu, $\kappa(\theta)$ kümülanant üreten fonksiyon, θ ise kanonik parametre olarak adlandırılır. θ , dağılımın şekli ve yapısı üzerinde önemli bir rol oynar (Sellers, 2023: 17).

Doğal üstel aile önemli bir dağılım sınıfıdır. Bu aile, Bernoulli, binom, Poisson gibi birçok bilinen dağılımı içerir ve istatistiksel modelleme, özellikle genelleştirilmiş doğrusal modeller (GLM) gibi alanlarda yaygın olarak kullanılır. İstatistiksel modelleme için güçlü bir çerçeve sunar. Üstel ailenin özel bir durumudur. Doğal üstel aileler, üstel ailelerin bir alt kümesidir. Doğal üstel ailedeki dağılımlar, yeterli istatistikler ve kanonik form özellikleri sayesinde geniş bir uygulama yelpazesinde etkilidir.

Doğal üstel ailelerin, varyans fonksiyonlarıyla karakterize edildiği bilinmektedir. Doğal üstel ailelerin varyans fonksiyonlarına göre sınıflandırmasına yönelik bir dizi makale yazılmıştır. Birkaç sınıflandırma elde edilmiştir. Bu bağlamda doğal üstel bir ailenin varyans fonksiyonu en uygun araç olarak görünmektedir, bu nedenle istatistik teorisinde büyük ilgi görmüştür (Masmoudi, 2007).

COM-Poisson dağılımı için yaygın olarak kullanılan bağlantı fonksiyonları arasında logaritmik (log) fonksiyon bulunur. Log bağlantı fonksiyonu $g(\mu) = \log(\mu)$ eşitliği ile tanımlanır. Kanonik bağlantı fonksiyonu, bir dağılımın doğal bağlantısını kuran ve log-olabilirlik fonksiyonunun doğrusal bir fonksiyon olarak ifade edilmesini sağlayan fonksiyondur. Poisson dağılımı için kanonik bağlantı fonksiyonu log-bağlantı fonksiyonudur, bu fonksiyon doğal parametre ile doğrusal tahminleyici arasında doğrudan bağlantı kurduğu için kanoniktir, ancak COM-Poisson dağılımı daha karmaşık bir yapıya sahip olduğu için COM-Poisson dağılımında durum farklıdır. COM-Poisson dağılımında kullanılan bağlantı fonksiyonu, kanonik bir bağlantı fonksiyonu değildir. COM-Poisson dağılımında, modelin esnekliği için ek bir parametre gerekli olduğundan, standart GLM çerçevesinde tek bir doğal parametre ile uyum sağlayan bir kanonik bağlantı fonksiyonu oluşturmak zorlaşır. COM-Poisson dağılımında, bağlantı fonksiyonu genellikle logaritmik formda ifade edilir, çünkü bu form parametrelerin daha doğal ve kolay yorumlanmasını sağlar.

$$\log(\mu) = X \beta$$

Burada μ bağımlı değişkenin ortalaması, X bağımsız değişkenlerin matrisi, β model parametrelerinin vektörüdür.

Genelleştirilmiş Lineer Modellerde parametre kestirimleri en çok olabilirlik fonksiyonuna dayanmaktadır. Parametre kestirimlerinin en çok olabilirlik tahminleyicileri Newton-Raphson yöntemi ve Fisher-Scoring algoritması kullanılarak bulunmaktadır (Korucu, 2010).

COM-Poisson Regresyon modeli, verinin az veya aşırı yayılım gösterdiği durumlarda kullanılmak üzere geliştirilmiş bir modeldir. Conway-Maxwell-Poisson (COM-Poisson) dağılımı olarak adlandırılan Poisson dağılımının iki parametrelili genelleştirilmiş formunun,

geniş bir sayma verisi yayılım aralığını tanımlamak için yeterince esnek olduğu belirtilmiştir. COM-Poisson dağılımı Poisson, Bernoulli ve geometrik dağılımların yanı sıra bu dağılımlar arasındaki yayılım seviyelerine sahip dağılımları da içermektedir. Ayrıca COM-Poisson dağılımının üstel aileye ait olduğu ve bu nedenle tahmin, eşlenik önsel gibi avantajlara sahip olduğu belirtilmiştir (Sellers ve Shmueli, 2010).

CMP (λ, ν) dağılımı Eş. 3.3'de tanımlandığı gibi üstel aile formuna sahiptir. Burada x ve $\ln(x!)$ sırasıyla λ ve ν için ortak yeterli istatistiklerdir (Sellers, 2023: 25).

$$P(X = x; \theta, \phi) = \Gamma(x; \phi) \exp[\eta(\theta)T(x) - \Psi(\theta; \phi)], \quad x \in \mathbb{N} \quad (3.3)$$

Üstel aile formu, geniş bir dağılım sınıfını kapsayan ve istatistiksel analizlerde çeşitli kolaylıklar sağlayan bir yapıdır. Bu yapının avantajları arasında şunlar bulunmaktadır:

1. Tahmin kolaylığı: Üstel aile formundaki dağılımlar, maksimum olabilirlik tahmin edicilerinin kolayca hesaplanabilmesi gibi hesaplama kolaylıkları sunar.
2. Eşlenik önsel: Bayesci istatistikte, eşlenik önsel dağılımların kullanımı daha kolaydır ve güncellenmiş tahminler elde etmek daha pratiktir.
3. Lineer model uyumlaması: Üstel aile formu, genelleştirilmiş lineer modeller gibi yöntemlere kolayca entegre edilebilir.

COM-Poisson dağılımının üstel aile formuna ait olduğu bilgisi, bu dağılımın hesaplama ve analiz açısından önemli avantajlara sahip olduğunu ifade etmektedir.

4. SAYMA VERİLERİ

Herhangi bir olayın belirlenen süreç içerisinde yapılan denemeler sonucunda meydana gelme sayısına sayma verisi denir. Sayma verileri, sıfırdan büyük belirlenmemiş bir değere kadar değişen, negatif olmayan tamsayı değerlerine ait gözlemleri ifade etmektedir. Teorik olarak bu veriler sıfırdan sonsuza kadar değişebilmektedir ancak genellikle modellenen sayma verileri maksimum bir değere sahiptirler (Hilbe, 2014: 2).

Sayma verilerinin modellenmesi aktüerya bilimlerinde yaygın olarak kullanılmıştır, Aitkin ve diğerleri (1989) ve Renshaw (1994), Poisson regresyonunu iki farklı Birleşik Krallık motorlu taşıt hasar verisine uyarlamıştır. Biyoistatistik ve demografide Frome (1983), düzenli sigara içen İngiliz doktorlar arasındaki akciğer kanseri ölüm oranlarını modellemiştir. Son yıllarda bu model ekonomi, siyaset bilimi ve sosyolojide sıklıkla kullanılmıştır (Çelik, 2019). Sayma veri regresyon modelleri kesikli regresyonun özel bir türüdür. Negatif olmayan kesikli değerlerden oluşur. Bu verilerde doğrusal regresyon modelinin uygulanması tahmin edilen katsayıların yanlış olmasına neden olmaktadır. Sayma verisi için kullanılan en yaygın regresyon modeli "Poisson Regresyon", varyansların eşitliği varsayımı sağlanmadığında ise "Negatif Binom Regresyon" modelidir.

Poisson dağılımı, sayma verilerini modellemek için yaygın olarak kullanılan bir dağılımdır. Poisson regresyon modelinin avantajlarına rağmen, Poisson dağılımının yalnızca bir parametresi vardır ve bu parametre sayma rastgele değişkeninin hem ortalamasını hem de varyansını temsil eder. Poisson dağılımının neden olduğu ortalama ve varyans arasındaki ilişkiye getirilen bu kısıtlamaya eşit yayılım denir. Ancak, uygulamada böyle bir kısıtlama uygun olmayabilir, çünkü gözlenen verilerin varyansı ortalamadan daha küçük ise az yayılım, ortalamadan daha büyük ise aşırı yayılıma neden olabilir. Bu da Poisson dağılımını, aşırı ve az yayılım gösteren verileri modellemek için ideal olmaktan uzaklaştırmaktadır. Poisson regresyon modelinin eşit dağılımlı olmayan sayma verilerine uygulanmasının temel sorunu, regresyon katsayılarıyla ilişkili standart hataların tutarsız bir şekilde tahmin edilmesidir ve bu da yanıltıcı çıkarımlara yol açabilir (Ribeiro Jr., Zeviani, Bonat, Demetrio ve Hinde 2019). Bu durum, aşırı yayılımı yakalayabilen negatif binom dağılımının popülerleşmesine neden olmuştur. Başlangıçta negatif binom dağılımının kullanılması bazı hesaplama zorlukları oluştursa da, günümüzde bu sorun ortadan kalmış olup negatif binom dağılımı ve regresyon, istatistiksel yazılım paketlerine dahil edilmiştir (Sellers, Borle ve

Shmueli, 2011). Negatif binom dağılımı, aşırı yayılım gösteren veriler için çözüm sağlamaktadır ancak her ne kadar literatürde aşırı yayılıma ilişkin daha fazla örnek bulunsa da az yayılım durumu da oldukça yaygındır. Örnekler arasında Birleşik Krallık kömür madenciliği endüstrisinde 1948 ile 1959 arasındaki ardışık dönemlerde meydana gelen grev sayısı ve bir kuş türü için yuva başına düşen yumurta sayısı yer almaktadır (Sellers ve diğerleri, 2011).

Bu gibi durumlarda ne Poisson ne de negatif binom dağılımları yeterli yaklaşımlar sağlamamaktadır. Hem aşırı yayılım hem de az yayılım gösteren verileri modellemek için çeşitli dağılımlar önerilmiştir. Bunlar, del Castillo ve Perez-Casany (2005)'nin ağırlıklı Poisson dağılımlarını ve Consul (1989)'un Genelleştirilmiş Poisson (GP) dağılımını içerir. Her ikisi de ek bir parametrenin eklendiği Poisson dağılımını genellemektedirler. Ancak GP modelinin eksikliği, dağılım parametresine ilişkin belirli koşullar altında dağılımın kesilmesi nedeniyle bazı yayılım seviyelerini yakalayamaması ve dolayısıyla gerçek bir olasılık modeli olmamasıdır (Sellers ve diğerleri, 2011).

4.1. Aşırı Yayılım

Aşırı yayılım (karşılaştırılabilir bir Poisson modeline göre), varyansı ortalamadan daha büyük olan dağılımları tanımlar (Sellers, 2023: 5). Poisson dağılımında ortalama ve varyans eşittir. Ancak aşırı yayılım durumunda varyans ortalamadan büyüktür. Aşırı yayılım, veri heterojenliği (gözlemler arasındaki farklılıklar), bağımlı değişkenler arasındaki pozitif korelasyonun varlığı (bir gözlemin sonucu diğerini etkiliyorsa), bağımlı değişken olasılıkları veya sayıları arasındaki aşırı değişkenlik (bazı gözlemler beklenenden çok daha yüksek veya düşük değerler alabilir) ve veri dağılım varsayımlarındaki ihlaller dâhil olmak üzere birçok durumdan kaynaklanabilir. Veri setlerinde aykırı değerler nedeniyle de belirgin aşırı yayılım mevcut olabilir veya regresyon modellerinde model, önemli açıklayıcı değişkenleri veya yeterli sayıda etkileşim terimini içermeyebilir veya açıklayıcı değişkenler ile bağımlı değişken arasındaki ilişkiyi tanımlayan bağlantı fonksiyonu doğru seçilmemiş olabilir. Bu koşullar altında, aşırı dağılım sorunlara neden olur çünkü parametre tahminiyle ilişkili standart hatalar eksik tahmin edilebilir ve dolayısıyla yanlış çıkarımlar yapılabilir (Sellers, 2023: 5, 6). Aşırı yayılım gösteren verileri tanımlayan en popüler dağılım negatif binom dağılımıdır ancak az yayılıma sahip verileri ele alamaz (Sellers, 2023: 18).

Conway-Maxwell Poisson dağılımında aşırı yayılım, dağılım parametresi olan v parametresinin 1'den küçük olduğu durumlarda gözlemlenir. $v < 1$ olduğunda, dağılım Poisson dağılımından daha fazla bir varyansa sahip olur ve bu da aşırı yayılımı temsil eder. Bu, COM-Poisson dağılımının önemli avantajlarından biridir; çünkü bu dağılım, Poisson dağılımının sınırlamalarını aşarak daha geniş bir varyans aralığını modelleyebilir.

Lord, Guikema ve Geedipally (2008), COM-Poisson GLM'yi aşırı yayılım gösteren motorlu taşıt kazası veri setinde kullanarak, negatif binom ve COM-Poisson dağılımlarını karşılaştırmış, araştırma sonucunda COM-Poisson modelinin negatif binom modeline göre daha iyi bir alternatif sunduğu sonucuna varılmıştır.

4.2. Az Yayılım

Verilerin aşırı yayılımı sayma verilerindeki aşırı değişkenliği tanımlarken, az yayılım sayma verilerindeki yetersiz değişkenliği ifade eder. Az yayımlı veriler (Poisson modeline göre) varyansın ortalamadan daha küçük olacağı şekilde dağılan sayma verilerini ifade eder.

Bazı araştırmacılar, az yayılımı veri üretimine (örneğin, küçük örnek değerleri) veya modelleme sürecine (örneğin, modelin aşırı uyumu) bağlayarak varış sürecinin, doğum-ölüm sürecinin veya binom inceleme mekanizmalarının da yetersiz dağılıma yol açabileceğini belirtmektedir (Sellers, 2023: 10).

Veri yayılımını (aşırı veya az yayılım) barındırabilen en popüler model, özel bir durum olarak Poisson dağılımını içeren esnek, iki parametrelili bir dağılım olan Genelleştirilmiş Poisson dağılımıdır (Sellers, 2023: 10).

GP dağılımı aşırı veya az yayılıma izin verir; ancak aşırı az yayılım, temel olasılık aksiyomlarını karşılamayan olasılık modelleriyle sonuçlanabilir (Sellers, 2023: 11).

Ridout ve Besbeas (2004), az yayılım gösteren verileri modellemek için COM-Poisson dağılımını da içeren çeşitli ağırlıklı Poisson dağılımlarını ele almıştır. Çalışmada belirli bir kuş türüne ait her yuvada bulunan yumurta sayılarını kullanılmıştır. Araştırma sonucunda COM-Poisson modelinin veriye uygunluk açısından mükemmel sonuç vermediği, modelin veriye yeterince iyi uyum sağlamadığı, ancak diğer birçok ağırlıklı Poisson dağılımına göre

daha iyi performans sergilediği, bu bulgunun ise COM-Poisson modelinin düşük yayılımı ele alabilme kapasitesini artırdığı, ancak belirli veri setlerinde sınırlı bir uygunluğa sahip olabileceği ortaya koyulmaktadır.

4.3. Poisson Regresyon Modeli

Poisson regresyon, genelleştirilmiş lineer modellerden biridir ve sayma verisinin analizi için kullanılan en yaygın regresyon modelidir.

X değişkenleri verilmişken Poisson dağılımına uyan y değişkeninin olasılık fonksiyonu;

$$\Pr[Y = y] = \frac{e^{-\mu} \mu^y}{y!}, \quad y = 0, 1, 2, \dots$$

şeklinde gösterilir (Cameron ve Trivedi, 2013: 10). Poisson dağılımı, dağılım parametresi μ olan tek parametrelili bir dağılımdır. Bağımlı değişken y , belirli bir zaman aralığı içerisindeki olay sayısını, X ise y 'yi belirlediği düşünülen doğrusal bağımsız değişkenlerin vektörünü göstermektedir. Poisson regresyon analizi bir olayın gerçekleşme sayısını tahmin etmek için kullanılan istatistiksel bir yöntemdir. Log-lineer model için ortalama parametresi $\mu = \exp(X\beta)$ üstel ortalama fonksiyonu olarak adlandırılmaktadır.

Poisson regresyon modelinin log-olabilirlik fonksiyonu;

$$\ln L(\beta) = \sum_{i=1}^n \{y_i X_i' \beta - \exp(X_i' \beta) - \ln y_i!\}$$

eşitliği ile gösterilir.

Eğer dağılımda aşırı yayılım durumu varsa, Poisson regresyon modeli sonucu elde edilen tahminler tutarlı olacaktır ancak tahmin edicilerin standart hataları yanlış olacaktır. Bu nedenle aşırı yayılım durumunda negatif binom regresyon modeli kullanılır.

4.4. Negatif Binom Regresyon Modeli

Negatif binom regresyonunda bağımlı değişken negatif binom dağılımından gelmektedir (Cameron ve Trivedi, 2013: 80). Negatif binom modeli, aşırı yayılmış verilerin modellenmesinde tek parametrelili Poisson modelinden daha fazla esneklik sağlamaktadır. Negatif binom regresyon modelinin μ ortalama ve $\mu(1 + \alpha\mu)$ varyans ile Poisson– gamma modelinin karışımı olduğu kabul edilmektedir. Ortalama (μ) ve yayılım (α) parametrelerine sahiptir. α sıfıra yaklaştıkça model Poisson modeline yaklaşmaktadır.

Negatif binom dağılımının en yaygın uygulaması NB2 modelidir, bu modelin olasılık fonksiyonu;

$$f(y; \mu, \alpha) = \binom{y + \frac{1}{\alpha} - 1}{\frac{1}{\alpha} - 1} \left(\frac{1}{1 + \alpha\mu} \right)^{\frac{1}{\alpha}} \left(\frac{\alpha\mu}{1 + \alpha\mu} \right)^y$$

şeklinde gösterilir. $\alpha = 0$ olduğunda Poisson dağılımına indirgenir. Negatif binom dağılımının log-olabilirlik fonksiyonu ise aşağıdaki eşitlikte verilmiştir (Hilbe, 2014: 130).

$$L(y; \mu, \alpha) = \sum_{i=1}^n y_i \log \left(\frac{\alpha\mu_i}{1 + \alpha\mu_i} \right) - \frac{1}{\alpha} \log(1 + \alpha\mu_i) + \log \Gamma \left(y_i + \frac{1}{\alpha} \right) - \log \Gamma(y_i + 1) - \log \Gamma \left(\frac{1}{\alpha} \right)$$

Negatif binom regresyonu kullanmanın en önemli amacı varyans değerinin ortalamayı aştığı veya gözlemlenen varyansın beklenen varyansı aştığı verileri modellemektir (Kılıç, 2023).



5. COM-POISSON DAĞILIMI, DAĞILIMSAL VE İSTATİSTİKSEL ÖZELLİKLER

Conway - Maxwell – Poisson dağılımı ilk kez Richard W. Conway ve William L. Maxwell tarafından 1962 yılında esnek bir kuyruk sistemi geliştirme amacıyla ortaya çıkmıştır (Conway, Maxwell, 1962). Kuyruk teorisi, bekleme hatlarının ya da kuyukların matematiksel anlamda incelenmesidir. Kuyruk teorisi yöneylem araştırmasının bir kolu olarak kuyruk modellemesiyle ilgilenir. Genellikle işletmelerde, müşterilere yönelik, servis sağlayıcı kaynak ihtiyacına karar verme aşamalarında sıklıkla kullanılmaktadır. Kuyruk teorisinin temel amacı, müşterilerin servis ihtiyacına yönelik bir analitik ya da matematiksel model oluşturmak ve bu model sayesinde kuyruk uzunluğunu ve bekleme sürelerini tahmin etmektir (Sundarapandian, 2009: 686; Tekin, 2011). Bir kuyruk sisteminde, hizmet görmek için hizmet noktasına sabit veya değişken oranda gelen müşteriler vardır. Müşteriler bir insan olabileceği gibi tamir için gelen bir araç veya makine, pistte iniş için veya bakım için bekleyen bir uçak, işlem için gelen bir sipariş de olabilir. Eğer müşteriler geldiklerinde hizmet noktasına girebiliyorlarsa hizmet alırlar. Eğer hizmet için beklemleri gerekiyorsa, kuyruğa girerler ve hizmet görene dek kuyrukta beklerler. Daha sonra sabit ve değişken oranda hizmet görürler ve sistemi terk ederler. Kuyruk sisteminde, hem bekleme hem de hizmet noktaları vardır (Ada, 1990: 1). Hizmet görmek amacıyla sisteme gelişler, düzgün gelişler, tamamen tesadüfi gelişler, genel bağımsız gelişler vb. şekillerde karşımıza çıkar (Kara, 1976: 13). Sisteme gelen ve servis görmek için bekleyen müşterilerden seçilenlerin sırasını ifade eden kuyruk disiplini, kuyruk sistemlerinin analizinde önemli bir faktördür. Yaygın olarak kullanılan kuyruk disiplini İlk Gelen İlk Hizmet Görür (FIFO) disiplini'dir. Kuyruk disiplini seçilecek bir sonraki müşterinin hangi kurala göre servise alınacağını ifade eder. FIFO kuyruk disiplini müşterilere öncelik vermeden geliş sıralarına göre hizmet verilmesini sağlar (Tekin, 2011).

Her ne kadar bu dağılım sayma verileri için esnek olması yönüyle önemli bir gelişme olsa da uzun bir süre bu dağılımın olasılık ve istatistik özellikleri araştırılmamıştır. Shmueli ve diğerlerinin (2005) yayını ile beraber dağılımın gelişimi de artmıştır (Sellers ve diğerleri, 2011).

Conway ve Maxwell (1962), rastgele geliş zamanlarına ve gelişler arası sürelerinin λ ortalamalı üstel dağılımlı olduğu ilk gelen ilk hizmet alır politikasına sahip tek kuyruklu ve servis kanalı sayısının tek olduğu bir sistemi ele almaktadır (Sellers, 2023: 23).

COM-Poisson dağılımı, olağan Poisson ortalaması λ 'ya ek olarak ikinci bir parametre olan ν parametresinin eklenerek Poisson dağılımını genelleştiren ve aşırı yayılım ile az yayılım gösteren verileri modellemek için kullanılan kesikli bir olasılık dağılımıdır. CMP (λ, ν) dağılımı üstel ailenin bir üyesidir.

5.1. Olasılık Dağılımı

Com Poisson dağılımı iki parametrelili, esnek bir dağılımdır. Aşırı ve az yayılıma izin verir. COM-Poisson olasılık kütle fonksiyonu aşağıdaki gibi tanımlanır.

$$P(X = x) = \frac{\lambda^x}{(x!)^\nu Z(\lambda, \nu)} \quad x = 0, 1, 2, \dots, \quad (5.1)$$

Olasılık kütle fonksiyonunun logaritması alındığında,

$$\begin{aligned} \ln [P(X = x)] &= x \ln \lambda - \nu \ln(x!) - \ln[Z(\lambda, \nu)] \\ &= x \ln \lambda - \nu \sum_{j=1}^x \ln(j) - \ln[Z(\lambda, \nu)] \end{aligned} \quad (5.2)$$

eşitliği elde edilir.

Eş. 5.1'de X rastgele değişkeni, $Z(\lambda, \nu)$ normalleştirme sabitini, ν dağılım parametresini temsil etmektedir. λ , gözlemlerin ortalamasıyla doğrudan ilişkili olan bir merkezleme parametresidir.

Normalleştirme sabiti $Z(\lambda, \nu) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\nu} = \sum_{j=0}^M \exp[j(\log(\lambda) - \nu \log(j!))]$, $\lambda > 0$ ve $\nu \geq 0$ ile tanımlanır. Veri $\nu > 1$ olduğunda az, $\nu < 1$ olduğunda aşırı yayılım göstermektedir. COM-

Poisson yoğunluk veya konum parametresi $\lambda = E(X^v)$, klasik Poisson dağılımındaki ortalama parametresinin ($\lambda = E(X)$) genelleştirilmiş halidir. COM-Poisson dağılımında bulunan x ve $\ln(x!)$ sırasıyla λ ve v için ortak yeterli istatistikleri belirtmektedir.

n sayıdaki gözlemin olabilirlik ve log-olabilirlik fonksiyonu sırasıyla;

$$L(\lambda, v; x) = \prod_{i=1}^n \frac{\lambda^{x_i}}{(x_i!)^v Z(\lambda, v)} = \frac{\lambda^{\sum_{i=1}^n x_i}}{[Z(\lambda, v)]^n \left(\prod_{i=1}^n x_i! \right)^{-v}} \quad (5.3)$$

$$\ln L(\lambda, v; x) = \sum_{i=1}^n x_i \ln \lambda - v \sum_{i=1}^n \ln(x_i!) - n \ln Z(\lambda, v) \quad (5.4)$$

olarak tanımlanır. (Sellers, 2023: 25).

COM-Poisson dağılımı, bilinen bazı kesikli dağılımların genelleştirilmiş halidir. Normalleştirme sabiti $Z(\lambda, v)$, sonsuz toplamın tam olarak belirlenebileceği üç özel durum içermektedir. Çizelge 5.1, $Z(\lambda, v)$ için çözümleri ve ortaya çıkan ilgili dağılımları özetlemektedir (Sellers, 2023: 26).

Çizelge 5.1. COM-Poisson parametrelendirme dağılımlarının özel durumları

Kısıt	$Z(\lambda, \nu)$	Olasılık Kütle Fonksiyonu	Ortalama	Dağılım
$\nu = 1$	e^λ	$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, 2, \dots,$	λ	Poisson (λ)
$\nu = 0$ $\lambda < 1$	$\frac{1}{1-\lambda}$	$P(X = x) = (1-\lambda)\lambda^x, \quad x = 0, 1, 2, \dots$	$\frac{\lambda}{1-\lambda}$	Geometrik ($p = 1-\lambda$)
$\nu = 0$	$1+\lambda$	$P(X = x) = \begin{cases} \frac{1}{1+\lambda} & x = 0 \\ \frac{\lambda}{1+\lambda} & x = 1 \\ 0 & x = 2, 3, \dots \end{cases}$	$\frac{\lambda}{1+\lambda}$	Bernoulli $\left(p = \frac{\lambda}{1+\lambda} \right)$

Normalleştirme sabiti için Çizelge 5.1'deki sonuçlar göz önüne alındığında, COM-Poisson dağılımının üç klasik özel durum dağılımı içerdiği sonucuna varılabilir. $\nu = 1$ için, ortalama parametresi λ olan Poisson modeli, $\nu = 0$, $\lambda < 1$ için başarı olasılığı $1-\lambda$ olan geometrik dağılım ve $\nu \rightarrow \infty$ için başarı olasılığı $\lambda / (1+\lambda)$ olan Bernoulli dağılımı ortaya çıkar.

COM-Poisson dağılımının üretim fonksiyonları Z fonksiyon gösterimleri cinsinden tanımlanır. Bir rastgele değişken X için olasılık üreten fonksiyon $\Pi_x(t) = E(t^x) = \frac{Z(\lambda t, \nu)}{Z(\lambda, \nu)}$

eşitliği ile, moment üreten fonksiyonu ise $M_x(t) = E(e^{xt}) = \frac{Z(\lambda e^t, \nu)}{Z(\lambda, \nu)}$ şeklinde belirtilir.

Dağılımın beklenen değeri ve varyansının, $\log \lambda$ 'ya göre türevleri alınarak sunulan formları

$$E(X) = \frac{\partial \log Z(\lambda, \nu)}{\partial \log \lambda} \quad (5.5)$$

$$Var(X) = \frac{\partial E(X)}{\partial \log \lambda} \quad (5.6)$$

Eş. 5.5 ve Eş. 5.6'da gösterildiği gibi verilmiştir (Sellers ve diğerleri, 2011).

Beklenen değer ve varyans için açık formüller ise

$$E(X) = \frac{\partial \ln Z(\lambda, \nu)}{\partial \ln \lambda} = \lambda \frac{\partial \ln Z(\lambda, \nu)}{\partial \lambda} \quad (5.7)$$

$$Var(X) = \frac{\partial^2 \ln Z(\lambda, \nu)}{\partial (\ln \lambda)^2} = \frac{\partial E(x)}{\partial \ln \lambda} \quad (5.8)$$

olarak verilmiştir.

COM-Poisson dağılımının momentler için özyinelemeli formu ;

$$E(X^{r+1}) = \begin{cases} \lambda [E(X+1)]^{1-\nu} & r = 0 \\ \lambda \frac{\partial}{\partial \lambda} E(X^r) + E(X)E(X^r) & r > 0 \end{cases}$$

şeklinde gösterilir. $E(X^r) = \sum_{x=0}^{\infty} x^r P(X=x)$, $P(X=x)$ olasılığına sahip bir X rastgele

değişkeninin $E(X^r)$ momentini ifade eder.

COM-Poisson dağılımı, iki parametrelili güç serisi dağılımları sınıfına dahil olup, değiştirilmiş güç serisi dağılımının bir özel hali olarak tanımlanmaktadır, yani olasılık kütle fonksiyonu Eş. 5.9'da belirtildiği gibi olan bir dağılımdır:

$$P(X=x) = \frac{A(x)(g(\theta))^x}{f(\theta)}, \quad x \in B, \quad (5.9)$$

B , negatif olmayan tam sayılar kümesinin bir alt kümesidir. $A(x) > 0$, $f(\theta)$ ve $g(\theta)$ pozitif, sonlu, türevlenebilir fonksiyonlardır.

5.2. Parametre Tahmini ve İstatistiksel Çıkarım

COM-Poisson dağılım parametrelerini tahmin etmek için üç yöntem mevcuttur. Birincisi, basit bir grafik tekniğinin ağırlıklandırılmış en küçük kareler yöntemiyle birleştirilmesi, ikincisi, en çok olabilirlik yöntemi, üçüncüsü ise Bayesci yöntemidir (Shmueli ve diğerleri, 2005).

5.2.1.COM-Poissonness grafiğinin ağırlıklandırılmış en küçük kareler ile birleştirilmesi

Shmueli ve diğerleri (2005) tarafından ağırlıklandırılmış en küçük kareler yönteminin basit ve hesaplama açısından verimli olduğu, COM-Poisson dağılımının bazı veri setleri için uygun bir model olup olmadığının belirlenmesine ve eğer öyleyse ν ve λ parametrelerinin basit bir şekilde tahmin edilmesine olanak sağladığı belirtilmiştir. Ayrıca en çok olabilirlik yönteminde başlangıç değerlerini belirlemek için kullanışlı olabilir.

Bu yöntem COM-Poisson dağılımının ardışık olasılıkları arasındaki ilişkinin lineer olmasına dayanmaktadır. Ardışık olasılık oranlarının özelliklerini inceleyerek, bir dağılımın COM-Poisson dağılımına uygunluğunu test etmeye yarar. COM-Poisson ardışık olasılıkları

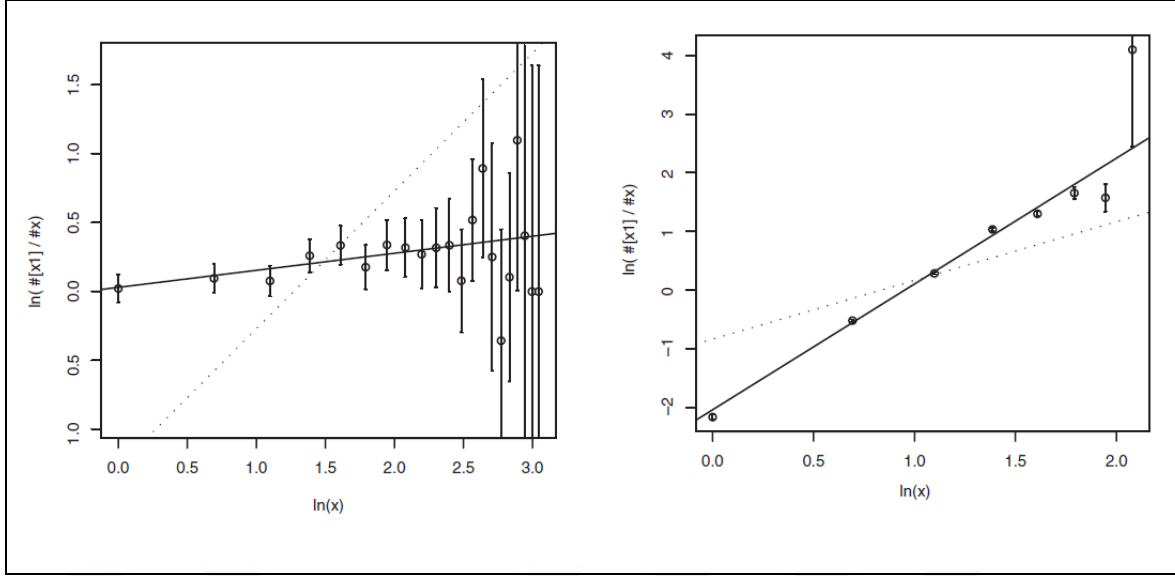
arasındaki ilişki $\frac{P(X = x-1)}{P(X = x)} = \frac{x^\nu}{\lambda}$ şeklinde tanımlanmaktadır. Bu oran ν parametresine

bağlı olarak değişir ve COM-Poisson dağılımının esnekliğini yansıtır. Bu denklemin her iki tarafının logaritması alınarak, log oranları ve $\log(x)$ arasında doğrusal bir ilişki elde edilir.

$$\log(p_{x-1} / p_x) = -\log(\lambda) + \nu \log(x)$$

p_x , $P(X = x)$ ’i belirtir. $P(X = x)$, x ’in veriden tahmin edilen olasılığını gösterir.

Shmueli ve diğerleri (2005), COM-Poisson dağılımının esnekliğini ve kullanışlılığını göstermek için iki veri seti ele almıştır ve bu verilere göre COM-Poissonness grafiği oluşturulmuştur. Birinci veri seti, iyi bilinen bir markanın giyim eşyasının üç aylık satışını, ikinci veri seti bir Macarca sözlükteki kelime uzunluklarını içermektedir. Şekil 6.1’de iki veri seti için sırasıyla COM-Poissonness grafikleri verilmiştir.



Şekil 5.1. COM-Poissonness grafikleri

Şekiller incelendiğinde, her iki veri setinde de doğrusal bir ilişkinin geçerli olabileceği anlaşılmaktadır. Ancak veriler her iki durumda da Poisson modeli ile yeterince iyi açıklanamamaktadır. Çünkü Poisson çizgisinin eğimi satış verilerinde 45 dereceden çok daha küçük ($\nu < 1$ aşırı yayılım), kelime uzunluklarında ise 45 dereceden daha büyük ($\nu > 1$ az yayılım) olarak gözlemlenmektedir. Ardından, ağırlıklı en küçük kareler yöntemiyle verilere düz bir çizgi uyarlanmıştır. Satış verileri için tahminler $\hat{\lambda} = 0,97$ ve $\hat{\nu} = 0,124$, kelime uzunluğu verileri için $\hat{\lambda} = 7,68$ ve $\hat{\nu} = 2,14$ ’dür.

Sıfır sayıları içermeyen ardışık olasılık oranları, $\log(x)$ 'e göre çizilir. COM-Poissonness grafiğindeki noktaların düz bir çizgi üzerinde olması durumunda COM-Poisson dağılımının yeterli bir dağılım olduğu gösterilmiş olacaktır. Poisson dağılımı söz konusu olduğunda, çizginin kesişim noktası 0 olacaktır (Shmueli ve diğerleri, 2005). COM-Poissonness grafiği yalnızca bu çizginin kesişim noktasını belirlerken, Poissonness grafiği (Hoaglin, 1980) hem kesişimi hem de eğimi belirlemektedir. COM-Poissonness grafiği, ardışık olasılık oranlarına odaklanan Ord’un grafiği (Ord, 1967) ile daha çok benzerlik göstermektedir.

5.2.2. En çok olabilirlik tahmini

COM-Poisson dağılımının olabilirlik fonksiyonu Eş. 5.3 ile ifade edilmektedir. Bu olabilirlik fonksiyonunu maksimum yapan değerler en çok olabilirlik tahmin edicilerini verir. Ancak

olabilirlik fonksiyonunun maksimizasyonu yerine genellikle fonksiyonun logaritması maksimize edilir. COM-Poisson dağılımının log-olabilirlik fonksiyonu Eş. 5.4 ile ifade edilir. Bu fonksiyonun λ ve ν 'ye göre türevleri alınıp sıfıra eşitlendiğinde aşağıdaki denklemler elde edilmektedir.

$$\frac{\partial \ln L}{\partial \lambda} = \frac{\sum_{i=1}^n x_i}{\lambda} - \frac{n \frac{\partial Z(\lambda, \nu)}{\partial \lambda}}{Z(\lambda, \nu)} = 0$$

$$\frac{\partial \ln L}{\partial \nu} = -\sum_{i=1}^n \ln(x_i!) - \frac{n \frac{\partial Z(\lambda, \nu)}{\partial \nu}}{Z(\lambda, \nu)} = 0$$

Burada

$$\frac{\partial Z(\lambda, \nu)}{\partial \lambda} = \sum_{j=0}^{\infty} \frac{j \lambda^{j-1}}{(j!)^{\nu}} = \frac{Z(\lambda, \nu)}{\lambda} \sum_{j=0}^{\infty} j \cdot \frac{\lambda^j}{(j!)^{\nu} Z(\lambda, \nu)} = \frac{Z(\lambda, \nu) E(X)}{\lambda}$$

$$\frac{\partial Z(\lambda, \nu)}{\partial \nu} = -\sum_{j=0}^{\infty} \frac{\lambda^j \ln(j!)}{(j!)^{\nu}} = -Z(\lambda, \nu) \sum_{j=0}^{\infty} (j!) \cdot \frac{\lambda^j}{(j!)^{\nu} Z(\lambda, \nu)} = -Z(\lambda, \nu) (\ln(X!))$$

eşitlikleri yerine koyulup denklemler çözüldüğünde maksimum noktadaki λ ve ν değerleri aşağıdaki eşitlikleri karşılar.

$$E(X) = \lambda \bar{X}$$

$$E(\ln(X!)) = \overline{\ln(X!)}$$

Bu denklemler analitik olarak çözülemediğinden Newton-Raphson yöntemi gibi yinelemeli bir yöntem kullanılabilir. Her yinelemede, X ve $\log(X!)$ 'in beklenen değerleri, varyansları ve kovaryansları, önceki yinelemedeki λ ve ν tahminlerinin ifadeye eklenmesiyle yaklaşık olarak hesaplanır.

Newton-Raphson metodu ile maksimum olabilirlik tahmini ve algoritması

Newton-Raphson metodu, kök bulma ve optimizasyon problemlerinde etkili, yinelemeli bir yöntemdir. Özellikle, maksimum olabilirlik tahmininde kullanılan log-olabilirlik fonksiyonunun türevlerinin bulunup çözülmesinde çok kullanışlıdır. Maksimum olabilirlik tahmin yöntemi, parametreleri belirlemek için log-olabilirlik fonksiyonunu maksimize etmeyi gerektirir. Bu fonksiyonun türevlerini almak ve sıfıra eşitlemek, parametrelerin tahmin edilmesini sağlar. Ancak, bu türevlerin kapalı formda çözümleri genellikle bulunamaz. Bu aşamada Newton-Raphson yöntemi devreye girer. Newton-Raphson metodu için kullanılan formül;

$$\theta_{\text{yeni}} = \theta_{\text{eski}} - H^{-1} \nabla \ell(\theta)$$

şeklinde tanımlanır. Burada θ parametre vektörünü temsil eder. $\theta = \begin{pmatrix} \lambda \\ \nu \end{pmatrix}$ eşitliği ile gösterilir.

Formülün uygulanması ile yeni parametre değerleri, eski değerlerin bir fonksiyonu olarak iteratif olarak güncellenir. Bu yöntem özellikle log-olabilirlik fonksiyonunun türevlerinin hesaplanabildiği ve ikinci türevlerin matrisinin (Hessian) tersinin alınabildiği durumlarda hızlı bir yakınsama sağlar.

COM-Poisson dağılım parametrelerinin Newton-Raphson metoduna göre tahmin algoritmasının adımları aşağıda listelenmiştir.

1. λ ve ν parametreleri için uygun başlangıç değerleri belirlenir.
2. Log-olabilirlik fonksiyonu hesaplanır.
3. Log-olabilirlik fonksiyonunun birinci ve ikinci türevleri hesaplanır.
4. Hessian matris ve gradient vektör hesaplanır.

- Gradient Vektör $\nabla \ell(\lambda, \nu) = \left(\frac{\partial \ell}{\partial \lambda}, \frac{\partial \ell}{\partial \nu} \right)$ ile
- Hessian matris $\begin{bmatrix} \frac{\partial^2 \ell}{\partial \lambda^2} & \frac{\partial^2 \ell}{\partial \lambda \partial \nu} \\ \frac{\partial^2 \ell}{\partial \nu \partial \lambda} & \frac{\partial^2 \ell}{\partial \nu^2} \end{bmatrix}$ ile elde edilir

5. Newton-Raphson güncellemesi uygulanır.

$$\begin{pmatrix} \lambda^{(k+1)} \\ \nu^{(k+1)} \end{pmatrix} = \begin{pmatrix} \lambda^{(k)} \\ \nu^{(k)} \end{pmatrix} - H^{-1} \nabla \ell(\lambda^{(k)}, \nu^{(k)})$$

6. Yakınsama kontrolü yapılır. Eğer $\|\theta^{(k+1)} - \theta^{(k)}\| < \epsilon$ ise iterasyon durur. Değilse, adım 2 ye dönerek işleme devam edilir.

7. Yakınsama sağlandığında $\lambda^{(k+1)}$ ve $\nu^{(k+1)}$ değerleri, parametre değerleri olarak kabul edilir.



6. COM-POISSON DAĞILIMININ PARAMETRELENDİRİLMESİ

6.1. Model Formülasyonları

COM-Poisson regresyonu için çeşitli model formülasyonları önerilmiştir. Bu alandaki ilk çalışmalar, dağılış parametresini sabit tutarak, yalnızca ilgili konum parametresine (λ) odaklanmıştır. Ancak COM-Poisson regresyon analizi, yalnızca konum parametresine odaklanmakla sınırlı kalmayıp, hem gözlemlerin yoğunluk seviyesini hem de yayılımını modellemeye imkan tanıyacak şekilde genişletilebilir.

Sellers ve Shmueli (2010) orijinal CMP(λ, ν) parametrelendirmesi ile çalışmış ve bu parametreleri doğrudan bağımsız değişkenlerle ilişkilendirmişlerdir. Sabit dağılış parametresi varsayımı altında bağımlı ve bağımsız değişkenler arasındaki ilişkiyi tanımlamak için logaritmik bağlantı fonksiyonunu (log-link) $\ln(\lambda_i) = X_i\beta$ kullanmışlardır. Bu formülasyon aşağıdaki şekilde tanımlanır;

$$\log E(Y_i) = \log \mu_i = \eta_i = \beta' X_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip}, \quad i = 1, \dots, n$$

Burada $\text{var}(Y_i) = \sigma^2 E(Y_i)$ ve σ^2 bir dağılış parametresini belirtir.

Her ne kadar bu formülasyon katsayıların kolay yorumlanmasına imkan vermese de, lojistik ve Poisson regresyon için tanımlanan formülasyonları genelleştirmesi nedeniyle oldukça yararlı olduğu belirtilmiştir. Poisson regresyonunda, bu bağlantı fonksiyonu ortalama ve bağımsız değişkenleri tam olarak ilişkilendirmektedir. Lojistik regresyonda ise, başarı olasılığı $p = \lambda / (1 + \lambda)$ dir ve COM-Poisson bağlantı fonksiyonu $\text{logit}(p) = \ln(\lambda)$ 'ya indirgenir. COM-Poisson regresyon link fonksiyonu $\ln(\lambda)$ uygun tahminler ve çıkarımlar yapılmasını, modelin güvenilirliğini artırır ve sonuçların geçerliliğini destekler.

Bazı araştırmacılar COM-Poisson regresyonuna karşı çıkmışlardır çünkü araştırmacılar genelleştirilmiş doğrusal modelleme için belirlenen herhangi bir bağlantı fonksiyonunun, bağımlı değişken ile bağımsız değişkenler arasında doğrudan bir ilişki sunmadığını ve katsayıların yorumlanmasını zorlaştırdığını belirtirler. Ne λ ne de ν net bir merkezi

parametre sunar, ν 1'e yakinken λ ortalamaya yakındır, ancak bu önemli yayılım durumları için geçerli değildir. Özellikle verinin aşırı yayılım gösterdiği durumlarda Lord ve diğerleri (2008), küçük ν değerinin COM-Poisson modelinin parametre tahminini zorlaştırdığını öne sürmektedir. Lord ve diğerleri (2008) bunun yerine $\ln(\mu_*) = \ln(\lambda^{1/\nu}) = X\beta$ ile yaklaşık COM-Poisson (ACMP(μ_*, ν)) parametrelendirmesini kullanırlar ve ACMP regresyonunun, μ_* merkezleme parametresi ile ilişkilerini tanımlayarak katsayıların daha iyi yorumlanabildiğini savunurlar. Dağılım parametresinin (ν) sabit olduğu veya $\ln(\nu) = G\gamma$ bağlantı fonksiyonuna göre değiştiği varsayılabilir (Sellers, 2023: 158).

Diğer çalışmalar μ_* için daha geniş bir logaritmik doğrusal modeli ele almıştır, yani $\ln\left(\frac{\mu_{*i}}{h_i}\right) = x_i\beta + \phi_i$ olarak tanımlanan log-lineer modelin, değişen yayılıma izin veren bir yapıya sahip olması için $\ln(\nu_i) = -g_i\gamma$ şeklinde olması sağlanmaktadır.

ACMP formülasyonları bazı faydalı değerlendirmeler sunsa da, dikkatli kullanılmalıdır. Lord ve diğerleri (2008), ACMP yapısının ortalamaya veya varyansa bağlandığını, bu nedenle regresyonun aşırı ve ya az yayılmış verileri veya iki yayılımında bulunduğu durumu modelleyebileceğini belirtmişlerdir. Chaniyalidis, Evers, Neocleous ve Nobile (2017) de benzer şekilde, μ_* ve ν yeterince büyük olduğunda, ACMP regresyonu yoluyla elde edilen ortalama ve varyans yaklaşımlarının makul olduğunu vurgulamaktadırlar, ancak bu, ACMP regresyonunun az yayılmış veride her zaman uygun olmadığı anlamına gelir. $\mu_* = \lambda^{1/\nu}$ parametrelendirmesi, zaten yaklaşık bir kapalı forma sahip olan $E(X) = \frac{\partial \log Z(\lambda, \nu)}{\partial \log \lambda}$

beklenen değerinin, daha kesin bir gösterimini sunmaktadır, bu yalnızca veri seti $\nu \leq 1$ olduğunda veya $\lambda > 10^\nu$ olduğunda geçerlidir. Bu nedenle bu yaklaşım yalnızca kısıtlı bir alanda karşılanır. Francis, Geedipally, Guikema, Dhavala, Lord ve LaRocca (2012), aslında, ACMP genelleştirilmiş doğrusal model çıkarımlarının, veriler düşük bir ortalamaya sahip olduğunda ve az yayılım gösterdiğinde yanlışlık gösterdiğini belirtmektedir. Ayrıca ACMP parametrelendirmesi μ_* 'ı bir merkezleme parametresi olarak belirtse de, dağılım ortalaması değildir. Dolayısıyla ACMP regresyonu yapılırken, özellikle doğruluk ve katsayı yorumlama açısından bu hususların göz önünde bulundurulması gerekmektedir.

Ortalamanın daha iyi yaklaşımlarına dayanan alternatif parametrelendirmeler de benzer şekilde ilişkili regresyon modellerinin geliştirilmesiyle sonuçlanmıştır. MCMP1 regresyonu, Y 'nin (tasarım matrisi X 'e bağlı) $MCMP1(\mu, \nu)$ dağılımına sahip olduğu, bağlantı fonksiyonu $\ln(\mu) = X\beta$ olan ve sabit dağılış (ν) parametresine sahip olan modeli ele alır. Bu yapının arkasındaki temel motivasyon, MCMP1 parametrelendirmesinin ortogonal ortalama ve dağılış parametreleri (μ ve ν) sağlamasıdır, bu da regresyon analizinde kullanışlı bir özellik olarak öne çıkar (Sellers, 2023: 159).

Ortogonallık, modeldeki her bir parametrenin diğerlerinden bağımsız olarak yorumlanmasını sağlar. Örneğin μ parametresi değiştirilirken ν parametresi sabit kalır, bu da iki parametrenin ayrı ayrı yorumlanmasını kolaylaştırır.

Regresyon katsayıları $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_{p1})$ için en çok olabilirlik tahmin edicileri asimptotik olarak etkili ve $\hat{\nu}$ 'den asimptotik olarak bağımsız olacaktır (Sellers, 2023: 159). Diğer bir deyişle veri seri büyüdükçe β katsayılarının tahminleri, $\hat{\nu}$ 'lerden bağımsız hale gelir. Bu durum parametrelerin ortogonal olduğu modellerde gözlemlenir. Bu yeniden parametrelendirme sabit terimlerin modele kolayca eklenmesine olanak tanır ve ortalama $\exp(\text{sabit terim}(\text{offset})) \times \mu(X\beta)$ şeklinde güncellenir. Yani MCMP1 dağılımının yeni yapısı artık $MCMP1(\exp(\text{sabit terim}(\text{offset})) \times \mu(X\beta), \nu)$ olarak güncellenir ve bir logaritmik bağlantı fonksiyonu varsayıldığında artık $\ln(\mu) = X\beta + \text{sabit terim}(\text{offset})$ elde edilir (Sellers, 2023: 160). Bu yaklaşım, sabit terimlerin modelde yer almasını ve mevcut yapıya zarar vermeden modeldeki bağımsız değişkenlerin daha sağlıklı tahmin edilmesini sağlar.

MCMP2 regresyonu bağımlı değişken Y 'nin bir $MCMP2(\mu, \phi)$ dağılımına sahip olduğu varsayılarak MCMP1 fonksiyonuna benzer şekilde geliştirilir. Burada $\mu \doteq \lambda^{1/\nu} - \frac{\nu-1}{\nu}$ için $\ln(\mu) = X\beta$ ve $\phi = \ln(\nu)$ bir sabittir. MCMP1 gibi, MCMP2 regresyonu da ortaya çıkan parametrelerin ortogonallığı nedeniyle hızlı bir şekilde yakınsar ve olağan genelleştirilmiş lineer model yapısı altında doğal katsayı yorumu sağlar. MCMP2 regresyonu yoluyla yakınsama, MCMP1'e göre daha hızlıdır çünkü MCMP2 modelinde μ , MCMP1 regresyonunda yapıldığı gibi hesaplamaya dayalı olarak değil cebirsel olarak elde edilir, bu

da daha basit bir form sağlar. Yine de MCMP2 regresyonunun dikkatli kullanılması önerilir. MCMP2 modelinin ortalama içi sağladığı yaklaşım, ACMP regresyonu için sağlanandan daha kesin olsa da ortalama yaklaşımın geçerliliği için $E(X) \approx \mu \doteq \lambda^{1/\nu} - \frac{\nu-1}{2\nu}$ denklemindeki kısıtlamaların sağlanmasına bağlıdır.

Tüm model formülasyonları sabit bir dağılış parametresi varsayılararak tanıtılırken, veri yayılımı da değişkenlik gösterebilir ve böylece yayılım miktarı ölçümler boyunca ilgili açıklayıcı değişkenlerle ilişkilendirmek için uygun bir bağlantı fonksiyonu (örneğin $n(\nu_i) = G_i \gamma$, $\gamma = (\gamma_0, \gamma_1, \dots, \gamma_{p_2-1})$) aracılığıyla modellenebilir. Dağılımın değişkenliğini dikkate alan bu yaklaşım, Miaou ve Lord (2003) tarafından açıklanan NB regresyonuyla uyumludur. Ayrıca, hem ortalama hem de varyans üzerinde farklı yardımcı değişkenlerin (covariate) görece etkilerini anlamaya olanak tanır (Sellers, 2023: 160).

6.2. COM-Poisson Parametrelendirmesi

COM-Poisson dağılımının parametrelendirilmesi, sayma verilerini daha esnek bir şekilde modellemek, modelin doğruluğunu arttırmak ve hesaplamaların karmaşıklığını azaltmak için kullanılır. Parametrelendirme ile verilerin yayılım seviyeleri kontrol edilir, bu sayede aşırı yayılım ya da az yayılım gösteren veriler modellenebilmektedir. Her bir parametrelendirmenin avantaj ve dezavantajlarına göre, belirli veri yapılarına ve yayılım özelliklerine sahip durumlarda uygun dağılımı seçmek önemlidir.

Bazı araştırmacılar, COM-Poisson parametrelendirmesinin sınırlı olduğunu savunmaktadır çünkü ne λ ne de ν açık bir merkezi eğilim ölçüsü sağlar, yani COM-Poisson dağılımının ortalama değeri doğrudan bir parametre ile ifade edilmez bu da merkezi eğilim ölçüsünün belirlenmesini zorlaştırmaktadır. Örneğin, Barriga ve Louzada (2014), λ değerinin, ν değeri 1'e yakın olduğunda beklenen değere yakın olduğunu, ancak ν değeri 1'den önemli ölçüde farklı olduğunda ortalamanın λ değerinden önemli ölçüde farklılık gösterdiğini belirtmiştir. Huang (2017), beklenen değerin $\nu \leq 1$ veya $\lambda > 10^6$ olduğunda doğru olduğunu, ancak bu aralığın dışında kalan senaryolar için, özellikle az yayılım gösteren sayma verilerini analiz ederken ($\lambda > 10^6$ çok büyük olmadığı sürece) hatalı olabileceğini

belirtmiştir. Bu durum, orijinal COM-Poisson formülasyonuna dayanan regresyon modelini, yayılmış veriler için yorumlamayı ve kullanmayı zorlaştırmaktadır.

COM-Poisson dağılımının ortalama ve varyansı için yaklaşık değerler;

$$E(X) \approx \mu = \lambda^{1/\nu} - \frac{\nu-1}{2\nu}$$

$$V(X) \approx \frac{1}{\nu} \lambda^{1/\nu} = \frac{\mu}{\nu} + \frac{\nu-1}{2\nu^2}$$

şeklinde verilir. Bu yaklaşımlar normalleştirme sabiti için asimptotik yaklaşım yoluyla belirlenir ve $\nu \leq 1$ veya $\lambda > 10^\nu$ için geçerlidir, ayrıca bu yaklaşımlar COM-Poisson dağılımının parametrelendirilmesi için bir dayanak oluşturmaktadır.

6.2.1. ACMP parametrelendirmesi (Approximate COM-Poisson parametrization)

COM-Poisson dağılımının ortalamasını daha doğru modelleyebilmek için geliştirilen bir parametrizasyon yöntemidir.

Bu parametrizasyon, λ ve ν parametrelerinin doğrudan ortalamayı vermediği klasik COM-Poisson modelinin eksikliklerini gidermeyi amaçlar. ACMP parametrizasyonu dağılımın ortalamasını daha doğru bir şekilde modelleyerek λ parametresinin verilerdeki ortalama sayıyı temsil etmesine yardımcı olur.

Aşırı yayılım gösteren veriler için bu sorunu aşmak amacıyla, Guikema ve Coffelt (2008), $\mu_* = \lambda^{1/\nu}$ yeniden parametrizasyonunu, COM-Poisson dağılımının "merkezi" olan ortalama değer için bir yaklaşım olarak önermektedir; bu noktada $\mu_* = \lambda^{1/\nu} > 10$, COM-Poisson ortalama yaklaşımını $\left(\mu = \lambda^{1/\nu} - \frac{\nu-1}{2\nu} \right)$ kısaltır ve $\lambda > 10^\nu$ kısıtlamasının yerine geçer. Bu yeniden parametrelendirme (ACMP parametrelendirmesi), COM-Poisson olasılık kütle fonksiyonunu Eş. 6.1'de belirtildiği gibi yeniden yapılandırır:

$$P(X = x) = \frac{1}{Z_1(\mu_*, \nu)} \left(\frac{\mu_*^x}{x!} \right)^\nu, \quad x = 0, 1, 2, \dots, \quad (6.1)$$

Burada $Z_1(\mu_*, \nu) = \sum_{j=0}^{\infty} \left(\frac{\mu_*^j}{j!} \right)^\nu$ normalizasyon sabitinin değiştirilmiş halidir. Bu temsille, ardışık olasılıkların oranı şu şekli alır:

$$\frac{P(X = x-1)}{P(X = x)} = \left(\frac{x}{\mu_*} \right)^\nu$$

Ayrıca, X değişkeninin ortalama ve varyansı şu şekildedir:

$$E(X) = \frac{1}{\nu} \frac{\partial \ln Z_1(\mu_*, \nu)}{\partial \ln \mu_*} \approx \mu_* + \frac{1}{2\nu} - \frac{1}{2} \quad (6.2)$$

$$V(X) = \frac{1}{\nu^2} \frac{\partial^2 \ln Z_1(\mu_*, \nu)}{\partial (\ln \mu_*)^2} = \frac{1}{\nu} \frac{\partial E(Y)}{\partial \ln \mu_*} \approx \frac{\mu_*}{\nu} \quad (6.3)$$

Buradaki asimptotik yaklaşım, $\mu_* = \lambda^{1/\nu} > 10$ için doğrudur ve sadece belirli dağılım koşulları altında iyi sonuç verir.

Chanialidis, Evers, Neocleous ve Nobile (2014), μ_* değerinin ortalamaya çok yakın olduğunu doğrulamaktadır, ancak bu durumun yalnızca μ_* ve ν küçük olmadığında doğru olduğunu belirtirler. Guikema ve Coffelt (2008) benzer şekilde bu yaklaşımların $\nu > 1$ veya $\mu_* < 10$ için doğru olmayabileceğini vurgulamaktadır.

Bu dağılım $\nu = 0$ için tanımsızdır ve bu nedenle $\nu > 0$ ile sınırlıdır. Dolayısıyla bu parametrelendirme geometrik dağılım özel durumu içermemektedir.

ACMP parametrizasyonu, özellikle aşırı yayılım gösteren verilerde modelin ortalamasını daha doğru bir şekilde modellemeye olanak tanır. Dağılımın merkezi parametresi olarak λ 'yı daha doğrudan yorumlamayı kolaylaştırır. Bu sayede, model çıktıları gerçek ortalamaya

daha yakın olduğu için λ daha anlamlı hale gelir. Yayılımın fazla olduğu ya da çok düşük olduğu veriler için COM-Poisson modelini daha esnek hale getirir.

6.2.2. MCMP1 parametrelendirmesi (Mean-parametrized COM-Poisson)

Huang (2017), ACMP parametrisasyonunun ortalamayı tam olarak temsil edemediğini savunmaktadır. Çünkü $\mu_* = \lambda^{1/\nu}$ olarak tanımlanan parametrenin, sadece konum parametresi olarak işlev gördüğünü, gerçek ortalamayı yansıtmadığını, bu değer, ortalamaya eşit olmadığını ve herhangi bir ν için iyi bir yaklaşım sağlamadığını belirtmektedir. Bunun üzerine Huang (2017), ortalama-parametrelili bir COM-Poisson (MCMP1) dağılımını önermiş, bu parametrelendirme ile daha basit ve daha kolay yorumlanan modellerin kullanabileceğini belirtmiştir.

Ortalama-parametrelili COM-Poisson (MCMP1) dağılımı, COM-Poisson olasılık kütle fonksiyonunu Eş. 6.4'te belirtildiği gibi yeniden parametrelendirir:

$$P(X = x | \mu, \nu) = \frac{\lambda(\mu, \nu)^x}{(x!)^\nu Z(\lambda(\mu, \nu), \nu)}, \quad x = 0, 1, 2, \dots, \quad (6.4)$$

Burada $\mu = E(X)$ dağılımın ortalamasıdır, dolayısıyla $\nu = 1$ eş yayılımı, $\nu < 1$ aşırı yayılımı, $\nu > 1$ ise az yayılımı göstermek üzere $\lambda(\mu, \nu)$ ile denklem çözümü

$$\sum_{x=0}^{\infty} (x - \mu) \frac{\lambda^x}{(x!)^\nu} = 0 \text{ olur.}$$

COM-Poisson dağılımının tüm temel özellikleri MCMP1 dağılımı için de geçerlidir. MCMP1 dağılımı yalnızca Poisson dağılımını değil, diğer bilenen kesikli dağılımları da genelleştirir; MCMP1 dağılımı $\nu = 0$ olduğunda başarı olasılığı $p = 1/(\mu + 1) < 1$ olan geometrik dağılıma, $\nu = 1$ olduğunda $\text{Poisson}(\mu)$ dağılımına indirgenir ve ν sonsuza yaklaştıkça Bernoulli ($p = \mu$) dağılımına yakınsar. Bu durum da, bu dağılımların kullanıldığı birçok alanda MCMP1'in alternatif bir seçenek olarak değerlendirilmesini sağlar (Huang, 2017).

MCMP1 ayrıca iki parametrelili bir üstel aile yapısına sahiptir ve ν sabit olduğunda μ ile ilgili bir parametrelili üstel dağılım ailesi oluşturur. Bu yapıda, ortalama ve dağılım parametreleri μ ve ν ortogonaldır. Bu, regresyon analizinde yararlı olur çünkü ortak değişkenleri, ortalama ile ilişkilendiren tahmini regresyon katsayıları asimptotik olarak etkilidir ve dağılım tahmini $\hat{\nu}$ 'den bağımsızdır. MCMP1 parametrelendirmesi, iki parametrelili üstel aileye sahip olması sayesinde istatistiksel çıkarımlar için güçlü bir yapı sunar.

Ancak, MCMP1 modelinde ortalama için hala kapalı bir form yoktur, yani belirli bir formüle dayalı basit bir çözüm sunulamamaktadır. Bu durum, modelin genellikle ortalama üzerinden kurulan, basit ve kolay yorumlanabilir log-lineer modellerle doğrudan uyumlu çalışmasını zorlaştırır (Huang, 2017).

MCMP1 ortalama ve dağılım parametrelerinin ilgili en çok olabilirlik tahmin edicileri

$\hat{\mu} = \bar{X}$ ve $\hat{\nu}$ 'dir. Bu, $E_{\bar{X},\nu}(\ln(x!)) = \frac{1}{n} \sum_{i=1}^n \ln(x_i!) = \overline{\ln(X!)}$ fonksiyonunu çözerek elde edilir.

$\hat{\mu}$ için kapalı form, μ için doğrudan bir tahmin sunar, böylece EÇO'lar için genel hesaplama süresini azaltır.

MCMP1'in önsel dağılımı,

$$h_+(\mu, \nu) = \lambda(\mu, \nu)^{a-1} \exp(-\nu b) [Z(\lambda(\mu, \nu), \nu)]^c \kappa_+(a, b, c), \quad \mu > 0, \nu \geq 0 \text{ formunda yazılır.}$$

Burada $\kappa_+^{-1}(a, b, c) = \int_0^\infty \int_0^\infty \lambda(\mu, \nu)^{a-1} e^{-b\nu} [Z(\lambda(\mu, \nu), \nu)]^{-c} d\mu d\nu$ normalleştirme sabitidir. Sonsal

form, önsel dağılım ile aynı formu korumaktadır.

Bayesci istatistiksel yöntemlerde kullanılan eşlenik önsel dağılıma sahip olması, MCMP1'i bu alanda popüler bir tercih haline getirir.

COM-Poisson dağılımını ortalama parametresi üzerinden yeniden yapılandırarak, COM-Poisson modelleri, diğer popüler ve rekabetçi sayma modelleriyle aynı seviyede anlaşılabilirlik ve yalınlığa ulaşmıştır. Bu yeni parametrelendirme sayesinde COM-Poisson,

yayılım esnekliğini koruyarak, aşırı veya az yayılımlı verilerin analizi için tercih edilen özelliklerini muhafaza etmektedir ve hesaplama süresi önemli ölçüde daha hızlı hale gelmektedir (Huang, 2017). Bu, COM-Poisson'ın sayma verilerinde yayılım derecesini bağımsızca ayarlayabilmesi ve geniş bir esneklik sağlaması ile birlikte, modelin kullanımı ve yorumlanabilirliği açısından önemli bir avantaj sunar. COM-Poisson'ın ortalama bazında parametrelendirilmesi, özellikle verilerin merkezi eğilimi hakkında daha doğrudan bir bakış açısı sağlar, bu da pratik uygulamalarda COM-Poisson modelinin diğer yaygın kullanılan modellerle daha kolay karşılaştırılabilir olmasını sağlar. Bu sayede, COM-Poisson'ın karmaşık yapısı ve esnekliği, daha net bir şekilde yorumlanabilir ve kullanımı kolay hale gelmiştir. Bu tezde MCMP1 parametrelendirilmesine dayalı Conway–Maxwell–Poisson regresyonu ele alınmıştır

COM–Poisson regresyonu, sayma regresyon modellerinde mevcut olan performans kısıtlamalarını ve ilişkili etkileri ortadan kaldıran, verilerin aşırı veya az yayılmasını ele alan alternatif bir modeldir.

COM-Poisson genelleştirilmiş lineer modeli, negatif binom regresyonu ile performans açısından karşılaştırılabilir bir şekilde, az yayılmış verileri daha iyi barındırabilen ve aşırı yayılmış verileri modelleyebilen esnek bir modeldir (Guikema ve Coffelt, 2008; Sellers ve Shmueli, 2010; Sellers, 2023: 158).

COM-Poisson regresyon modeli, bağımlı değişkenin COM-Poisson dağılımına uygun olduğu varsayımı altında çalışır. Modelin bağımsız değişkenlerle ilişkisi bir bağlantı fonksiyonu aracılığıyla belirlenir. COM-Poisson modeli, dağılış parametresi sayesinde hem az hem de aşırı yayılımı esnek bir şekilde modelleyebilir. Poisson ve binom dağılımlarını özel durumlar olarak içerir.

Tez çalışmasının simülasyon ve uygulama bölümlerinde R programlama ile yapılan analizlerde, MCMP1 parametrelendirmesi için geliştirilen mpcmp paketi kullanılmıştır. Ortalama parametrelili COM-Poisson (MCMP1) regresyon modelinde, kantil artıkların kullanımı ilk kez bu tez çalışması ile önerilmektedir. Kantil artık hesaplamaları için yazılan kodlar tamamen özgündür, hesaplamalar yazılan özel fonksiyonlarla gerçekleştirilmiştir. mpcmp paketi içeriğinde ya da herhangi bir R kütüphanesinde bulunmamaktadır. Kullanılan

mpcmp paketine ek olarak, tarafımızdan yazılan kantil artık fonksiyonları analiz ve simülasyon çalışmasında kullanılmıştır.

6.2.3. MCMP2 parametrelendirmesi (Alternative mean-parametrized COM-Poisson)

Ribeiro Jr. ve diğerleri (2019) ortalama yaklaşımı kullanmışlardır. Bu yaklaşımda modelin parametreleri iteratif bir çözüm süreciyle tahmin edilir. Önceki parametre (λ) tahminleri üzerinden geriye doğru giderek son parametreler belirlenir. Diğer bir deyişle, ilk başta bazı başlangıç değerleri tahmin edilir ve daha sonra bu değerler, modelin tüm parametreleri için en uygun sonucu bulana kadar güncellenir.

$\lambda = \left(\mu + \frac{\nu - 1}{2\nu} \right)^\nu$ ve $\vartheta = \ln(\nu)$ sırasıyla yoğunluk ve dağılış parametrelerini temsil eder;

$\vartheta = 0$ olduğunda model Poisson dağılımına yakınsar, $\vartheta < 0$ olduğunda aşırı yayılım, $\vartheta > 0$ olduğunda ise modelde az yayılım görülür. Buna göre, $\mu > 0$ için olasılık kütle fonksiyonu;

$$P(X = x | \mu, \vartheta) = \left(\mu + \frac{e^\vartheta - 1}{2e^\vartheta} \right)^{xe^\vartheta} \frac{(x!)^{-e^\vartheta}}{Z_2(\mu, \vartheta)}, \quad x = 0, 1, 2, \dots, \quad (6.5)$$

olan, μ ve ϑ parametrelili ikinci bir ortalama parametrik COM-Poisson (MCMP2) dağılımı elde edilir. Burada Z_2 ;

$$Z_2(\mu, \vartheta) = \sum_{j=1}^{\infty} \left(\mu + \frac{e^\vartheta - 1}{2e^\vartheta} \right)^{je^\vartheta} / (j!)^{e^\vartheta} \text{ şeklinde gösterilir.}$$

Ribeiro Jr. ve diğerleri (2019) de bu parametrelendirmenin μ ve ϑ arasında ortogonalite sağladığını belirtmektedirler; ancak, bunların parametrelendirmesi Huang (2017) modelinden daha basittir çünkü μ değeri "basit cebir" ile elde edilmektedir. Ayrıca, bu parametrelendirme, ACMP parametrelendirmesine kıyasla yaklaşık ortalamaya dayalı daha doğrudan bir parametrelendirme sunar. MCMP2 dağılımının ortalama ve varyansın asimptotik yaklaşımlarının, "parametre alanının büyük bir kısmı" için doğru olduğu gösterilmiştir; ancak ortalama yaklaşımı, sayma verileri küçük ve güçlü bir şekilde aşırı

yayılımlı olduğunda hatalı olabilir (Ribeiro Jr. ve diğerleri, 2019). Aşırı yayılım durumunda, MCMP2 modeli bu durumu yönetmek için, sınırlı sıfır yığılmalı (limited ZI) modeli kullanabilir, az yayılım durumunda ise sıfır yığılmalı (zero-deflation) modeli kullanabilir.

Bu parametrelendirmelerin tümü teorik olarak aynı olasılıkları üretmek amacıyla tasarlandığından teknik olarak geçerli olmakla birlikte, farklı parametrelendirmelerin istatistiksel metodolojide ve hesaplamada kullanılması, sonuçların yorumlanması ve çıkarımlarında farklılıklara yol açabilir. Francis ve diğerleri (2012), ACMP parametrelendirmesi için en çok olabilirlik tahmin doğruluğunu, tahmin yanlılığını ve yaklaşık asimptotik ortalamanın doğruluğunu değerlendiren çeşitli simülasyon çalışmaları yürütmüştür. En çok olabilirlik tahmin doğruluğu açısından, çoğu senaryoda güven aralığı, belirlenen parametreyi beklenen oranda içermektedir. Ancak bu durum dağılış parametresi için geçerli değildir. Bu doğrultuda, "dağılış parametresinin gerçek değeri, seçilen α -seviyesinin varsaydığı güven düzeyiyle belirlenmesi zor olabilir." Ayrıca, yapılan çalışmalar, gözlemlerin ortalamasının yüksek olduğu aşırı yayılımlı veri setlerinde, dağılış parametresinin olduğundan daha yüksek tahmin edildiğini göstermiştir. Tahmin edilen değerlerin yanlılığı ile ilgili olarak, veri az-yayılımlı veya eş-yayılımlı olduğunda, kesim parametresi önemli bir yanlılık göstermiştir; ancak yanlılık ortalamayla azalmış ve en büyük yanlılık, veri aşırı yayılımlı olduğunda meydana gelmiştir.

Shmueli ve diğerleri (2005) tarafından belirtilen kısıtlamalar, ACMP ortalama yaklaşımının $\mu_* > 10$ olduğu durumlar için uygun olduğunu öne sürmektedir. Ancak Francis ve diğerleri (2012), simülasyon çalışmalarının asimptotik ortalamanın $5 < \mu_* < 10$ aralığında bile gerçek ortalamaya yakın olduğunu, ancak $\mu_* < 5$ olduğunda yaklaşımın doğruluğunun azaldığını savunmaktadır. Ayrıca, asimptotik ortalama yaklaşımının, μ_* orta büyüklükte veya büyük olduğunda tüm yayılım türleri için doğru olduğu ve veri az yayılımlı olduğunda küçük μ_* için daha doğru olduğu gösterilmiştir. Ancak μ_* küçük olduğunda, yani dağılımın ortalamasının küçük ve verinin daha sıkışmış olduğu durumlarda, asimptotik doğruluk kaybolmaktadır. Francis ve diğerleri (2012), $\mu_* > 10$ olduğunda gerçek ortalama ile asimptotik yaklaşım arasındaki farkın çok küçük olduğunu, asimptotik doğruluğun oldukça yüksek olduğunu kabul etmektedir. Bununla birlikte, Ribeiro Jr. ve diğerleri (2019), Shmueli ve diğerlerinin (2005) bu kısıtlamalara dair iddialarına karşı çıkarak MCMP2

parametrelendirmesini savunmaktadırlar. Sayısal çalışmalar yoluyla, yazarlar, ortalama yaklaşımının doğru olduğunu, ancak varyans yaklaşımının en büyük hatalarının $\nu \leq 1$ olduğunda meydana geldiğini ve bildirilen kısıtlama bölgesi $\lambda > 10^\nu$ 'nın gerekli görünmediğini, çünkü yeniden parametrelendirilen parametrelerin varyansa etkisinde bir ilişki bulunmadığını belirtmektedirler. Bu değerlendirmeler, birkaç varsayıma dayanmaktadır. Yazarlar, gerçek ortalama ve varyansı, sonsuz toplamdan ilk 500 terimi toplayarak tahmin etmektedirler; ancak, sonsuz toplamın yakınsamasının uzun sürebileceğini kabul etmektedirler. Dolayısıyla, bu tür bir çalışma için $M = 500$ yeterli olup olmadığı belirsizdir. Bu yaklaşımların geçerliliği için belirlenen kısıtlama alanı, normalleştirme fonksiyonunu yaklaşık olarak hesaplama yeteneğinden kaynaklanmaktadır; ortalama ve varyans yaklaşımları bu sonucun yan ürünleridir. Bununla birlikte, Ribeiro Jr. ve diğerlerinin (2019) sonuçlarının, aşırı yayılım kısıtlaması $\nu \leq 1$ 'e karşıt olarak elde edilmesi şaşırtıcıdır. Diğer çalışmalarda (örneğin, Guikema ve Coffelt, 2008; Lord ve diğerleri, 2008), Ribeiro Jr. ve diğerlerinin (2019) sunduğu ortalama yaklaşımının daha kısıtlı bir temsilini sağlayan, ACMP yeniden parametrelendirmesinin iyi çalıştığı aşırı yayılım gösteren örnekler verilmiştir.

Bu arada, pozitif bir yaklaşık ortalamayı varsaymak, $\lambda^{1/\nu} > \frac{\nu-1}{2\nu}$ olduğu anlamına gelir, bu da küçük ve az yayımlı sayma verilerini içeren alanı yasaklar. Bununla birlikte, normalleştirme sabitinin yaklaşık olarak hesaplanabilmesi için, $\lambda^{1/\nu} > 10$ kısıtının olması gerekmektedir.

MCMP2 parametrelendirmesinin avantajları aşağıdaki şekilde sıralanmıştır (Ribeiro Jr. ve diğerleri, 2019):

Parametre Uzayının Basit Dönüşümü: Bu yeni parametrelendirme, orijinal parametre uzayını daha basit bir forma dönüştürür, bu da modellerin anlaşılmasını ve uygulanmasını kolaylaştırır.

Parametrelerin Ortogonal Olması: Bu parametrelendirme, parametrelerin birbiriyle bağımsız olmasını sağlar; yani, bir parametredeki değişiklik diğerini doğrudan etkilemez. Bu durum deneysel olarak gözlemlenmiştir ve modelde istatistiksel kararlılığı artırır.

Tam Parametrik Yaklaşım: Model tam parametrik bir yapıda olduğundan, teorik olarak daha sağlam bir çerçeve sunar.

Tahmin Ediciler Arasında Sıfıra Yakın Korelasyon: Deneysel olarak, tahmin ediciler arasında çok düşük (neredeyse sıfır) bir korelasyon bulunmuştur. Bu durum, modelin parametrelerini tahmin ederken daha az çakışma yaşanmasını sağlar.

Daha Hızlı Uyum Sağlama: COM-Poisson ortalama parametrelendirmesi, modeli veri setine uyarlarken daha hızlı bir performans gösterir. Bu, özellikle büyük veri setlerinde avantaj sağlar.

Katsayıların Doğrudan Yorumlanabilirliği: Bu parametrelendirme, GLM-Poisson modelinde olduğu gibi katsayıların doğrudan yorumlanabilmesine olanak tanır. Bu özellik, araştırmacıların modelin çıktısını daha kolay anlamasına ve yorumlamasına yardımcı olur.



7. ARTIKLARIN ANALİZİ

Artık, gözlem ile model tarafından tahmin edilen değer arasındaki farktır, aynı zamanda regresyon modeli tarafından açıklanamayan bağımlı değişkendeki değişkenliğin de bir ölçüsüdür. Artıkların analizi, çeşitli model yetersizliklerini keşfetmenin etkili bir yoludur. Artıkların grafiksel analizi, bir regresyon modelinin veriye ne kadar iyi uyduğunu kontrol etmek, yeterliliğini araştırmak ve temel varsayımları kontrol etmek için kullanılan güçlü bir araçtır (Montgomery, Peck, Vining, 2012; 232-243).

Normal doğrusal regresyon modellerinin incelenmesinde hem Pearson hem de sapma (deviance) artıkları sıklıkla kullanılır; bunlar, model verilere yeterince uyduğunda eşdeğer ve yaklaşık olarak standart normal dağılım gösterir. Bununla birlikte, bağımlı değişken kesikli olduğunda, artıkların dağılımı normallikten uzaklaşır ve farklı kesikli bağımlı değişken değerlerine göre neredeyse paralel eğriler oluşturulur. Bu durum artıkların görsel olarak incelenmesini önemli ölçüde zorlaştırır. Hesaplanan artık değerleri bağımlı değişkenin gerçek değerine karşı çizilebilmektedir. Ancak böyle bir çizim, sayma verileri için bilgilendirici değildir. Ayrıca, bağımlı değişken yalnızca birkaç değer alıyorsa bu grafiğin yorumlanması da zor olacaktır. Modelin belirlenmesi, uygun bir modelin veriler için yeterli olduğundan emin olmak için önemli bir adımdır. Doğrusal modeller için artık, gerçek ve uygun değerler arasındaki fark olarak kolayca tanımlanır. Doğrusal olmayan modeller için tek bir tanım yapmak mümkün değildir (Kılıç, 2023).

Artıklar, farklı bağlamlarda ve analizlerde kullanılan birden fazla türe sahiptir. Bunlardan bazıları pearson, sapma (deviance), ham ve kantil artıklardır. Bu bölümde Pearson, sapma ve kantil artıklar açıklanmıştır.

7.1. Pearson Artıklar

Heteroskedastisiteye yönelik en belirgin düzeltme Pearson artıklarında görülmektedir. Pearson artıkları, modelin tahmin edilen ve gözlenen değerler arasındaki farkı standartlaştırarak hesaplar.

$r_i^p = \frac{y_i - \hat{\mu}_i}{\sqrt{\text{Var}(\hat{y}_i)}}$ ile ifade edilir (Cameron ve Trivedi, 2013; 179). Pearson artıklar, gözlenen değer (y_i) ile tahmin edilen değer ($\hat{\mu}_i$) arasındaki farkın, standart hatasına oranı olarak tanımlanır. Pearson artıklar, modelin varyans yapısını göz önünde bulundurur.

7.2. Sapma (Deviance) Artıklar

Sapma artıkları, modelin tahmin ettiği değerler ile gözlenen değerler arasındaki olasılıksal uyumsuzluğu ölçer. Modelin maksimum olabilirlik tahmini çerçevesindeki performansını değerlendirir. Sapma artıkları daha karmaşık bir yapıya sahiptir ve genelde negatif logaritmik olabilirlik üzerinden hesaplanır. Özellikle Poisson ve diğer sayma veri modellerinde sıklıkla kullanılır.

Sapma artıkları, GLM için en uygun artık analizlerinden biridir. Sapma artıkları (deviance residuals), toplam sapma (summed deviance) olarak adlandırılan, tam model ile önerilen modelin olabilirlik fonksiyonları arasındaki farkın incelenmesine dayanmaktadır (Sarul, 2012).

y_i noktasındaki sapma artıkları,

$$r_i^D = \text{sign}(y_i - \hat{\mu}_i) \sqrt{d_i}$$

şeklinde tanımlanmaktadır. Burada $d_i, 2\{\log[p(y_i | \hat{\theta}_s)] - \log[p(y_i | \hat{\theta})]\}$ ifadesine eşittir.

$\theta = (\mu, \nu)$ modelin parametrelerini, $\hat{\theta}_s$ doymuş model için parametrelerin tahminini temsil eder.

7.3. Kantil (Quantile) Artıklar

Kantil artıklar, karmaşık regresyon modellerinde tanısal analiz yapmak için kullanılan bir yöntemdir. İlk olarak Dunn ve Smyth (1996) tarafından tanıtılmıştır. Kantil artıklar, model parametreleri tutarlı bir şekilde tahmin edildiği sürece, büyük örneklem durumunda asimptotik olarak standart normal dağılım gösterir (Pereira, 2017).

Kantil artıklar MCMP1 parametrelendirmesine dayalı COM-Poisson regresyon modeli için ilk olarak bu tezde $r_{q,i} = \Phi^{-1}\{F(y_i; \hat{\mu}_i, \hat{\nu})\}$ eşitliği ile tanımlanmıştır. Burada $\Phi()$, standart normal dağılımın kümülatif fonksiyonudur, $(\hat{\mu}_i, \hat{\nu})$ ise COM-Poisson regresyonundan tahmin edilen parametreleri ifade eder. Kantil artıklar, bilgisayar paketleri aracılığıyla kolayca hesaplanır ve normal olmayan dağılımlar, kesikli veriler veya karmaşık modeller gibi geleneksel yöntemlerin yetersiz kaldığı durumlarda güvenilir bir çözüm sunmaktadır (Dunn ve Smyth, 1996). Bu tezde MCMP1 parametrelendirmesine dayalı COM-Poisson regresyon modeli için kantil artıklar R programlama dilinde fonksiyonlar yazılarak hesaplanmıştır (EK-1).





8. SİMÜLASYON ÇALIŞMASI

Bu çalışmada COM-Poisson, Poisson ve negatif binom dağılımlarından ürettiğimiz dört farklı sayma verisi yapılarında, Pearson, sapma ve kantil artıklarının performansları karşılaştırılmıştır. Burada amaç, her modelden elde edilen Pearson, sapma ve kantil artıklarının normallik testine göre (Anderson-Darling testi) dağılımlarını değerlendirmek ve sayma verisi modellerinde kantil artıkların sayma regresyon modelleri için yanlış belirlenmiş dağılım varsayımlarını tespit etmede etkili olduğu göstermektir. Ayrıca, COM-Poisson modelinin uyum sağlama, çıkarım yapma ve esneklik açısından diğer istatistiksel modellerle karşılaştırılması yapılmıştır. Bu çalışma COM-Poisson'un, farklı veri dağılımlarına nasıl uyum sağladığını, modelin esnekliğini ve çeşitli dağılımlarla karşılaştırıldığında ne tür çıkarımlar sunduğunu incelemeyi hedefler. Bu karşılaştırmalar, modelin potansiyelini ve uygulama alanlarını daha iyi anlamamıza olanak tanır. Simülasyonda Huang (2017) tarafından geliştirilen ortalama parametrelili COM-Poisson (MCMP1) parametrelendirmesine dayalı regresyon modeli ele alınmıştır.

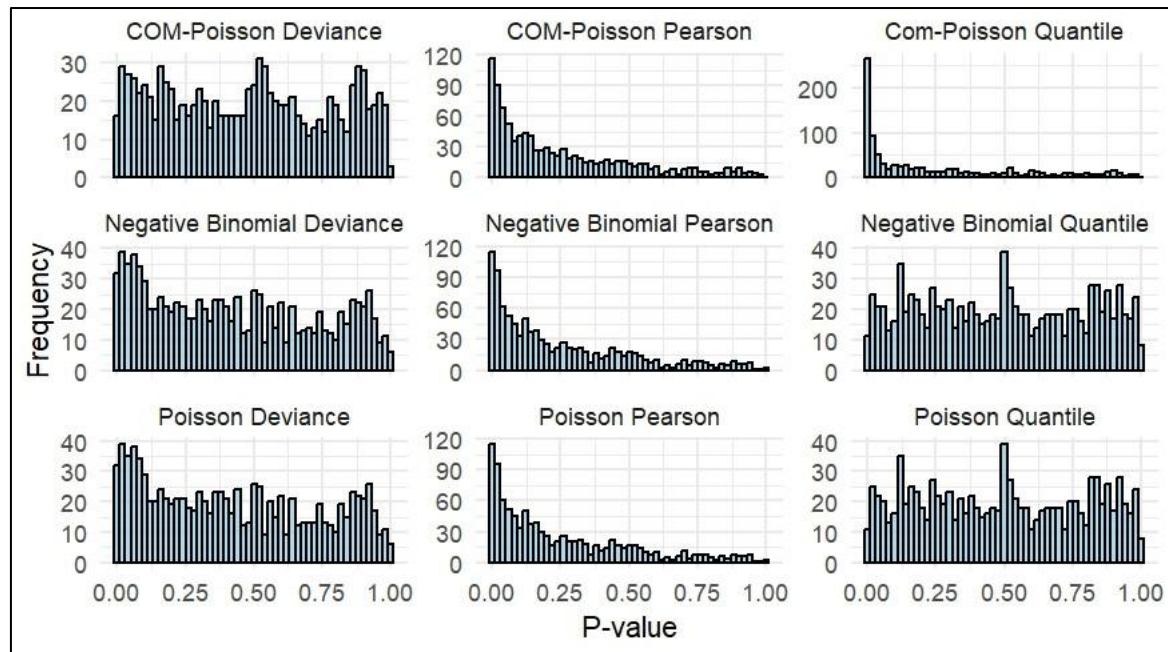
Bu inceleme için dört simülasyon döngüsü oluşturulmuştur. İlk döngüde az yayılım gösteren COM-Poisson dağılımından, ikinci döngüde aşırı yayılım gösteren COM-Poisson dağılımından veriler üretilmiş, üçüncü döngüde Poisson dağılımından, dördüncü döngüde ise negatif binom dağılımında elde edilen veriler üzerinden modeller değerlendirilmiştir. Her durum için 100 gözlem içeren veri setleri elde edilmiş olup, simülasyon 1000 tekrar üzerinden gerçekleştirilmiştir. Her durumda AD (Anderson-Darling) testinden elde edilen p değerlerine göre histogramlar oluşturulmuştur. Az yayılım durumunda COM-Poisson dağılımının dağılış parametresi 1.5 ve 2, aşırı yayılım durumunda ise 0.5 ve 0.8 olarak belirlenmiş ve her bir parametre değeri için ayrı ayrı histogram grafikleri oluşturulmuştur. Modellerin performansı yayılım koşullarına göre karşılaştırılmıştır. Simülasyon çalışmasıyla kantil artıkların sayma regresyon modellerini teşhis etmekteki performansı geleneksel artıklarla karşılaştırılmıştır.

Tüm bu analizler R programlama dilinde gerçekleştirilmiş olup, verilerin modellenmesi, analiz edilmesi ve görselleştirilmesi süreçleri R programlama dili ile yapılmıştır. COM-Poisson regresyon modeli için hesaplamalar gerçekleştirilirken mpcmp paketi kullanılmıştır (Fung, Alwan, Wishart ve Huang, 2019).

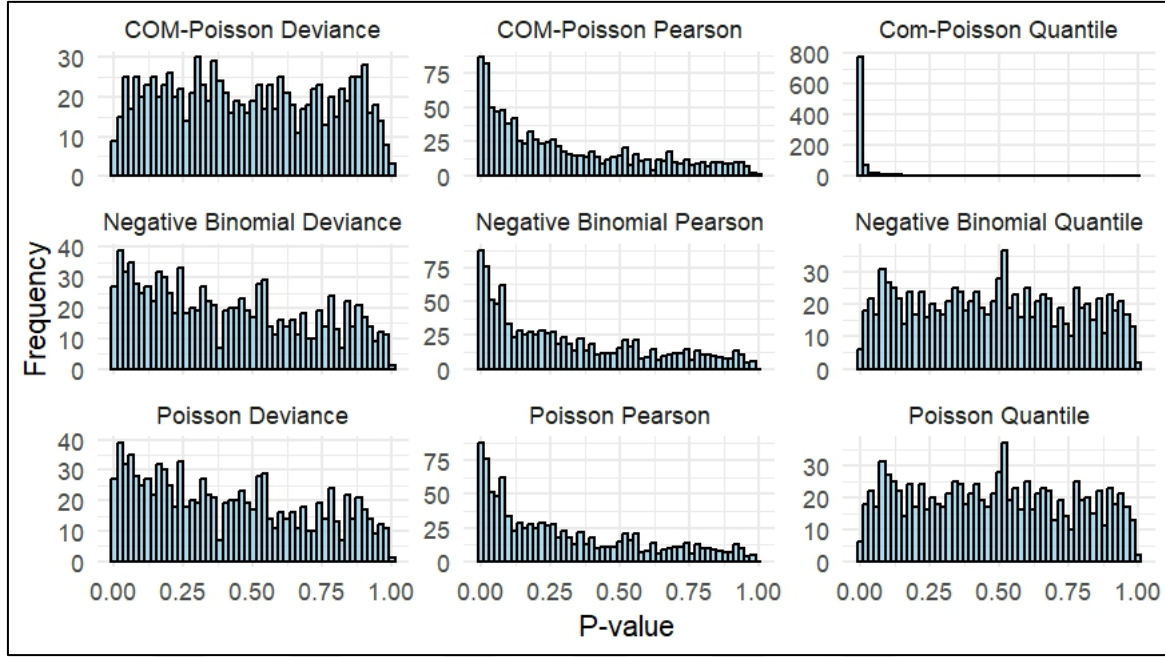
Simülasyon çalışmasında kullanılan kantil artık hesaplamalar için yazılan kodlar tamamen özgündür, hesaplamalar yazılan özel fonksiyonlarla gerçekleştirilmiştir. mpcmp paketi içeriğinde ya da herhangi bir R kütüphanesinde bulunmamaktadır. Kullanılan mpcmp paketine ek olarak, tarafımızdan yazılan pakete dahil olmayan kantil artık fonksiyonları simülasyon koduna eklenmiştir. Yazılan simülasyon kodunda dört farklı dağılım durumu için kullanılan parametreler Çizelge 8.1’de sunulmuştur. Kullanılan parametreler az yayılım ve aşırı yayılım durumlarına uygun olarak seçilmiş olup, az yayılım için $\nu = 1,5$ ve $\nu = 2$ değerleri, aşırı yayılım için $\nu = 0,5$ ve $\nu = 0,8$ değerleri kullanılmıştır.

Çizelge 8.1. Kullanılan parametreler

Durum	Parametreler
COM-Poisson (az yayılım)	$\mu = \exp(X\beta), n = 100, x_i \sim N(0,1), \beta = (1; 0,5), \nu = 1,5; 2$
COM-Poisson (aşırı yayılım)	$\mu = \exp(X\beta), n = 100, x_i \sim N(0,1), \beta = (1; 0,5), \nu = 0,5; 0,8$
Poisson	$\mu = \exp(X\beta), n = 100, x_i \sim N(0,1), \beta = (1, 0,5)$
Negatif Binom	$\mu = \exp(X\beta), n = 100, x_i \sim N(0,1), \beta = (1, 0,5), \theta = 1,5$

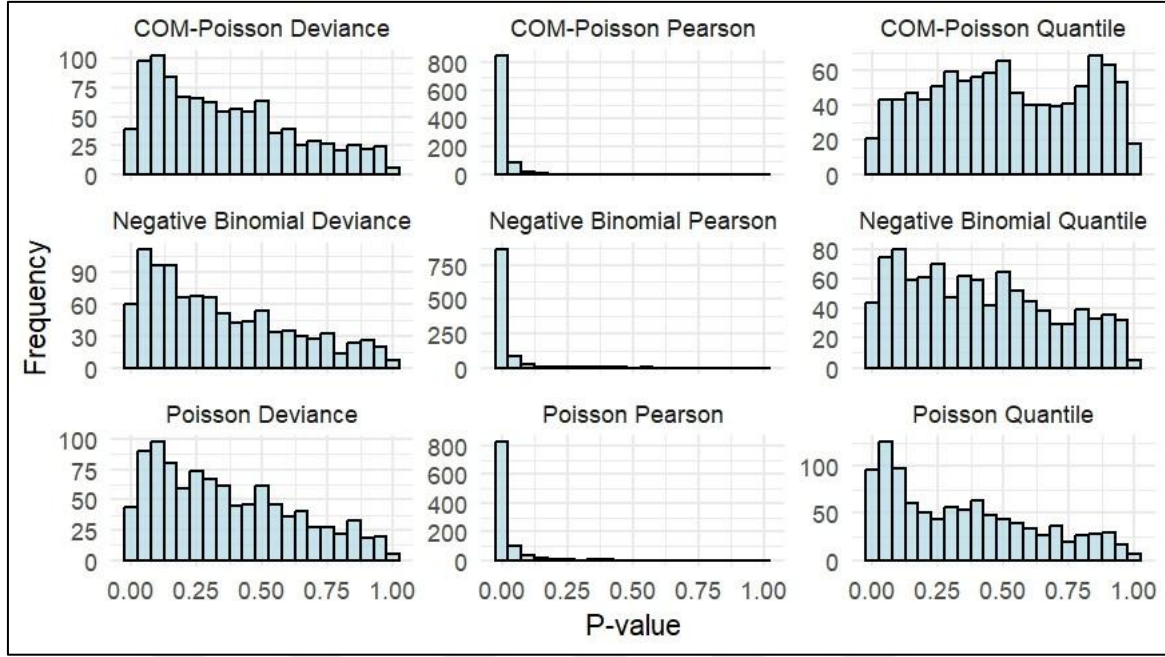


Şekil 8.1. Az yayılım ($\nu = 1,5$) COM-Poisson dağılımı altında veri seti üretildiğinde sayma modelleri için artıkların Anderson-Darling testinin p-değerlerinin histogramı

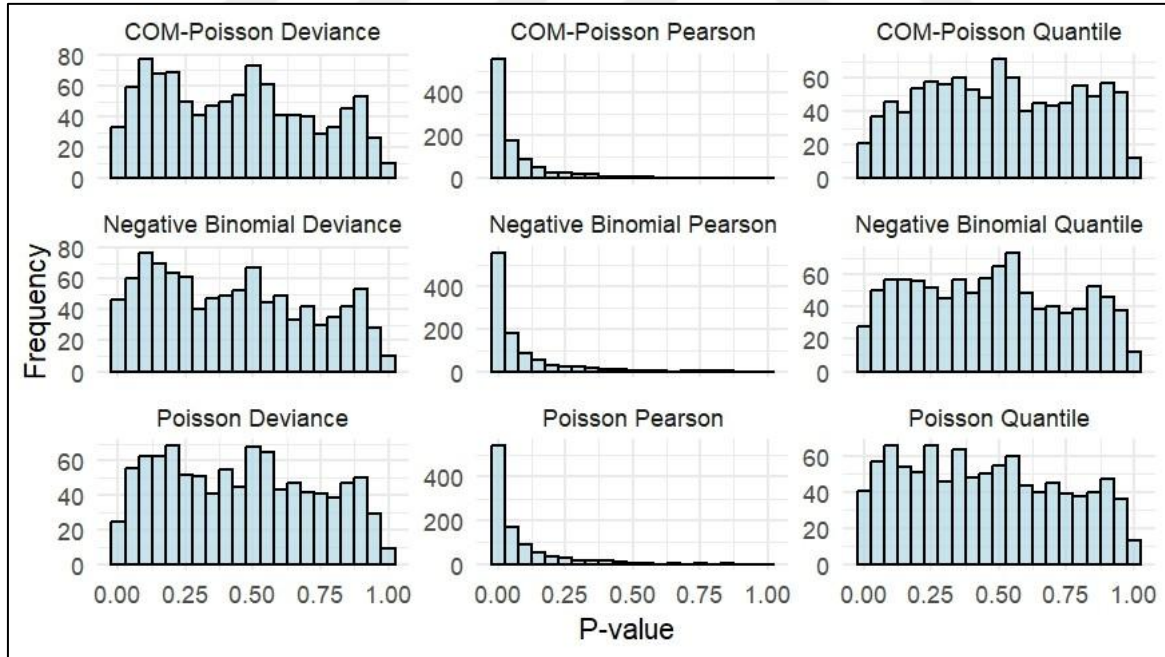


Şekil 8.2. Az yayılım ($\nu = 2$) COM-Poisson dağılımı altında veri seti üretildiğinde sayma modelleri için artıkların Anderson-Darling testinin p-değerlerinin histogramı

Şekil 8.1 ve Şekil 8.2’de her sütun bir artık türüne, her satır bir model türüne karşılık gelmektedir. COM-Poisson modelinden üretilen az yayılım gösteren ve örneklem büyüklüğü 100 birim olan veriler Poisson, negatif binom ve COM-Poisson modellerine uydurularak artıkların Anderson-Darling normallik testi p değerlerinin histogramları elde edilmiştir. Gerçek model altında p değerlerinin sapma artıklar için $[0,1]$ aralığında neredeyse tekdüze olduğu görülmektedir. Sapma artıklar normallik testine göre gerçek modelle en uyumlu sonucu vermektedir. Doğru modeli yanlış modellerden ayırt etmede etkili olduğunu göstermiştir. Pearson artıklarına ait p değerlerinin 0’a oldukça yakın olduğu, normallik varsayımını sağlamadığı görülmektedir. Bu da Pearson artıklarının doğru modeli yanlış modellerden ayırt etmede başarılı olamadığını göstermektedir. Kantil artıklar da doğru modeli ayırt etmede başarılı olamamıştır. Sapma artıkların normalliğe en yakın artıklar olduğu, COM-Poisson modelinin az yayılım koşullarında diğer modellere kıyasla daha esnek ve uyumlu olduğu sonucuna varılmaktadır.



Şekil 8.3. Aşırı yayılım ($\nu = 0,5$) COM-Poisson dağılımı altında veri seti üretildiğinde sayma modelleri için artıkların Anderson-Darling testinin p-değerlerinin histogramı

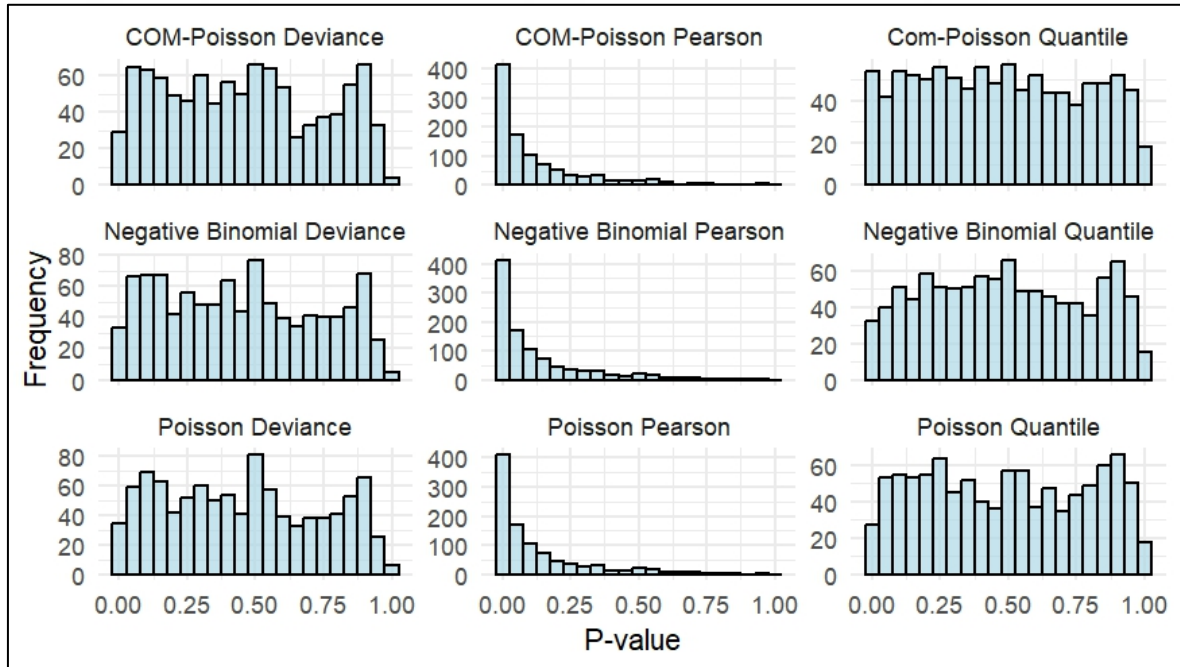


Şekil 8.4. Aşırı yayılım ($\nu = 0,8$) COM-Poisson dağılımı altında veri seti üretildiğinde sayma modelleri için artıkların Anderson-Darling testinin p-değerlerinin histogramı

Şekil 8.3 ve Şekil 8.4’de, COM-Poisson modelinden üretilen dağılış parametreleri sırasıyla ($\nu = 0,5$) ve ($\nu = 0,8$) olan, aşırı yayılım gösteren ve örneklem büyüklüğü 100 birim olan

verilerin Poisson, negatif binom ve COM-Poisson modellerine uydurularak elde edilen artıkların Anderson-Darling normallik testi p değerlerinin histogramları verilmektedir. Pearson artıklarına ait p değerlerinin 0'a oldukça yakın olduğu, normallik varsayımını sağlamadığı görülmektedir. Bu da Pearson artıklarının doğru modeli yanlış modellerden ayırt etmede başarılı olamadığını göstermektedir.

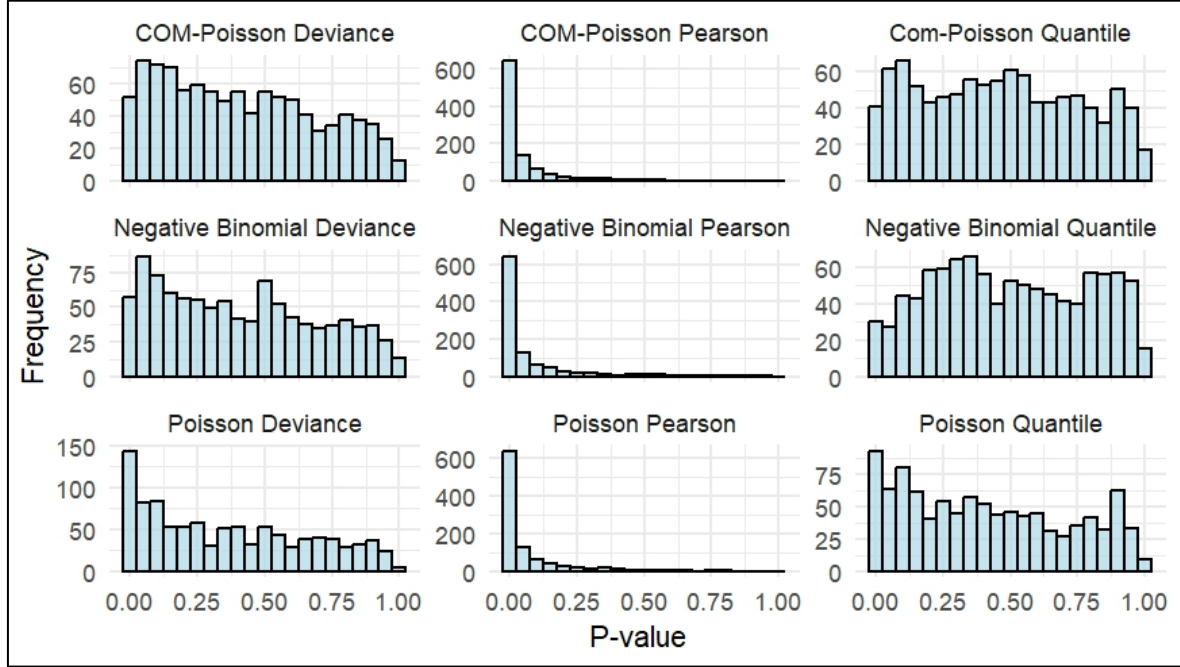
COM-Poisson dağılımı ve negatif binom dağılımının aşırı yayılım gösteren verileri modellemede başarılı olduğu, COM-Poisson modeline ait kantil artıkların doğru modeli yanlış modellerden ayırt etmede etkili olduğu görülmüştür.



Şekil 8.5. Poisson dağılımı altında veri seti üretildiğinde sayma modelleri için artıkların Anderson-Darling testinin p -değerlerinin histogramı

Şekil 8.5, Poisson modelinden üretilen ve örneklem büyüklüğü 100 olan veriler Poisson, negatif binom ve COM-Poisson modellerine uydurularak artıkların Anderson-Darling normallik testi p değerlerinin histogramları elde edilmiştir. Pearson artıklarına ait p değerlerinin 0'a oldukça yakın olduğu, normallik varsayımını sağlamadığı görülmektedir. Bu da Pearson artıklarının doğru modeli yanlış modellerden ayırt etmede başarılı olamadığını göstermektedir. Sapa artıklarına ait p değerleri net olarak tekdüze olmasa da Pearson artıklarına göre daha iyi sonuç vermiştir ancak doğru modeli yanlış modellerden ayırmada başarılı olamamıştır. COM-Poisson modeline ait kantil artıklarının neredeyse

tekdüze dağıldığı görülmektedir. Gerçek model, yanlış modellerden ayırt edilememiştir. Üç artık türü de doğru modeli tespit etmede yetersiz kalmıştır.



Şekil 8.6. Negatif Binom dağılımı altında veri seti üretildiğinde sayma modelleri için artıkların Anderson-Darling testinin p-değerlerinin histogramı

Şekil 8.6, Negatif Binom modelinden üretilen ve örneklem büyüklüğü 100 olan veriler Poisson, negatif binom ve COM-Poisson modellerine uydurularak artıkların Anderson-Darling normallik testi p değerlerinin histogramları elde edilmiştir.

Pearson artıklarına ait p değerlerinin 0'a oldukça yakın olduğu, normallik varsayımını sağlamadığı görülmektedir. Bu da Pearson artıklarının doğru modeli yanlış modellerden ayırt etmede başarılı olamadığını göstermektedir.

Sapma artıklarına ait p değerleri tekdüze olmasa da Pearson artıklarına göre daha iyi sonuç vermiştir ancak doğru modeli yanlış modellerden ayırt etmede başarılı olamamış, normallik varsayımını sağlayamamıştır. COM-Poisson ve negatif binom modeline ait kantil artıkların p değerlerinin benzer şekilde dağıldığı görülmektedir. Gerçek model, yanlış modellerden net olarak ayırt edilememiştir. Kantil artıklarla diğer artık türlerine kıyasla daha iyi sonuçlar verse de üç artık türü de doğru modeli tespit etmede yetersiz kalmıştır.

AIC, modelin hem uyumunu hem de karmaşıklığını değerlendirir. Daha düşük AIC değeri, modelin daha iyi bir performansa sahip olduğunu ve veri ile daha iyi uyum sağladığını göstermektedir.

Çizelge 8.2. dört farklı dağılımından (COM-Poisson az yayılım, COM-Poisson aşırı yayılım, Poisson, negatif binom) üretilen verilerin COM-Poisson, Poisson ve negatif binom modellerine uydurulup, her model için hesaplanan AIC değerlerinin ortalamasının alınmasıyla oluşturulmuştur.

Çizelge 8.2. COM-Poisson, Poisson ve Negatif Binom modellerine ait AIC değerleri

Durum	Model	Yayılgı Parametresi	Akaike Bilgi Kriteri (AIC) Değeri
Az Yayılım	COM-Poisson	$\nu = 1,5$	348,2614
		$\nu = 2$	325,6690
	Poisson	$\nu = 1,5$	353,1743
		$\nu = 2$	340,5141
	Negatif Binom	$\nu = 1,5$	355,1757
		$\nu = 2$	342,5159
Aşırı Yayılım	COM-Poisson	$\nu = 0,5$	409,5093
		$\nu = 0,8$	390,5111
	Poisson	$\nu = 0,5$	420,0958
		$\nu = 0,8$	390,7774
	Negatif Binom	$\nu = 0,5$	410,3865
		$\nu = 0,8$	390,8686
Poisson Dağılımı	COM-Poisson		376,8351
	Poisson		375,8679
	Negatif Binom		377,5399
Negatif binom dağılımı	COM-Poisson		395,5041
	Poisson		407,4751
	Negatif Binom		395,4784

Çizelge 8.2’de az yayılım ve aşırı yayılım gösteren verileri modellemede veriyle en uygun modelin COM-Poisson regresyon modeli olduğu görülmektedir. Poisson dağılımından

retilen verileri modellemede 375,8679 AIC deęeri ile Poisson modelinin en iyi seęim olduęu aynı zamanda COM-Poisson regresyon modelinin de iyi bir alternatif olduęu grlmřtr. Negatif Binom daęılımından retilen verileri modellemede 395,4784 AIC deęeri ile negatif binom modelinin en iyi seęim olduęu aynı zamanda COM-Poisson regresyon modelinin de ok yakın bir alternatif olduęu grlmřtr.



9. UYGULAMA

Bu bölümde iki farklı veri seti kullanılarak, aşırı ve az yayılım gösteren dağılımları modellemek amacıyla üç farklı sayma regresyon modeli kullanılmıştır. COM-Poisson, Poisson ve negatif binom modellerinin veriye uygunluğu karşılaştırılmıştır. Analizler sırasında, modellerin doğruluğunu ve geçerliliğini incelemek için kantil, Pearson ve sapma artık analizleri yapılmış ve her bir modelin artık yapıları değerlendirilmiştir. Ayrıca her model için QQ grafikleri oluşturulmuş ve bu grafikler aracılığıyla modellerin normal dağılıma uyumu incelenmiştir. Modellerin performanslarının karşılaştırılmasında Akaike Bilgi Kriteri (AIC) kullanılmıştır.

Uygulamada kullanılan tüm analizler R programlama dilinde gerçekleştirilmiş olup, verilerin modellenmesi, analiz edilmesi ve görselleştirilmesi süreçleri R programı ile yapılmıştır (EK-1). Bu bölümde, uygulamanın her aşaması detaylı bir şekilde ele alınarak elde edilen bulgulara dayalı çıkarımlar sunulmuştur. Ayrıca uygulamada kullanılan iki veri seti için de MCMP1 parametrelendirmesine dayalı regresyon modelleri kullanılmıştır. Kantil artık hesaplamaları için yazılan kodlar tamamen özgündür, hesaplamalar yazılan özel fonksiyonlarla gerçekleştirilmiştir. mpcmp paketi içeriğinde ya da herhangi bir R kütüphanesinde bulunmamaktadır. Kullanılan mpcmp paketine ek olarak, tarafımızdan yazılan kantil artık fonksiyonları analize eklenmiştir.

9.1. Teklif Sayısı Veri Seti

Tez çalışmasının bu bölümünde, R programında AER paketi içerisinde yer alan teklif sayısı (takeover bids) veri seti kullanılarak, COM-Poisson ve Poisson modellerinden hangisinin veriye daha iyi uyum sağladığını belirlemek amaçlanmıştır (Kleiber ve Zeileis, 2008: 9). Ayrıca artık analizi ile uygun modelin belirlenmesi sağlanmıştır.

Teklif sayısı veri seti 1978-1985 yılları arasında ABD’de başarılı ihale tekliflerin hedefi olan 126 firmanın aldığı teklif sayısını ve bu teklifleri etkileyen çeşitli firma özelliklerini içeren bir veri setidir. Veri setinin orijinal kaynağı Jaggia ve Thosar (1993)’ın çalışmasına dayanmaktadır. Çalışma, 1978’den 1985’e kadar olan dönemde ihale tekliflerinin hedefi

olan ve aslında ilk tektiftten itibaren 52 hafta içinde 126 ABD firmasının alım teklif sayısının çeşitli faktörlerle nasıl ilişkilendiğini modellemeyi amaçlamıştır (Jaggia ve Thosar, 1993).

Teklif sayısını modellemek için COM-Poisson ve Poisson regresyon modelleri uygulanmıştır. Bağımlı değişken hedef firma tarafından alınan ilk tektiftten sonraki teklif sayısı, bağımsız değişkenler ise Çizelge 9.1’de listelendiği gibi olacak şekilde modeller kurulmuştur.

Çizelge 9.1. Modelde kullanılan bağımsız değişkenler ile ortalama ve standart sapmaları

Değişken	Tanım	Ortalama	Standart Sapma
NUMBIDS	Devralma tekliflerinin sayısı	1,738	1,432
LEGLREST	Şirketin davalar yoluyla kendini savunup savunmadığı (Eğer hukuki savunma dava yoluyla yapılmışsa 1)	0,429	0,497
REAREST	Şirketin herhangi bir varlık yeniden yapılandırmasına dahil olup olmadığı (Eğer varlık yapısında önerilen değişiklikler varsa 1)	0,183	0,388
FINREST	Şirketin mülkiyet değişikliğine gidip gitmediği (Eğer mülkiyet yapısında önerilen değişiklikler varsa 1)	0,103	0,305
WHTKNIGHT	Şirketin dostane üçüncü taraf teklifine davet gönderip göndermediği (Eğer yönetim dostane üçüncü taraf teklifine davet göndermişse 1)	0,595	0,493
BIDPREM	İlk teklif fiyatının, tektiftten 14 gün önceki hisse fiyatına oranı	1,347	0,189
INSTHOLD	Kurumlar tarafından tutulan hisse yüzdesi	0,252	0,186
SIZE	Şirketin toplam varlıklarının milyar dolar cinsinden büyüklüğü	1,219	3,097
SIZESQ	Şirket büyüklüğünün karesi	10,999	59,915
REGULATN	Adalet Bakanlığı tarafından herhangi bir müdahale olup olmadığı (Eğer federal düzenleyici tarafından müdahale varsa 1)	0,270	0,446

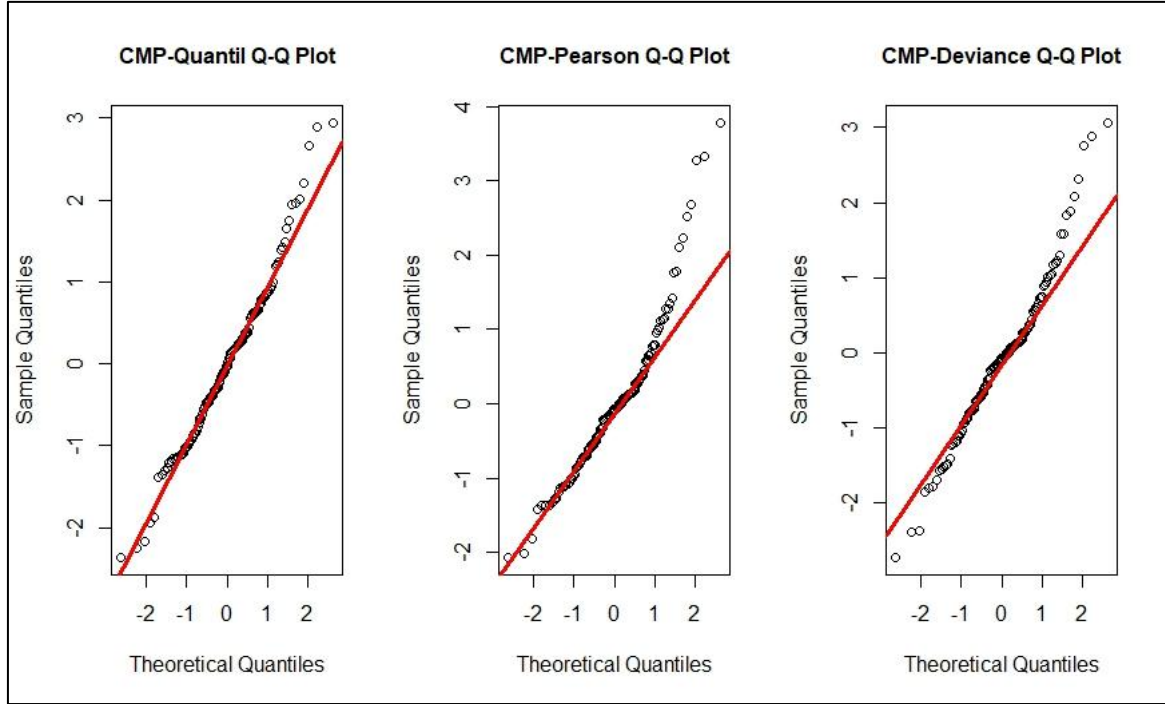
Teklif sayısı ile ilgili veriler Çizelge 9.2’de verilmiştir. Firmaların %10’undan azı sıfır teklif, firmaların yarısı tam olarak bir teklif (ilk tektiftten sonra), dörtte biri tam olarak 2 teklif, örneklemin geri kalanı 3 ila 10 teklif almıştır.

Çizelge 9.2. Teklif sayısı verileri

Teklif Sayısı	0	1	2	3	4	5	6	7	8	9	10
Frekans	9	63	31	12	6	1	2	1	0	0	1
Göreceli Frekans	0.071	0.500	0.246	0.095	0.048	0.008	0.016	0.008	0.001	0.00	0.008

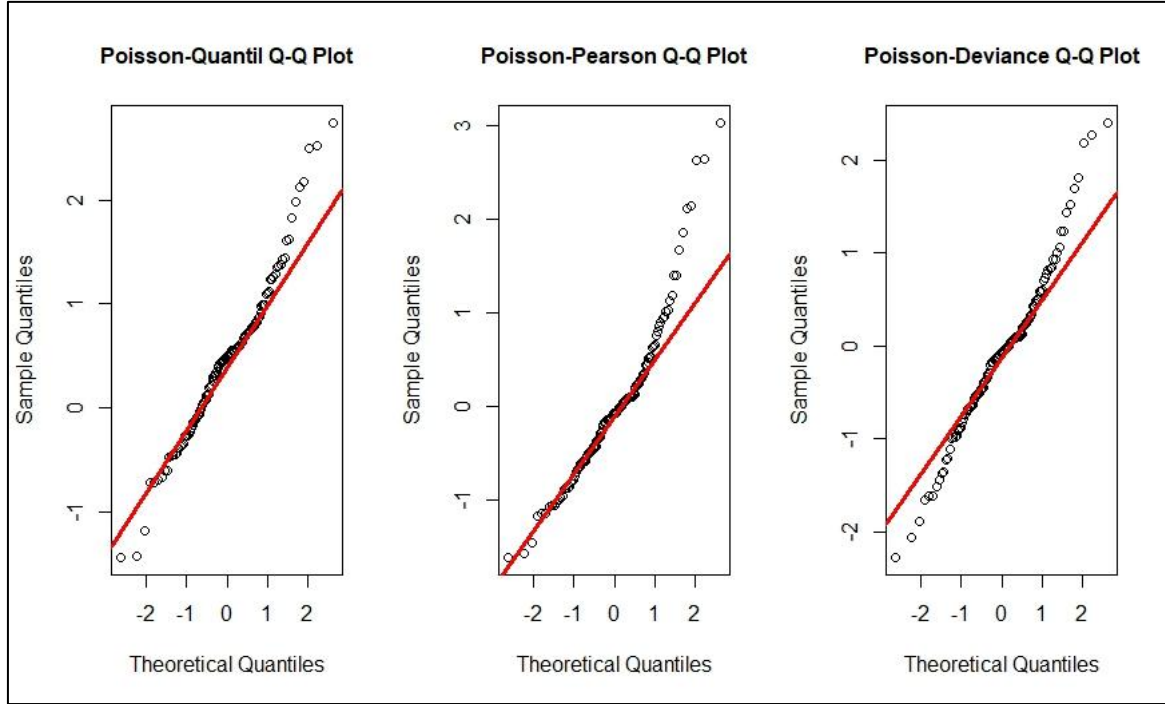
Teklif sayısı bağımlı değişkeninin yayılım durumu incelenmiştir. Poisson modeline dayalı yayılım testi yapılmıştır. Test sonucuna göre Z değeri = -3,0281, p değeri = 0,9988, tahmini yayılım katsayısı = 0,6824405 olarak elde edilmiştir. Bu durumda tahmini yayılım katsayısı 1'den küçük olduğu için (0,682), veri az yayılım göstermektedir (Kleiber ve Zeileis, 2019; Cameron ve Trivedi, 1990). Ayrıca, p -değerinin oldukça yüksek olması (0,9988), yayılımın 1'den büyük olma hipotezini açıkça reddetmekte ve yayılımın 1'den küçük olduğuna işaret etmektedir. Ayrıca ortalama parametrelili COM-Poisson dağılışı parametresi $\nu = 1,752$ olarak elde edilmiştir. Bu sonuç dağılışı parametresi $\nu = 1,752 > 1$ olduğu için modelin az yayılım gösterdiğini ifade etmektedir. Bağımlı değişkenin ortalaması 1,738 ve örneklem varyansı 2,050 olarak elde edilmiştir.

Bu doğrultuda, her bir model için uygunluk testleri yapılmış, kantil, Pearson ve sapma artıkları hesaplanmış ve bu artıklara ilişkin QQ grafikleri çizilmiştir. Artıklar üzerinden yapılan Anderson-Darling testi ile normal dağılımdan sapmalar incelenmiş ve her bir modelin uygunluğu değerlendirilmiştir. Model karşılaştırması için ise Akaike Bilgi Kriteri (AIC) kullanılarak en düşük AIC değerine sahip olan model seçilmiştir.



Şekil 9.1. COM-Poisson modeli artık değerlerinin QQ grafikleri

Şekil 9.1’de bulunan QQ grafikleri COM-Poisson modelinin artıklarının normalliğini değerlendirmek için oluşturulmuştur. Kantil artıklar, kırmızı çizgi üzerinde ve yakınında yer almakta, özellikle orta kısımda normalliğe iyi uyum göstermektedir. Uç değerlerde hafif sapmalar gözlenmektedir. Pearson ve sapma artıklar normallik çizgisi ile büyük ölçüde uyumsuzdur, uç değerlerde normallikten sapmalar oldukça belirgindir.



Şekil 9.2. Poisson modeli artık değerlerinin QQ grafikleri

Şekil 9.2 Poisson modelinin farklı türdeki artıklarının normalliğe uyumunu değerlendirmek için oluşturulmuştur. COM-Poisson ve Poisson modellerine ait artıkların QQ grafikleri incelendiğinde en iyi modelin COM-Poisson modeli olduğu, genel performans açısından en iyi sonucu sunduğu görülmektedir.

Çizelge 9.3. COM-Poisson ve Poisson modellerine ait AIC değerleri

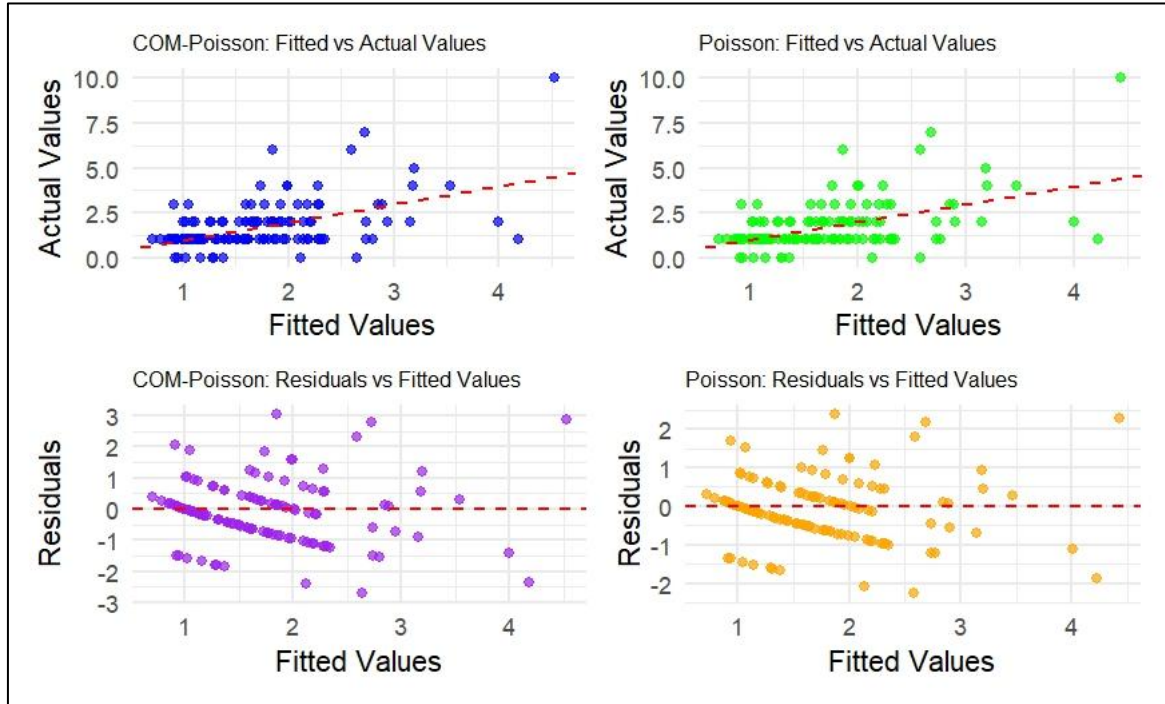
Model	Akaike Bilgi Kriteri (AIC) Değeri
COM-Poisson	382,1753
Poisson	389,8967

Çizelge 9.3’de açıkça görüldüğü üzere COM-Poisson modelinin veriye daha iyi uyum sağladığı ve Poisson modeline göre daha üstün olduğu gösterilmektedir.

Artıkların Anderson-Darling normallik testi sonucu Çizelge 9.4’de verilmiştir. Çizelgede COM-Poisson modeline ait kantil artıkların p değerinin, anlamlılık seviyesi ($p > 0,05$) üzerinde olduğu, modelin kantil artıklarının normallik varsayımına diğer artık türlerinden daha yakın olduğu gösterilmiştir.

Çizelge 9.4. Artıkların Anderson-Darling normallik testi p değerleri

Model	p - değeri
COM-Poisson Kantil	0,1666
COM-Poisson Pearson	< 0,001
COM-Poisson Sapma	0,0079
Poisson Kantil	0,01192
Poisson Pearson	< 0,001
Poisson Sapma	0,00965



Şekil 9.3. COM-Poisson ve Poisson modellerinin tahmin, gerçek değer ve sapma artık analizi karşılaştırması

Simülasyon çalışması bulgularına göre az yayılım COM-Poissonda durumunda sapma artıklarının diğer artıklara göre nispeten daha iyi sonuç verdiğinin tespit edilmesi nedeniyle Şekil 9.3’de COM-Poisson ve Poisson modellerine ait sapma artıklar analiz edilmiştir.

Hem COM-Poisson hem de Poisson modelleri için sapma artıklarının Anderson Darling normallik testi sonuçları (p - değerleri: 0,0079 ve 0,00965), sapma artıklarının normal dağılım göstermediğini ortaya koymaktadır. Bu, her iki modelin de tahmin edilen değerlere karşı sistematik hatalar içerdiğini ve artıkların normal dağılım varsayımını karşılamadığını göstermektedir. Şekil 9.3’de alt grafiklerde görüldüğü gibi, Poisson modelinin artıkları daha dar bir aralıkta (-2 ile +2) yoğunlaşırken, COM-Poisson modelinin artıkları daha geniş bir aralıkta (-3 ile +3) yayılmaktadır. Ancak her iki modelde de artıklar sistematik bir eğilim

göstermekte ve bu, test sonuçlarıyla desteklenmektedir. Hem COM-Poisson hem de Poisson modelleri artıkları sistematik hatalar ve rasgelelik ihlalleri göstermektedir.

Çizelge 9.5. COM-Poisson model için parametre tahmini

Değişkenler	Tahmin Değeri	Standart Hata	Z değeri	p - değeri	e^{β}
Sabit Terim	0,989630	0,435366	2,273	0,023020	2,69
Leglrest	0,267879	0,122873	2,180	0,029248	1,30
Rearest	-0,173177	0,154779	-1,119	0,263197	0,84
Fınrest	0,067744	0,174403	0,388	0,697693	1,07
Whtkngh	0,481281	0,131721	3,654	0,000258	1,61
Bidprem	-0,684822	0,307627	-2,226	0,026005	0,50
Insthold	-0,367886	0,346799	-1,061	0,288780	0,69
Size	0,179325	0,047627	3,765	0,000166	1,19
Sizesq	-0,007582	0,002485	-3,052	0,002276	0,99
Regulatn	-0,037569	0,130303	-0,288	0,773101	0,96

Modelden elde edilen regresyon katsayıları logaritmik ölçüde tahmin edilir, ancak katsayıları doğrudan yorumlamak yerine katsayıların üstel formu (e^{β}) kullanılır. $e^{\beta} = 1$ ise bağımsız değişkenin etkisi yoktur. $e^{\beta} < 1$ ise negatif bir etkisi vardır, $e^{\beta} > 1$ ise pozitif bir etkisi vardır.

Çizelge 9.5 incelendiğinde leglrest (şirketin davalar yoluyla kendini savunup savunmadığı), whtkngh (şirketin dostane üçüncü taraf teklifine davet gönderip göndermediği), bidprem (ilk teklif fiyatının, tekliften 14 gün önceki hisse fiyatına oranı), size (şirketin toplam varlıklarının milyar dolar cinsinden büyüklüğü), sizeq (şirket büyüklüğünün karesi) değişkenlerinin p değeri 0,05'ten küçük olduğu için %5 hata payı ile istatistiksel olarak anlamlı olduğu görülmüştür. Diğer değişkenler ise istatistiksel olarak anlamlı bulunamamıştır. Yani bu değişkenler, teklif sayısı üzerinde istatistiksel olarak anlamlı bir etki göstermektedir.

Dostane teklifler (whtkngh) değişkeni, teklif sayısında en büyük artışı sağlamaktadır, Teklif primi değişkeni ise teklif sayısında anlamlı bir azalmaya neden olmaktadır. Teklif primi arttıkça, teklif sayısı azalmaktadır.

Çizelge 9.6. COM-Poisson ve Poisson modellerinin bağımsız değişkenler üzerindeki etkileri

Değişken	COM-Poisson Katsayısı (Standart hata)	Poisson Katsayısı (Standart hata)
(Intercept)	0,990* (0,435)	0,986+ (0,534)
leglrest	0,268* (0,123)	0,260+ (0,151)
rearest	- 0,173 (0,155)	- 0,196 (0,193)
finrest	0,068 (0,174)	0,074 (0,217)
whtknght	0,481*** (0,132)	0,481** (0,159)
bidprem	- 0,685* (0,308)	- 0,678+ (0,377)
insthold	- 0,368 (0,347)	- 0,362 (0,424)
size	0,179*** (0,048)	0,179** (0,060)
sizesq	- 0,008** (0,002)	- 0,008* (0,003)
regulatn	- 0,038 (0,130)	- 0,029 (0,161)
+ p < 0,1, * p < 0,05, ** p < 0,01, *** p < 0,001		

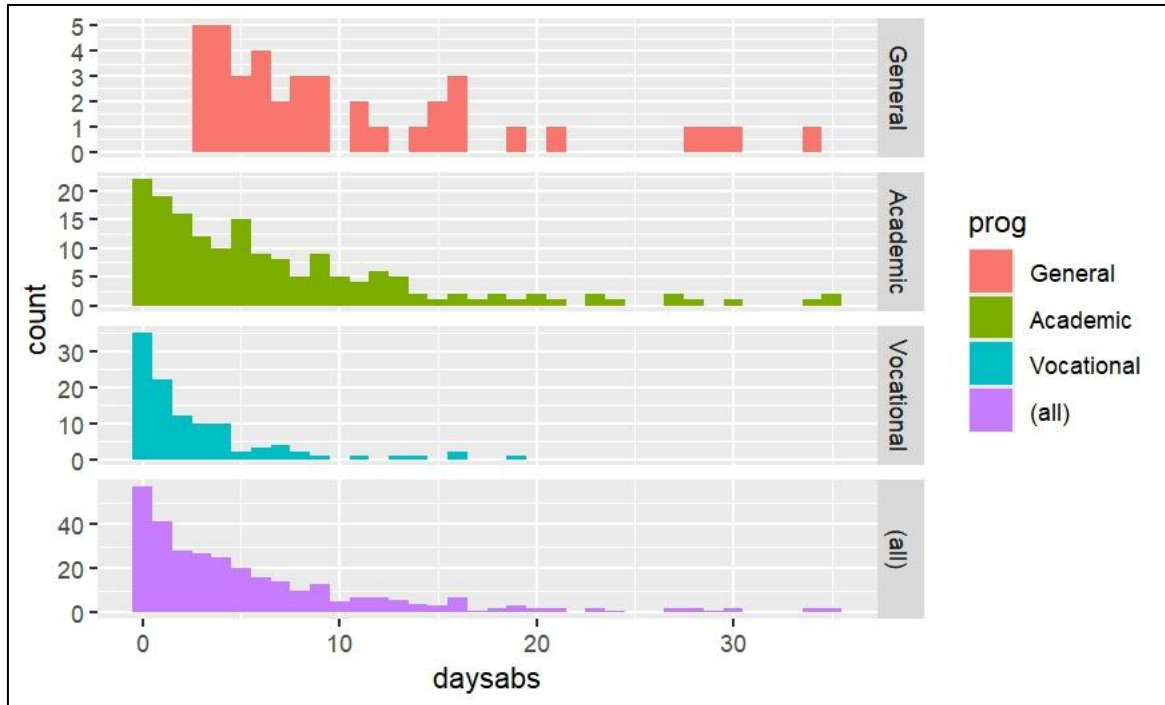
Çizelge 9.6.’da COM-Poisson ve Poisson modellerini karşılaştırdığımızda, COM-Poisson modelinin veriye daha iyi uyum sağladığı görülmektedir. Ayrıca, COM-Poisson modelinde tahmin edilen standart hataların daha küçük olduğu ve bunun sonucunda leglrest ve bidprem değişkenleri için p -değerlerinin Poisson modelindeki 0,1 yerine 0,05’in altına düştüğü anlaşılmaktadır.

9.2. Devamsızlık Veri Seti

Tez çalışmasının bu bölümünde, R programında mpcmp paketi içerisinde yer alan devamsızlık (attendance) veri seti kullanılarak, COM-Poisson, Poisson ve negatif binom modellerinden hangisinin veriye daha iyi uyum sağladığını belirlemek amaçlanmıştır. Ayrıca artık analizi ile uygun modelin belirlenmesi sağlanmış ve artıkların rasgeleliği analiz edilmiştir.

Bu veri seti iki liseden örnekleme yöntemiyle alınan 314 öğrencinin liseden devamsızlık gün sayısını, cinsiyetini, matematik puanını (100 üzerinden) ve program türünü (genel, akademik, mesleki) içermektedir (Huang, 2017).

Devamsızlık gün sayısını modellemek için COM-Poisson, Poisson ve negatif binom regresyon modelleri uygulanmıştır. Devamsızlık gün sayısı bağımlı değişken, cinsiyet, matematik puanı ve program türü bağımsız değişken olacak şekilde modeller kurulmuştur. Şekil 9.5, devamsızlık gün sayısının dağılımını, program türlerine göre gruplandırılmış olarak gösteren bir histogram sunmaktadır.



Şekil 9.4. Devamsızlık gün sayısı, akademik program türlerine göre gruplandırılmış olarak gösteren histogram

Devamsızlık gün sayısı bağımlı değişkeninin yayılım durumu incelenmiştir. Poisson modeline dayalı yayılım testi yapılmıştır. Test sonucuna göre $Z = 7,0546$, p - değeri = $8.653e-13$, tahmini yayılım katsayısı = $6,40697$ olarak elde edilmiştir (Fung ve diğerleri, 2019; Huang, 2017; Fung, Alwan, Wishart ve Huang, 2022). Bu durumda tahmini yayılım katsayısı 1'den büyük olduğu için, veri aşırı yayılım göstermektedir. Ayrıca ortalama parametrelili COM-Poisson dağılış parametresi $\nu = 0,02024$ olarak elde edilmiştir. Bu sonuç

dağılım parametresi $\nu = 0,02024 < 1$ olduğu için modelin aşırı yayılım gösterdiğini ifade etmektedir.

Bu doğrultuda, her bir model için uygunluk testleri yapılmış, simülasyon çalışmasında aşırı yayılım durumu için kantil artık değerleri daha iyi sonuçlar verdiği için bu veri setinde de kantil artıklar hesaplanmış ve bu artıklara ilişkin QQ grafikleri çizilmiştir. Artıklar üzerinden yapılan Anderson-Darling testi ile normal dağılımdan sapmalar incelenmiş ve her bir modelin uygunluğu değerlendirilmiştir. Model karşılaştırması için ise Akaike Bilgi Kriteri (AIC) kullanılarak en düşük AIC değerine sahip olan model seçilmiştir.

Çizelge 9.7. COM-Poisson, NB ve Poisson modellerine ait AIC değerleri

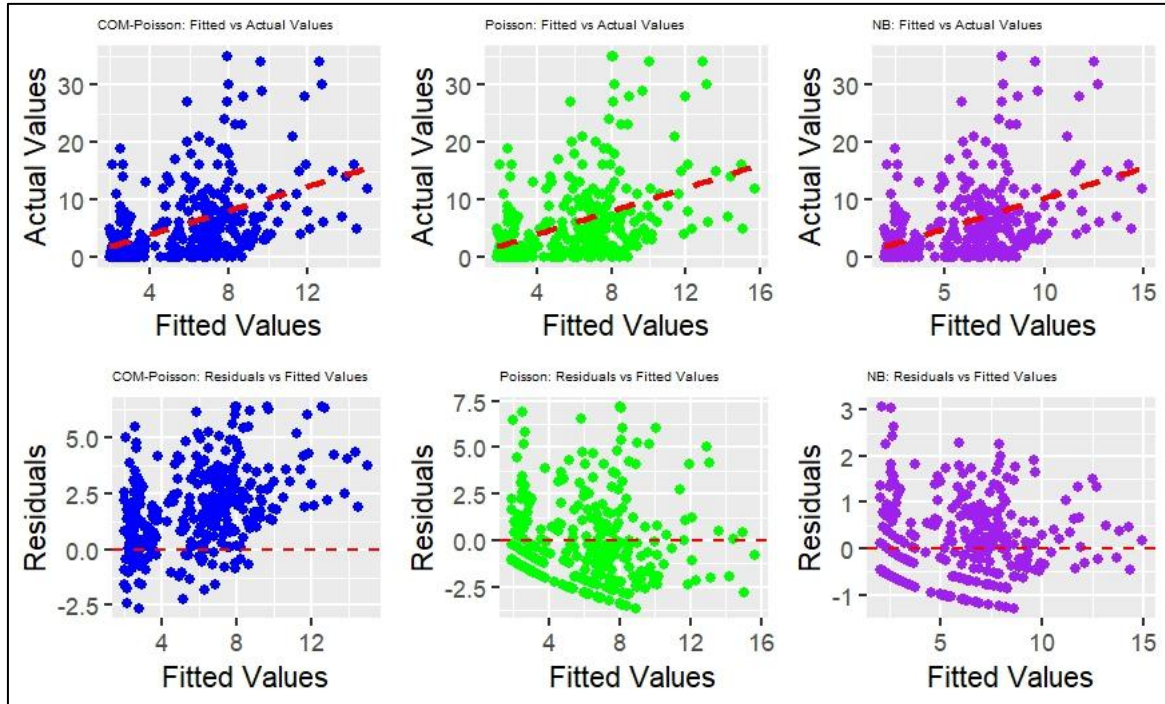
Model	Akaike Bilgi Kriteri (AIC) Değeri
COM-Poisson	1739,026
NB	1740,307
Poisson	2640,178

Çizelge 9.7’de açıkça görüldüğü üzere COM-Poisson modelinin veriye daha iyi uyum sağladığı ve Poisson modellere göre daha üstün olduğu gösterilmektedir.

Çizelge 9.8. COM-Poisson, NB ve Poisson modellerine ait kantil artıkların AD test p-değerleri

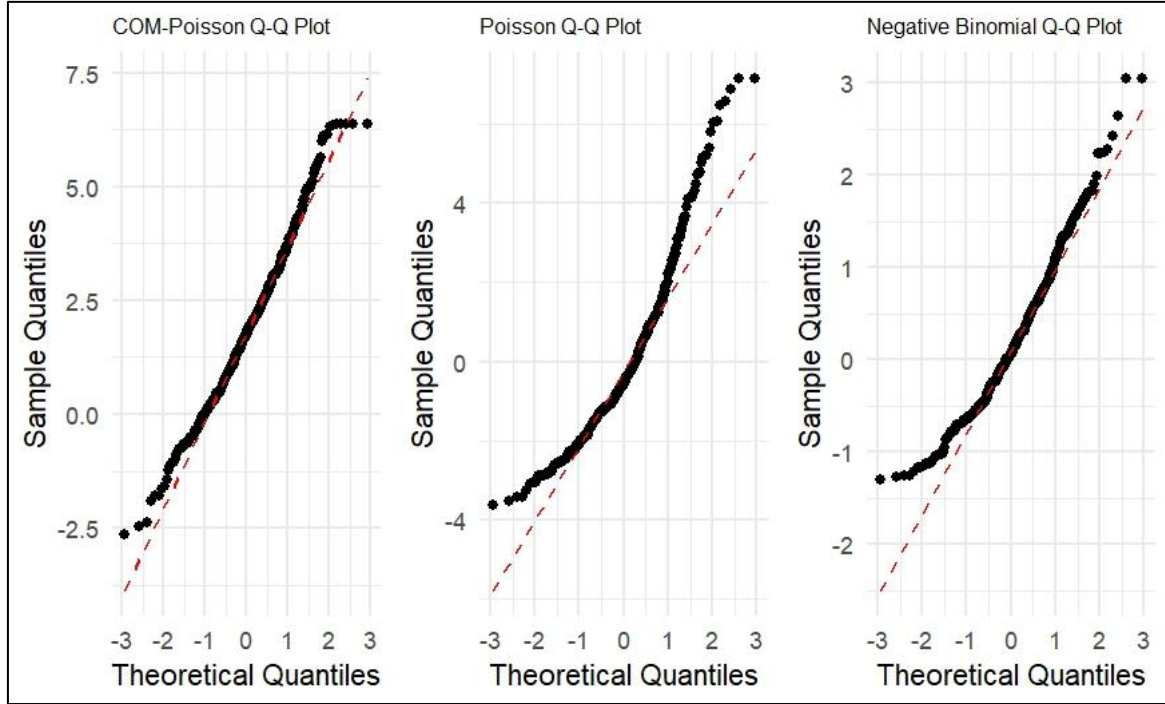
Model	AD Test p-değerleri
COM-Poisson	0,1047
NB	< 0,001
Poisson	< 0,001

Çizelge 9.8, COM-Poisson regresyon modeline ait kantil artıkların normal dağılıma uygunluk gösterdiğini, NB ve Poisson modeline ait kantil artıkların normal dağılıma uygunluk açısından başarısız olduğunu ifade etmektedir.



Şekil 9.5. COM-Poisson, Poisson ve NB modellerinin tahmin değerleri ve kantil artık analizi grafikleri

Şekil 9.5’da görüldüğü gibi, Poisson modelinin artıkları daha geniş bir aralıkta (-2,5 ile +7,5) yoğunlaşırken, COM-Poisson modelinin artıkları (-2,5 ile +5) Poisson modeline kıyasla daha dar bir aralıkta yayılmaktadır. NB modelinin artıkları ise (-1 ile 3) aralığında yayılmaktadır.



Şekil 9.6. COM-Poisson, Poisson ve NB modellerinin kantil artık değerlerinin QQ grafikleri

Şekil 9.6’da bulunan QQ grafikleri COM-Poisson, Poisson ve NB modellerinin kantil artık değerlerinin normalliğini değerlendirmek için oluşturulmuştur. COM-Poisson ve negatif binom modellerine ait kantil artıklar, kırmızı çizgi üzerinde ve yakınında yer almakta, özellikle orta kısımda normalliğe iyi uyum göstermektedir. Uç değerlerde hafif sapmalar gözlenmektedir. Poisson modeline ait kantil artıkların ise normal dağılıma uymadığı ve uç değerlerde ciddi sapmalar olduğu görülmektedir. Üç modele ait kantil artıklar, COM-Poisson ve NB modellerinin aşırı yayılım durumunu daha iyi ele alabildiğini göstermektedir.

Çizelge 9.9. COM-Poisson model için parametre tahmini

Değişkenler	Tahmin Değeri	Standart Hata	Z değeri	p - değeri	e^{β}
Sabit Terim	2,714645	0,190407	14,257	< 0,001	15,099
Cinsiyet	-0,214720	0,117148	-1,833	0,06682	0,806
Matematik puanı	-0,006323	0,002386	-2,650	0,00804	0,993
Akademik program türü	-0,425322	0,169524	-2,509	0,01211	0,653
Mesleki program türü	-1,253895	0,189478	-6,618	< 0,001	0,285

Çizelge 9.9. incelendiğinde matematik puanı, akademik program türü ve mesleki program türü değişkenlerinin p değeri 0,05'ten küçük olduğu için %5 hata payı ile istatistiksel olarak anlamlı olduğu görülmüştür. Cinsiyet değişkeni ise istatistiksel olarak anlamlı bulunmamıştır. Yani cinsiyet değişkeni devamsızlık gün sayısı üzerinde istatistiksel olarak anlamlı bir etki göstermemektedir.

Sabit terim, bağımsız değişkenlerin hepsi sıfır olduğunda devamsızlık gün sayısının temel seviyesini verir. Yani diğer etkiler yokken beklenen devamsızlık gün sayısı 15,10'dur. Matematik puanındaki her bir birimlik artış, devamsızlık gün sayısında %0,63 oranında bir azalma ile ilişkilidir. Program türü değişkeni için referans kategori genel program türü alınmıştır. Buna göre akademik program türünde olan öğrencilerin devamsızlık gün sayısı, genel program türünde olan öğrencilerin devamsızlık gün sayısından 0,653 kat daha azdır. Mesleki program türünde olan öğrencilerin devamsızlık gün sayısı ise referans kategoriye göre 0,285 kat daha azdır.



10. SONUÇ VE ÖNERİLER

Bu tez çalışmasında hem aşırı hem de az yayılım durumlarına uygun analiz yapma imkânı sağlayan Conway-Maxwell-Poisson regresyon modeli tanıtıldı ve simülasyon çalışmaları kullanılarak kantil artıkları, sapma ve Pearson artıkları karşılaştırıldı. Ayrıca, artıkların performansını ve sayma regresyon modellerinin veriye uygunluklarını belirlemek için gerçek veri setleri ile uygulama yapıldı.

Simülasyon çalışmasında COM-Poisson, Poisson ve negatif binom dağılımlarından ürettiğimiz dört farklı sayma verisi yapılarında, Pearson, sapma ve kantil artıklarının performansları karşılaştırıldı. Burada amaç, her modelden elde edilen Pearson, sapma ve kantil artıklarının normallik testine göre (Anderson-Darling testi) dağılımlarını değerlendirmek ve sayma verisi modellerinde kantil artıkların sayma regresyon modelleri için yanlış belirlenmiş dağılım varsayımlarını tespit etmede etkili olduğunu göstermektir. Ayrıca, COM-Poisson modelinin uyum sağlama, çıkarım yapma ve esneklik açısından diğer istatistiksel modellerle karşılaştırılması yapılmıştır.

Simülasyon çalışması, az yayılım durumlarında sapma artıkların, aşırı yayılım durumunda ise kantil artıkların daha iyi performans gösterdiğini ortaya koymuştur. Ayrıca, sayma verilerinde doğru modelin belirlenmesinde artık analizinin kullanılabilirliğini göstermiştir. Bununla birlikte bu bulgular sınırlı bir simülasyon çalışmasından elde edilmiştir.

Uygulama bölümünde yapılan ilk veri analizinde, R programında AER paketi içerisinde yer alan teklif sayısı veri seti kullanılarak, ihale teklif sayısına etki eden faktörlerin belirlenmesi amaçlanmış ve sayma veri modelleri kullanılmıştır. Bağımlı değişkenin yayılım durumu incelenmiş, Poisson ve COM-Poisson modellerine dayalı yayılım testleri yapılmış ve bağımlı değişkenin az yayılım gösterdiği belirlenmiştir (Kleiber ve Zeileis, 2019). Artık analizi yapılarak veriye uygun olan model belirlenmiştir. Uygun model derecesinin belirlenmesinde Akaike Bilgi Kriteri kullanılmış ve bu veri seti için en uygun modelin COM-Poisson modeli olduğu görülmüştür.

İkinci veri analizinde R programında mpcmp paketi içerisinde yer alan devamsızlık veri seti kullanılarak, öğrencilerin devamsızlık gün sayısına etki eden faktörlerin belirlenmesi

amaçlanmış ve sayma veri modelleri kullanılmıştır. Bağımlı değişkenin yayılım durumu incelenmiş, Poisson ve COM-Poisson modellerine dayalı yayılım testleri yapılmış ve bağımlı değişkenin aşırı yayılım gösterdiği belirlenmiştir. Artık analizi yapılarak veriye uygun olan model belirlenmiştir. Uygun model derecesinin belirlenmesinde akaike bilgi kriteri kullanılmış ve bu veri seti için en uygun modelin COM-Poisson modeli olduğu görülmüştür.

COM-Poisson modeli, hem az yayılım hem de aşırı yayılım koşullarında yüksek uyum performansı sergileyerek, alternatif modellere kıyasla daha esnek bir yapı sunmuştur. Özellikle kantil artıklar üzerinde yapılan normallik testlerinde, COM-Poisson modelinin kantil artıkları, normalliğe daha yakın p-değerleriyle başarılı bir performans göstermiştir. Bu durum, COM-Poisson modelinin farklı yayılım durumlarında güvenilir bir araç olduğunu göstermektedir. Simülasyon sonuçları, Poisson modelinin, sadece varsayımlarının karşılandığı durumlarda etkili olduğunu göstermiştir. Aşırı ve az yayılım koşullarında Poisson modelinden elde edilen artıklarda normallik varsayımı sağlanamamış ve p-değerleri düşük bulunmuştur. Bu da modelin yayılımın farklı olduğu durumlarda güvenilir bir analiz aracı olmadığını göstermektedir.

Aşırı yayılım durumlarında negatif binom modeli, COM-Poisson modeline benzer sonuçlar sunmuştur. Ancak, az yayılım durumlarında bu modelin etkinliği düşmüş ve normallik varsayımı sağlanamamıştır. Bu durum, negatif binom modelinin, az yayılımın söz konusu olduğu verilerde kullanılmaması gerektiğini göstermektedir. Sayma verileri için esnek bir çözüm sunan COM-Poisson modelinin avantajları, literatürdeki diğer modellerle kıyaslanarak detaylı bir şekilde incelenmiştir. Çalışmanın bulguları, özellikle yayılımın standart Poisson varsayımlarını ihlal ettiği durumlarda, COM-Poisson modelinin tercih edilmesi gerektiğini göstermektedir.

Bu sonuçlar, sosyal bilimlerden biyolojiye kadar geniş bir yelpazede sayma verisi analizi yapan araştırmacılar için rehber niteliğindedir. Gelecekteki araştırmalar, bu modelin diğer veri türlerine uygulanabilirliğini inceleyerek modelin genelleştirilebilirliğini daha da güçlendirebilir.

KAYNAKLAR

1. Hilbe, J.M. (2014). *Modeling Count Data*. New York: Cambridge University Press, 2-130.
2. Conway, R.W., Maxwell, W.L. (1962). A queuing model with state dependent service rates. *Journal of Industrial Engineering*, 12, 132-136.
3. Sellers, K.F., Borle, S. ve Shmueli, G. (2011). The Com-Poisson Model for count data: A survey of methods and applications. *Applied Stochastic Models in Business and Industry*, 28, 104-116.
4. Sellers, K.F., ve Shmueli, G. (2010). A Flexible regression model for count data. *The Annals of Applied Statistics*, 4 (2), 943-961.
5. Sellers, K.F. (2023). *The Conway-Maxwell-Poisson Distribution*. Cambridge: Cambridge University Press, 5-160.
6. Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., Boatwright, P. (2005). A useful distribution for fitting discrete data: revival of the Conway-Maxwell-Poisson distribution. *Applied Statistics*, 54, 127-142.
7. del Castillo J., Pérez-Casany M. (2005). Overdispersed and underdispersed Poisson generalizations. *Journal of Statistical Planning and Inference*, 134, 486–500.
8. Consul, P. C. (1989). *Generalized Poisson Distributions: Properties and Applications*. New York: Marcel Dekker Incorporated, 1-302.
9. Sundarapandian, V. (2009). *Probability, Statistics and Queueing Theory*. New Delhi: PHI Learning Private Limited, 686-709.
10. Tekin, B. (2011). *Sistemde Geçirilen Süre Açısından Kamu Hastanesi ve Özel Hastane Performanslarının Karşılaştırılması: Çankırı İli Örneği*, Yüksek Lisans Tezi, Çankırı Karatekin Üniversitesi Sosyal Bilimler Enstitüsü, Çankırı.
11. Ada, E. (1990). Bekleme Hattı Problemlerinin Analizinde Matematiksel Modeller. *Gazi Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 6, 1-54.
12. Kara, İ. (1976). *Servis Sistemleri ve Gelişler Zamana Bağlı Olduğunda Kapasite Sorununa Matematiksel Yaklaşım*. Eskişehir İktisadi ve Ticari İlimler Akademisi Yayınları. Yayın No:160/102.
13. Korucu, Ö. (2010). *Genelleştirilmiş Lineer Modellerde Model Seçimi Üzerine Alternatif Bir Yaklaşım*. Yüksek Lisans Tezi, İstanbul Üniversitesi Fen Bilimleri Fakültesi, İstanbul.
14. McCullagh, P., ve Nelder, J. A. (1989). *Genelleştirilmiş Doğrusal Modeller*. London: Chapman and Hall/CRC.

15. Myers, R. H., Montgomery, D. (1977). A Tutorial on Generalized Linear Models. *Journal of Quality Technology*, 29, 274-291
16. Hamada, M., Nelder, J. A. (1989). Generalized Linear Models for Quality-Improvement Experiments. *Journal of Quality Technology*, 21(3), 155-164.
17. Khuri, A. (2010). *Linear Models Methodology*. Chapman and Hall, USA, 473-509.
18. Agresti, A. (2002). *Categorical Data Analysis* (2nd Edition), New York: John Wiley and Sons, Inc., 117-149.
19. Dobson, A. J., Barnett, A. G. (2018). *An Introduction to Generalized Linear Models*. New York: CRC Press, 50-56.
20. Masmoudi, A. (2007). Steepness in Natural Exponential Families. *Turkish Journal of Mathematics*, 31(3), 319-331.
21. Aitkin, M., Anderson, D. A., Francis, B. J., and Hinde, J. P. (1989). *Statistical Modelling in GLIM*. Oxford: Clarendon Press.
22. Renshaw, A. E. (1994). Modelling the claims process in the presence of covariates. *ASTIN Bulletin*, 24(2), 265–285.
23. Frome, E. L. (1983). The analysis of rates using Poisson regression models. *Biometrics*, 39, 665-74.
24. Çelik, B. (2019). *Conway-Maxwell-Poisson Regresyon Modeli*. Yüksek Lisans Tezi, Giresun Üniversitesi Fen Bilimleri Enstitüsü, Giresun.
25. Ribeiro Jr., E. E., Zeviani, W. M., Bonat, W. H., Demetrio, C. G., and Hinde, J. (2019). Reparametrization of COM–Poisson regression models with applications in the analysis of experimental data. *Statistical Modelling*, 20(5), 443-466.
26. Lord, D., Guikema, S., Geedipally, S. R. (2008). Application of the Conway-MaxwellPoisson generalized linear model for analyzing motor vehicle crashes. *Accident Analysis and Prevention*, 40(3), 1123-1134.
27. Ridout, M.S., Besbeas, P. (2004). An empirical model for underdispersed count data. *Statistical Modelling*, 4, 77–89.
28. Hoaglin, D. C. (1980). A Poissonness plot. *The American Statistician*, 34, 146–149.
29. Ord, J. K. (1967). Graphical Methods for a Class of Discrete Disributions. *Journal of the Royal Statistical Society: Series A*, 130(2), 232-238.
30. Huang, A. (2017). Mean-parametrized Conway-Maxwell-Poisson regression models for dispersed counts. *Statistical Modelling*, 17(6), 359-380.

31. Madsen, H., Thyregod, P. (2011). *Introduction to General and Generalized Linear Models*. Boca Raton: CRC Press, 90-91.
32. Montgomery, D. C., Peck, E. A., Vining, G. G. (2012). *Introduction to Linear Regression Analysis*. Hoboken: John Wiley and Sons, 232-243.
33. Cameron, A. C., Trivedi, P. K. (2013). *Regression Analysis of Count Data*. New York: Cambridge University Press, 10-179.
34. Chaniavidis, C., Evers, L., Neocleous, T., and Nobile, A. (2014). Retrospective sampling in MCMC with an application to COM–Poisson regression. *Stat*, 3(1), 273–290.
35. Chaniavidis, C., Evers, L., Neocleous, T., and Nobile, A. (2017). Efficient Bayesian inference for COM–Poisson regression models. *Statistics and Computing*, 28(3), 595–608.
36. Francis, R. A., Geedipally, S. R., Guikema, S. D., Dhavala, S. S., Lord, D., and LaRocca, S. (2012). Characterizing the performance of the Conway–Maxwell–Poisson generalized linear model. *Risk Analysis*, 32(1), 167–183.
37. Miaou, S. P. and Lord, D. (2003). Modeling traffic crash-flow relationships for intersections: Dispersion parameter, functional form, and Bayes versus empirical Bayes. *Transportation Research Record*, 1840(1), 31–40.
38. Barriga, G. D. and Louzada, F. (2014). The zero-inflated Conway–Maxwell–Poisson distribution: Bayesian inference, regression modeling and influence diagnostic. *Statistical Methodology*, 21, 23–34.
39. Dunn, P. K., Smyth, G. K. (1996). Randomized Quantile Residuals. *Journal of Computational and Graphical Statistics*, 5(3), 236-44.
40. Kılıç, D. (2023). *Bell Tipi Regresyon Modelleri ve Artıkların İncelenmesi*. Doktora Tezi, Gazi Üniversitesi Fen Bilimleri Enstitüsü, Ankara.
41. Guikema, S. D. and Coffelt, J. P. (2008). A flexible count data regression model for risk analysis. *Risk Analysis*, 28(1), 213–223.
42. Sarul, L. S. (2012). *Aktüerya Alanında Genelleştirilmiş Lineer Modeller Kullanılarak Prim Fiyatlama ve Filo Verisi Üzerinde Bir Uygulama*. Doktora Tezi, İstanbul Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.
43. Pereira, G. H. A. (2017). On Quantile Residuals in Beta Regression. *Communications in Statistics – Simulation and Computation*, 48(1), 302-316.
44. Dunn, P. K., Smyth G. K. (1996) Randomized Quantile Residuals. *Journal of Computational and Graphical Statistics*, 5(3), 236-244.

45. Fung, T., Alwan, A., Wishart, J. and Huang, A. (2019). mpcmp: Mean-parametrized Conway-Maxwell Poisson Regression. R package version 0.2.0. URL: <https://CRAN.R-project.org/package=mpcmp>, Son Erişim Tarihi: 31.01.2025.
46. İnternet: Fung, T., Alwan, A., Wishart, J. and Huang, A. (2022). URL: <https://thomas-fung.github.io/mpcmp/articles/mpcmp.html>. Son Erişim Tarihi: 31.01.2025.
47. Jaggia, S., Thosar, S. (1993). Multiple Bids as a Consequence of Target Management Resistance. *Review of Quantitative Finance and Accounting*, 447-457.
48. Kleiber, C., Zeileis, A. (2008). *Applied Econometrics with R*. New York: Springer-Verlag, 9-10.
49. Kleiber, C., Zeileis, A. (2019). AER: Applied Econometrics with R. *R package version* 1.2-7. URL: <https://cran.r-project.org/web/packages/AER/index.html>. Son Erişim Tarihi: 31.01.2025.
50. Cameron, A.C., Trivedi, P.K. (1990). Regression-based Tests for Overdispersion in the Poisson Model. *Journal of Econometrics*, 46, 347–364.



EKLER

EK-1. Teklif sayısı veri seti R programlama kodu

```

library(COMPoissonReg)
library(MASS)
library(AER)
library(mpcmp)
library(ggplot2)
library(nortest)
library(dplyr)
rqres.cmp = function(y, lambda, nu)
{
  F = function(y) {
    pcomp(y, lambda, nu)
  }
  rqres(y, F)
}
rqres = function(y, F, eps = 1e-6)
{
  n = length(y)
  FL = F(y - eps)
  FU = F(y)
  u = pmin(1 - 1e-10, pmax(1e-10, runif(n, min = FL, max = FU)))
  qres = qnorm(u)
  return(qres)
}
residuals.cmp <- function(object, type = c("deviance", "pearson", "response"), ...) {
  type <- match.arg(type)
  y <- object$y
  mu <- object$fitted_values
  nu <- object$nu
  lambda <- object$lambda
  log.Z <- object$log_Z
  summax <- object$summax
  res <- switch(type,
    deviance = object$d_res,
    pearson = (y - mu) / sqrt(comp_variances(lambda, nu, log.Z, summax)),
    response = y - mu
  )
  return(res)
}
library(devtools)
devtools::install_github("thomas-fung/mpcmp")
data("takeoverbids")
ggplot(takeoverbids) +
  geom_bar(aes(x = numbids, y = ..count../sum(..count..))) +
  geom_pointrange(aes(x = numbids,
    y = dpois(numbids, mean(numbids)),
    ymin = 0,
    ymax = dpois(numbids, mean(numbids))),

```

EK-1. (devam) Teklif sayısı veri seti R programlama kodu

```

colour = "red") +
  labs(title = "Number of bids with Poisson fit", y = "Probability")
frequency <- table(takeoverbids$numbids)
relative_frequency <- prop.table(frequency)
freq_table <- data.frame(
  Count = as.numeric(names(frequency)),
  Frequency = as.numeric(frequency),
  Relative_Frequency = as.numeric(relative_frequency)
)
print(freq_table)

M.bids <- mpcmp::glm.cmp(numbids ~ leglrest + rearest + finrest + whtknight
  + bidprem + insthold + size + sizesq + regulatn, data=takeoverbids)
tidy(M.bids)
summary(M.bids)
M.bids.pois <- glm(numbids ~ leglrest + rearest + finrest + whtknight
  + bidprem + insthold + size + sizesq + regulatn, data=takeoverbids,
  family = poisson)
summary(M.bids.pois)
dispersiontest(M.bids.pois)
mnu<-M.bids $nu
mlambda<-M.bids$lambda
my=M.bids$y
quantile_res_cmp<-rqres.cmp(my, mlambda, mnu)
pearson_residuals_cmp_poisson <- residuals.cmp(M.bids,type="pearson")
deviance_residuals_cmp_poisson <- residuals.cmp(M.bids,type="deviance")
par(mfrow = c(1, 3))
qqnorm(quantile_res_cmp, main = "CMP-Quantil Q-Q Plot",cex.main = 1)
qqline(quantile_res_cmp, col = "red", lwd = 2)
qqnorm(pearson_residuals_cmp_poisson, main = "CMP-Pearson Q-Q Plot",cex.main = 1)
qqline(pearson_residuals_cmp_poisson, col = "red", lwd = 2)
qqnorm(deviance_residuals_cmp_poisson, main = "CMP-Deviance Q-Q Plot",cex.main =
1)
qqline(deviance_residuals_cmp_poisson, col = "red", lwd = 2)
par(mfrow = c(1, 1))
ad_test_quantile_cmp <- ad.test(quantile_res_cmp)
print(ad_test_quantile_cmp)
ad_test_pearson_cmp <- ad.test(pearson_residuals_cmp_poisson)
print(ad_test_pearson_cmp)
ad_test_deviance_cmp <- ad.test(deviance_residuals_cmp_poisson)
print(ad_test_deviance_cmp)
quantile_res_poi<-qnorm(ppois(M.bids.pois$y, lambda = fitted(M.bids.pois)))
pearson_residuals_poi <- residuals(M.bids.pois, type = "pearson")
deviance_residuals_poi<- residuals(M.bids.pois, type = "deviance")
par(mfrow = c(1, 3))
qqnorm(quantile_res_poi, main = "Poisson-Quantil Q-Q Plot",cex.main = 1)
qqline(quantile_res_poi, col = "red", lwd = 2)
qqnorm(pearson_residuals_poi, main = "Poisson-Pearson Q-Q Plot",cex.main = 1)

```

EK-1. (devam) Teklif sayısı veri seti R programlama kodu

```
qqline(pearson_residuals_poi, col = "red", lwd = 2)
qqnorm(deviance_residuals_poi, main = "Poisson-Deviance Q-Q Plot", cex.main = 1)
qqline(deviance_residuals_poi, col = "red", lwd = 2)
```

```
par(mfrow = c(1, 1))
```

```
ad_test_quantile_poi <- ad.test(quantile_res_poi)
print(ad_test_quantile_poi)
ad_test_pearson_poi <- ad.test(pearson_residuals_poi)
print(ad_test_pearson_poi)
ad_test_deviance_poi <- ad.test(deviance_residuals_poi)
print(ad_test_deviance_poi)
```

```
cmp_fitted <- fitted(M.bids)
cmp_residuals <- residuals(M.bids, type = "deviance")
```

```
ad.test(cmp_residuals)
```

```
pois_fitted <- M.bids.pois$fitted.values
pois_residuals <- residuals(M.bids.pois, type = "deviance")
ad.test(pois_residuals)
lawstat::runs.test(pois_residuals)
actual_values <- takeoverbids$numbids
cmp_data <- data.frame(
  Fitted = cmp_fitted,
  Residuals = cmp_residuals,
  Actual = actual_values
)
```

```
pois_data <- data.frame(
  Fitted = pois_fitted,
  Residuals = pois_residuals,
  Actual = actual_values
)
```

```
p1 <- ggplot(cmp_data, aes(x = Fitted, y = Actual)) +
  geom_point(color = "blue", alpha = 0.7) +
  geom_abline(intercept = 0, slope = 1, color = "red", linetype = "dashed") +
  labs(title = "COM-Poisson: Fitted vs Actual Values",
```

EK-1. (devam) Teklif sayısı veri seti R programlama kodu

```
  x = "Fitted Values",
  y = "Actual Values") +
  theme_minimal() + # Minimal tema ekleniyor
  theme(plot.title = element_text(size = 8))
```

```
p2 <- ggplot(pois_data, aes(x = Fitted, y = Actual)) +
  geom_point(color = "green", alpha = 0.7) +
```

EK-1. (devam) Teklif sayısı veri seti R programlama kodu

```
geom_abline(intercept = 0, slope = 1, color = "red", linetype = "dashed") +
  labs(title = "Poisson: Fitted vs Actual Values",
       x = "Fitted Values",
       y = "Actual Values") +
  theme_minimal() +
  theme(plot.title = element_text(size = 8)) # Ba
```

```
p3 <- ggplot(cmp_data, aes(x = Fitted, y = Residuals)) +
  geom_point(color = "purple", alpha = 0.7) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "COM-Poisson: Residuals vs Fitted Values",
       x = "Fitted Values",
       y = "Residuals") +
  theme_minimal()+
  theme(plot.title = element_text(size = 8))
```

```
p4 <- ggplot(pois_data, aes(x = Fitted, y = Residuals)) +
  geom_point(color = "orange", alpha = 0.7) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Poisson: Residuals vs Fitted Values",
       x = "Fitted Values",
       y = "Residuals") +
  theme_minimal()+
  theme(plot.title = element_text(size = 8))
```

```
library(patchwork)
(p1 | p2) / (p3 | p4)
```

```
qq_plot <- function(data, model_name) {
  ggplot(data, aes(sample = Residuals)) +
    stat_qq() +
    stat_qq_line(color = "red", linetype = "dashed") +
    labs(
      title = paste(model_name, "Q-Q Plot"),
      x = "Theoretical Quantiles",
      y = "Sample Quantiles"
    ) +
    theme_minimal()
}
qq_cmp <- qq_plot(cmp_data, "COM-Poisson")
qq_pois <- qq_plot(pois_data, "Poisson")
library(gridExtra)
grid.arrange(qq_cmp, qq_pois, ncol = 2)
```



Gazili olmak ayrıcalıktır