
GAZİ ÜNİVERSİTESİ
Gazi
1926

BELL TİPİ REGRESYON MODELLERİ VE ARTIKLARIN İNCELENMESİ

Duygu KILIÇ

DOKTORA TEZİ
İSTATİSTİK ANA BİLİM DALI

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

OCAK 2023

ETİK BEYAN

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

- 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 çalışmasında 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.

Duygu KILIÇ

12/01/2023

BELL TİPİ REGRESYON MODELLERİ VE ARTIKLARIN İNCELENMESİ

Doktora Tezi

Duygu KILIÇ

GAZİ ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

Ocak 2023

ÖZET

Bu tez çalışmasında veri setinde sayıma dayalı olarak elde edilen bağımlı değişken olması durumunda kullanılacak regresyon modellerinden bazıları ve bu modeller sonucunda oluşan artıklar incelenmiştir. Sayıma dayalı olarak elde edilen verilerde sıfır değerinde yığılma görülmesi durumuyla çoğunlukla karşılaşılır. Bu verilerle yapılan regresyon analizlerinde lineer regresyon varsayımları sağlanamayacaktır. Bu durum gözardı edilerek lineer regresyon ile analiz yapılırsa tahmin sonuçları hatalı veya yanlı olacaktır. Bu verileri modellemek için genellikle Poisson, negatif binom, sıfır yığılmalı Poisson ve sıfır yığılmalı negatif binom regresyon modelleri kullanılmaktadır. Yapılan tez kapsamında bu modellere alternatif olarak Bell ve sıfır yığılmalı Bell regresyon modelleri önerilmiştir. Önerilen iki regresyon modelinin bahsedilen diğer modellerle karşılaştırmasında hem uyum iyiliği kriterleri hem de Pearson, sapma ve rasgele kantil artıklara ait grafikler kullanılmıştır. Rasgele kantil artıklara ait grafiklerin çizdirilmesinde kullanılan R kodları herhangi bir paket içerisinde yer almamaktadır ve ilk kez bu tez çalışması için oluşturulmuştur. Pearson ve sapma artıkların kullanımının sayıma dayalı veriler için uygun olmadığı rasgele kantil artık türünün kullanımının uygun olduğu görülmüştür. 2018 yılı Türkiye Sağlık Araştırması'na ait veri seti kullanılarak ve simülasyon çalışmaları yapılarak bahsedilen regresyon modellerine ilişkin analizler sonucunda parametre tahminleri ve artıklara ait grafikler yorumlanmıştır. Çalışma sonucunda veride varyansın ortalamadan büyük olması durumu olarak ifade edilen aşırı yayılım ve sıfır yığılma olması durumda sıfır yığılmalı Bell regresyon modelinin kullanılan veriye uygunluk açısından diğer modellerden daha iyi olduğu kullanılan uyum iyiliği kriterleri ve rasgele kantil artık grafiklerinde görülmüştür. Bu tez çalışmasında rasgele kantil artıkların sayma regresyon modellerinde kullanımının önerilmesi ve klasik sayma modellerine alternatif olarak Bell/sıfır yığılmalı Bell regresyon modellerinin kullanılması istatistik literatürüne katkı sağlamıştır.

Bilim Kodu	:	20513
Anahtar Kelimeler	:	Sayma regresyon modelleri, Sıfır yığılmalı regresyon modellleri, Bell tipi regresyon modelleri, Artıkların incelenmesi
Sayfa Adedi	:	82
Danışman	:	Prof. Dr. Hülya BAYRAK

BELL TYPE REGRESSION MODELS AND ANALYSIS OF RESIDUALS

Ph. D. Thesis

Duygu KILIÇ

GAZİ UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

January 2023

ABSTRACT

In this thesis, some of the regression models to be used in the case of a dependent variable of the counting data type in the data set and the residuals formed as a result of these regression models were examined. It is often encountered in the case of zero value accumulation in counting type data. In the case of counting data, linear regression assumptions cannot be provided in the analyses. If this situation is ignored and analysis is made with linear regression, the estimation results will be erroneous or biased. Poisson, negative binomial, zero-stacked Poisson, and zero-stacked negative binomial regression models are generally used to model such data. In this study, Bell and zero-stack Bell regression models are proposed as alternatives to these models. In the comparison of these two proposed regression models with the other models mentioned, both the criteria of goodness of fit and the graphs of Pearson, deviation and random quantile residuals were used. It has been seen that the use of Pearson and deviation residuals from these three residue types used is not suitable for counting data, and the use of random quantile residue type gives the same result as the criteria used in model selection. By using the real data set of the 2018 Turkey Health Survey and by performing simulation studies, the analysis of the mentioned regression models was made and the graphs of the parameter estimates and the obtained residuals were interpreted. As a result of the study, it has been seen in the goodness of fit criteria and random quantile residual graphs that the zero-clustered Bell regression model is better than the other models in terms of compliance with the data used, in case of overdispersion, which is expressed as the case where the variance is greater than the mean, and in addition to the zero-scrambling condition. Pearson and deviation graphs were found to be inappropriate to use in counting data models. Suggesting the use of random quantile residuals in counting regression models and using Bell/zero-stacked Bell regression models as an alternative to classical counting models have contributed to the statistical literature.

Science Code	:	20513
Key Words	:	Count data regression models, zero inflated regression models, Bell type regression models, Residual diagnostic
Page Number	:	82
Supervisor	:	Prof. Dr. Hülya BAYRAK

TEŞEKKÜR

Bütün lisansüstü eğitim hayatım boyunca bana her türlü desteği veren yoğun çalışma temposuna rağmen benden yardımlarını, sabrını, bilgisini ve yol göstericiliğini esirgemeyen beni her zaman cesaretlendiren değerli danışman hocam Prof. Dr. Hülya BAYRAK' a ve doktora ders dönemimden itibaren geçen süre boyunca asla bıkmadan yardımlarını benden esirgemeyen, her çıkmaza düştüğümde değerli bilgilerini benden esirgemeyen bütün yoğunluğuna rağmen her zaman yanımda olan sevgili hocam Dr. Öğr. Üyesi Hatice Tül Kübra AKDUR' a sonsuz teşekkür ederim. Tez izleme komitelerinde her zaman yapıcı eleştiriler ve öneriler sunan sayın Prof. Dr. Birdal ŞENOĞLU ve sayın Prof. Dr. Fikri GÖKPINAR hocalarıma teşekkürü bir borç bilirim. Beni her zaman destekleyen ve yanımda olan canım aileme, tez sürecim boyunca benden yardımını ve desteklerini esirgemeyen canım arkadaşlarım ve değerli hocalarımın tamamına sonsuz teşekkürlerimi sunarım.

İÇİNDEKİLER

Sayfa

ÖZET iv

ABSTRACT v

TEŞEKKÜR vi

İÇİNDEKİLER vii

ÇİZELGELERİN LİSTESİ ix

ŞEKİLLERİN LİSTESİ x

SİMGELER VE KISALTMALAR xi

1. GİRİŞ 1

2. LİTERATÜR ÖZETİ 5

3. SAYMA REGRESYON MODELLERİ 11

3.1. Klasik Sayma Regresyon Modelleri 11

3.1.1. Poisson regresyon modeli 12

3.1.2. Negatif binom (NB) regresyon modeli 16

3.2. Sıfır Yığılmalı Regresyon Modelleri (Zero Inflated Regression Models) 19

3.2.1. Sıfır yığılmalı Poisson regresyon modeli (zero inflated Poisson regression-ZIP) 21

3.2.2. Sıfır yığılmalı negatif binom regresyon modeli (zero inflated negative binom-ZINB) 22

3.3. Bell Dağılımlı Regresyon Modelleri 23

3.3.1. Bell regresyon 24

3.3.2. Sıfır yığılmalı Bell regresyon (Zero inflated Bell regression-ZIBell) 28

4. ARTIKLARIN İNCELENMESİ 31

4.1. Pearson Artıklar 32

4.2. Sapma (Deviance) Artıklar 33

4.3. Pearson ve Sapma (Deviance) Artıklar ile İlgili Sorunlar 39

Sayfa

4.4. Rastgele Kantil Artıklar (RQR) 40
5. UYUM İYİLİĞİ ÖLÇÜTLERİ 43
5.1. Akaike Bilgi Kriteri (Akaike Information Criteria- AIC) 43
5.2. Bayesci Bilgi Kriteri (Bayesian Information Criteria- BIC) 44
5.3. Olabilirlik Oran Testi 45
5.4. Rootogram 45
5.5. Vuong Testi 46
6. UYGULAMA 47
6.1. Veri Setinin Tanıtılması ve Betimsel İstatistikler 47
6.2. Uygulama Sonuçları 54
7. SİMÜLASYON ÇALIŞMASI 63
7.1. Simülasyon Çalışması 1 63
7.2. Simülasyon Çalışması 2 69
8. SONUÇ VE ÖNERİLER 73
KAYNAKLAR 75
ÖZGEÇMİŞ 81

ÇİZELGELERİN LİSTESİ

Çizelge **Sayfa**

Çizelge 4.1. Kullanılan regresyon modellerine göre Pearson artıklar 32

Çizelge 4.2. Kullanılan regresyon modellerine göre sapma artıklar..... 34

Çizelge 4.3. Kullanılan regresyon modellerine göre RQR 42

Çizelge 5.1. Hilbe tarafından yapılan önerme 44

Çizelge 6.1. Kullanılan değişkenlere ait bilgiler..... 48

Çizelge 6.2. Cinsiyetin spor yapılan günlere dağılımı..... 49

Çizelge 6.3. Genel sağlık durumu spor yapılan gün dağılımı..... 50

Çizelge 6.4. Model uygunluğunun belirlenmesi..... 54

Çizelge 6.5. Aşırı yayılım testi sonucu 55

Çizelge 6.6. ZIBell model için parametre tahmini (log kısım)..... 55

Çizelge 6.7. ZIBell model için parametre tahmini (logit kısım)..... 55

Çizelge 6.8. Poisson model için parametre tahmini..... 57

Çizelge 6.9. NB model için parametre tahmini..... 57

Çizelge 6.10. Bell model için parametre tahmini 58

Çizelge 6.11. ZIP model için parametre tahmini (Log kısım) 58

Çizelge 6.12. ZIP model için parametre tahmini (Logit kısım)..... 58

Çizelge 6.13. ZINB model için parametre tahmini (Log kısım)..... 59

Çizelge 6.14. ZINB model için parametre tahmini (Logit kısım)..... 59

Çizelge 7.1. Farklı sıfır yığılma oranları altın kullanılan regresyon modellerine ait AIC değerleri..... 64

Çizelge 7.2. RQR için Shapiro- Wilk testi sonucu 72

ŞEKİLLERİN LİSTESİ

Şekil **Sayfa**

Şekil 3.1. Farklı λ parametreleri altında Poisson dağılımı 14

Şekil 3.2. Farklı k değerleri altında negatif binom dağılımı 18

Şekil 3.3. Farklı θ değerleri altında Bell dağılımı 25

Şekil 6.1. Spor_gun değişkenine ait dağılım 48

Şekil 6.2. Cinsiyet değişkenine ait dağılım 49

Şekil 6.3. Genel sağlık durumu değişkenine ait dağılım 50

Şekil 6.4. Yaş değişkenine ait dağılım 51

Şekil 6.5. Kilo değişkenine ait dağılım 52

Şekil 6.6. Sayma verilerinin modellenmesinde uygulanabilecek adımlar 53

Şekil 6.7. Pearson artıkların QQ Grafikleri 60

Şekil 6.8. Sapma (deviance) artıkların QQ grafikleri 61

Şekil 6.9. RQR QQ grafikleri 62

Şekil 7.1. Simülasyon sonucu elde edilen ortalama AIC değerleri 65

Şekil 7.2. RQR QQ grafikleri 66

Şekil 7.3. Pearson artıkların QQ grafikleri 67

Şekil 7.4. Sapma artıkların QQ grafikleri 68

Şekil 7.5. Pearson artıklara ait QQ grafiği 70

Şekil 7.6. Sapma artıklara ait QQ grafiği 70

Şekil 7.7. RQR ait QQ grafiği 71

SİMGELER VE KISALTMALAR

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

Simgeler	**Açıklamalar**
μ	Ortalama
k	Aşırı Yayılım Parametresi
p	Sıfır Yığılma Oranı
β	Regresyon Katsayıları

Kısaltmalar	**Açıklamalar**
AIC	Akaike Bilgi Kriteri
BIC	Bayesian Bilgi Kriteri
EM	Beklenti Maksimizasyon Algoritması
GLM	Genelleştirilmiş Lineer Model
LL	Log Olabilirlik
NB	Negatif Binom Regresyon
PR	Poisson Regresyon
RQR	Rasgele Kantil Artık
TÜİK	Türkiye İstatistik Kurumu
ZIBell	Sıfır Yığılmalı Bell Regresyon
ZINB	Sıfır Yığılmalı Negatif Binom Regresyon
ZIP	Sıfır Yığılmalı Poisson Regresyon

1. GİRİŞ

İstatistik biliminin en temel amaçlarından biri veriyi analiz etmek ve yorumlamaktır. Verileri doğru analiz etmek için analiz edilen verinin türünün (nicel,nitel vs.) ve dağılımının (normal dağılım, Poisson vb.) belirlenmesi gereklidir. İstatistikte kullanılan en temel dağılım normal dağılım olmakla beraber başka birçok dağılım türü de vardır. Sonuçların güvenilirliği verilerin dağılımı ile analizde varsayılan dağılım arasındaki farklılıktan etkilenebilir. Bu nedenle araştırmacının dağılımı doğru seçmesi oldukça önemlidir. Veriler her zaman normal dağılıma uyum göstermeyebilir.

Sayma verileri, normal dağılım varsayımına uyum göstermeyen bir veri tipidir. Normal dağılım sürekli veriler ile ilgilenirken sayma verileri kesikli veriler olup genellikle sağa çarpıktır. Bu tür verilere günlük hayat problemlerinde sıklıkla rastlanır. Çünkü adından da anlaşılabileceği gibi sayma verileri sayımdan kaynaklanan verilerdir. Bu verilere örnek olarak bir kişinin gün içerisinde tükettiği sigara sayısı, bir kavşakta belirli bir zaman aralığında meydana gelen kazaların sayısı, bir doktorun gün içerisinde muayene ettiği hasta sayısı verilebilir. Eğitimden, sağlığa, ekonomiye, zoolojiye varıncaya kadar farklı birçok disiplinde bu veri tipine rastlamak mümkündür.

Sayma verilerinin olduğu bir veri setinde normal dağılım varsayımı içeren klasik regresyon analizi ile analizi yapmak doğru bir yaklaşım olmayacaktır. Klasik doğrusal regresyon analizi yerine log bağlantı fonksiyonuna sahip olan sayma regresyon modellerinin kullanılması daha uygun olur. Eğer sayma verilerinin olduğu bir veri setinde bu veriler yokmuş gibi davranılarak klasik doğrusal regresyon yöntemi uygulanırsa yanlı ve tutarsız tahmin sonuçları elde edilecektir. Bu tür verilerin analizinde kullanılmak üzere literatürde farklı regresyon modelleri önerilmiş olup sayma verisinin yapısına bağlı olarak çeşitlenmektedir. Klasik sayma verilerinin modellenmesinde sıklıkla kullanılan regresyon modeli Poisson regresyondur. Ancak Poisson dağılımının en önemli özelliği olan varyans ve ortalamanın eşit olması durumuna gerçek hayatta rastlamak oldukça zordur. Sayma verilerinde genellikle gözlenen durum varyansın ortalamadan büyük olması yani aşırı yayılım (overdispersion) durumudur. Böyle durumlarda ise literatürde kullanılması önerilen regresyon modellerinden biri Negatif Binom (NB) regresyon modelidir. Bu tez çalışmasında aşırı yayılım olduğu durumda kullanılmak üzere yeni bir yöntem olan Bell regresyon modeli önerilmiştir. Bu regresyon modeli tek parametreli olduğundan parametre tahminlerinde

kullanılan en çok olabilirlik yöntemi sırasında oluşan denklemlerin çözümünde NB regresyon modeline göre kolaylık sağlamaktadır. Ayrıca çeşitli istatistiksel çıkarımlar içinde (skor fonksiyonu vb.) tek parametreli oluşu araştırmacıya kolaylık sağlamaktadır.

Sayma verilerinin regresyon kullanarak modellenmesinde sıklıkla görülen bir durum sıfır yığılma ile karşılaşılmasıdır. Sıfır yığılma durumunda beklenenden çok daha fazla sıfır vardır. Bu nedenle böyle verileri modellemek için farklı bir regresyon modeli olan sıfır yığılmalı regresyon modelleri önerilmiştir. Sıfır yığılmalı regresyon modellerini temelde iki parçalı bir yapı olarak düşünmek mümkündür. İlk parçasında veri seti içerisindeki sıfırlar göz önüne alınırken, ikinci parçasında farklı kesikli dağılımlardan gelen veriler yer almaktadır. Sıfır yığılmalı regresyon modeli buradaki bağımlı değişkenin geldiği dağılıma göre isim alır. En çok rastlanan regresyon modeli sıfır yığılmalı Poisson regresyon modeli olurken, aşırı yayılım varlığında klasik regresyon modellerinde olduğu gibi burada da sıfır yığılmalı negatif binom regresyon modeli kullanılmaktadır. Bu tez çalışmasında aşırı yayılım ve sıfır yığılma durumu olduğunda kullanılmak üzere yeni bir yöntem olan sıfır yığılmalı Bell regresyon modeli önerilmiştir.

Literatürde de görüldüğü gibi farklı veri tipleri için farklı modeller oluşturulabilmektedir. Bir araştırmacı için en iyi model, araştırdığı değişkene uygun şekilde cevap veren modeldir. Bu modeller gözlemlenen değerler ile model tarafından tahmin edilen değerler arasında yakın ilişki olan modellerdir. Yani ilgili araştırma için en iyi model, verilere iyi uyan (fitted) modeldir.

Farklı modellerin karşılaştırılması durumunda hangi modelin araştırma sonuçlarına daha iyi uyduğunun belirlenmesi amacıyla kullanılabilecek farklı istatistiksel kriterler / değerler vardır. Bunların başlıcaları Akaike Bilgi Kriteri (AIC), Bayesyen Bilgi Kriteri (BIC) ve log olabilirlik değeridir. Bunlar dışında kalan literatürde önerilmiş başka testler de vardır. Ancak en çok bunların kullanıldığı görülmektedir.

Uyum iyiliği ölçütleri dışında model seçimine karar verme de kullanılan bir diğer yöntem artık analizdir. Artıkları incelemek, modeller ve veriler arasındaki tutarsızlıkları belirlemenin bir yöntemidir. Sapma ve Pearson artıkları, genelleştirilmiş doğrusal modeller için doğru modelin belirlenmesinde sıklıkla kullanılır. Klasik doğrusal regresyonda bu artıklar normal dağılır. Ancak bununla birlikte sayma verileri gibi kesikli sonuç

değişkenlerinin modellenmesinde, artıklar normallikten uzak bir dağılım gösterir. Çizilen artıklar paralel eğriler oluşturarak görsel inceleme ve yorumlama için zorluğa yol açar. Bu nedenle, kesikli bağımlı değişkenleri modellemek için bu yöntemlerin kullanılması yanıltıcı olabilir.

1996'da Dunn ve Smyth tarafından tanımlanan rasgele kuantil artıklar (RQR) kesikli bağımlı değişkeni modellemek için yukarıda bahsedilen geleneksel artıklardan kaynaklanan sorunları çözmüştür. RQR, kesikli bağımlı değişkene sahip modellerin belirlenmesinde kullanılmak üzere önerilmiştir. RQR'nin ana fikri, kümülatif dağılım fonksiyonun (CDF) kullanılmasına dayanır. Ayrıca RQR'nin hesaplanması kolaydır, çünkü yöntem temelde yalnızca her gözlem için CDF'nin ters çevrilmesine dayanır. Bununla birlikte, sıfır yığılma durumunda da modellerin belirlenmesinde RQR uygulanabilir (Feng ve ark., 2020).

Bu tez çalışmasında sayma regresyon modelleri üzerine çalışan kişilere yeni bir bakış kazandırabileceği düşüncesiyle yeni bir yöntem olan Bell/ sıfır yığılmalı Bell regresyon modelleri üzerinde durulmuş ve sayma regresyon modellerinde artık analizi yapılmıştır. Literatürde incelendiği kadarıyla Bell/ sıfır yığılmalı Bell regresyon modellerinde RQR kullanarak model uyumuna ilk kez yer verilmiştir. Bu çalışmanın temel amaçları;

- Sayma regresyon modellerinde kullanılan yöntemleri karşılaştırmak,
- Yeni bir yöntem olan Bell ve sıfır yığılmalı Bell regresyon modellerini tanıtmak,
- Sayma regresyon modellerinde uygun modelin belirlenmesinde kullanılan RQR'ı tanıtmaktır.

Bu amaçlar kapsamında Türkiye İstatistik Kurumu'nda uygulanan araştırmalardan biri olan Türkiye Sağlık Araştırması'na ait veri seti kullanılarak Poisson, Neatif Binom, Bell, sıfır yığılmalı Poisson, sıfır yığılmalı negatif binom ve sıfır yığılmalı Bell regresyon modelleri incelenmiştir. Bu modellere ait Pearson, sapma ve RQR grafikleri çizdirilmiştir. Uyum iyiliği ölçütleri ve çizdirilen artıklara ilişkin grafikler göz önüne alındığında sıfır yığılmalı Bell modelinin uygunluğu görülmüştür. Bulunan sonuçlar yapılan simülasyon tasarımı ile de desteklenmiştir. Yapılan simülasyon çalışmasında ve gerçek veri uygulamasında artıkların elde edilmesinde kullanılan R kodları herhangi bir kütüphanede yer almamaktadır. Bu artıkların teorik çıkarımları bulunarak jenerik kodlar yazılmıştır.

Bu tez çalışmasının ikinci bölümünde literatürde sayma regresyon modelleri kullanılarak yapılan çalışmalar yer almaktadır. Üçüncü bölümünde klasik sayma regresyon modelleri, sıfır yığılmalı regresyon modelleri ve Bell dağılımlı regresyon modellerine ait bilgiler verilmiştir. Dördüncü bölümünde kullanılan artık türleri tanıtılmış ve teorik çıkarımlara yer verilmiştir. Beşinci bölümünde model karşılaştırmada kullanılan uyum iyiliği ölçütleri anlatılmıştır. Altıncı bölümde TÜİK tarafından yapılan Türkiye Sağlık Araştırması'ndan alınan veriler kullanılarak tanıtılan yöntemler veri setine uygulanmış elde edilen parametre sonuçları yorumlanmıştır. Ayrıca bahsedilen üç artık için grafikler çizilmiş, uyum iyiliği ölçütleri hesaplanmıştır. Yedinci bölümünde iki farklı simülasyon tasarımı anlatılmıştır ve bunlara ilişkin sonuçlara yer verilmiştir. Sekizinci ve son bölümünde ise tez çalışması sonucu elde edilen bulgular incelenmiş ve yorumlanmıştır.

2. LİTERATÜR ÖZETİ

Sayma verileri önerildiği ilk tarihten günümüze kadar olan süreçte son yıllarda daha da hızlı bir artış trendi ile modelleme çalışmalarında kullanılmaktadır. Sıfır yığılmalı modeller kullanılarak ilk yapılan çalışma Lambert (1992)'e aittir. Lambert (1992), bu çalışmasında sıfır yığılmalı Poisson regresyon modeli (Zero inflated Poisson- ZIP) baskı devre kartlarının üretiminde gözlenen lehimleme hata sayılarının bir model olarak ifade edilmesi konusunda kullanmıştır. Yaptığı çalışma sonucunda ZIP modelin kullanmış olduğu veriyi daha iyi açıkladığı sonucuna varmıştır. Sıfır yığılmalı negatif binom (Zero inflated negative Binom-ZINB) modelini ilk öneren ise Greene (1994) olmuştur. Greene bu yöntemi veride aşırı yayılım ve sıfırda yığılma olduğu durumlarda kullanılmak üzere yeni bir yöntem olarak önermiştir. Çalışmasında kullandığı ZINB modeli tüketici kredisine ilişkin bir veri setinde uygulamıştır. Ridout ve ark. (1998), yaptıkları çalışmada sayma verilerinde fazla sayıda sıfır olması durumunda kullanılabilecek modelleri araştırmışlardır. Bir elma çeşidinden elde edilen köklerin sayının modellenmesinde ZINB modelin daha uygun olduğunu görmüşlerdir. Melkersson ve Rooth (2000), çalışmalarında kadın doğurganlığını modellemek için sıfır yığılmalı sayma veri modeli kullanımını önermişlerdir. İsveçli kadınlardan oluşan bir örnek için tam doğurganlık verilerini kullanarak önerilen modeli klasik sayma veri modelleri ile karşılaştırmışlardır. Çalışma sonucunda geç yaşta evlenen kadınların erken yaşta evlenen kadınlara göre çocuksuz olma olasılığının daha yüksek olduğunu göstermişlerdir. Carrivick ve arkadaşları (2003), sayma regresyon modelleri kullanarak bir hastanede temizlik işçileri arasında mesleki yaralanma riskini azaltmada işyeri risk değerlendirme ekibinin etkinliğini değerlendirmişlerdir. Çalışma sonucunda ZIP modelin verilere uyum sağladığını görmüşlerdir. Yau ve ark. (2003), yaptıkları çalışmalarında bir dizi pankreas bozukluğu kalış süresi (LOS) verisini analiz etmek için sıfır yığılmalı negatif binom regresyon modelini kullanmışlardır ve parametre tahminlerini EM algoritması kullanılarak elde etmişlerdir. Yau ve ark. (2004), sıfır yığılma içeren olayları zaman serileri kullanarak analiz etmek için sıfır yığılmalı bir Poisson karma otoregresyon modeli önermişlerdir. Parametre tahminlerinde en çok olabilirlik yöntemi kullanmışlardır. Sezgin ve Deniz (2004) Türkiye'de grev sayılarına etki eden faktörleri bulmak için sayma veri modellerini kullanmışlardır. Çalışmalarında negatif binom regresyon modelini dikkate almışlardır. Lord ve ark. (2005), motorlu araç kazalarından elde edilen verilerin sayma regresyon modelleri ile uyumunu incelemişlerdir. Ardından, kaza verilerinde sıklıkla gözlemlenen "fazla" sıfırların veriyi nasıl etkilediğini göstermek için bir simülasyon

çalışması yapmışlardır. Sonuç olarak, sayma regresyon modellerinin bu tarz veriler için kullanılabilirliğini savunmuşlardır. Min ve Agresti (2005), sayma verilende gözlenen sıfır yığılma durumunun genellikle biyomedikal ve sosyolojik uygulamalarda ortaya çıktığından bahsetmişler ve sıfır yığılmalı sayma verilerinin tekrarlanan ölçümlerini modellemekte zor olduğunu söylemişlerdir. Bunun nedenin, fazladan sıfır sorununa ek olarak, aynı konuda farklı durumlarda yapılan ölçümler arasındaki korelasyonun da dikkate alınması gerektiği olduğunu vurgulamışlardır. Yaptıkları çalışma ile bu tür bir yanıt değişkeni üzerinde tekrarlanan ölçümler için rastgele etki modellerini tartışmışlardır. Alternatif bir yaklaşım olarak, kümülatif logit modelini, etkileri açıklamak için tek bir parametre kümesine sahip olan rastgele etkilerle uyarlamışlar ve önerdikleri yöntemleri örneklerle açıklamışlardır. Jung ve ark. (2005), çalışmalarında sıfır yığılmalı Poisson regresyon modelini sıfır yığılmalı negatif binom regresyon modeline karşı test etmek için bir skor testi önermişlerdir. Klasik yaklaşımı kullanan skor testlerinin küçük örneklemler için doğru tahmin vermeyebileceğini söylemişlerdir. Bu sorunu çözmek için parametrik bir bootstrap yöntemi önermişlerdir. Önerdikleri yöntemin anlamlılık düzeyini nominal değere yakın tuttuğu ve hipotezi test etmek için mevcut normal yaklaşımdan daha düzgün bir güce sahip olduğu gösterilmiştir. Simons ve ark. (2006), üniversite öğrencileri arasındaki esrar ile ilgili sorunlar ile sosyal normlar, dürtüsellik ve kullanım programı arasındaki ilişkinin incelenmesinde ZINB modelini kullanmışlardır. Yaptıkları çalışma sonucunun sosyal ve çevresel problemlere ilişkin teorilere uygun olduğunu göstermişlerdir. Famoye ve Singh (2006), aile içi şiddet içeren bir veri için ZIP ve genelleştirilmiş ZIP modellerini kullanmışlar ve ayrıca çalışmanın parametre tahminlerinin elde edilmesinde en çok olabilirlik yöntemi kullanmışlardır. Özmen ve Famoye (2007), çalışmalarında Dalyan Plajı'nda 1991-1993 yılları arasında güneşe maruz kalarak ölen caretta caretta yavrularının sayısının modellenmesinde sayma regresyon modellerini kullanmışlardır. Yapmış oldukları analiz sonucunda Genelleştirilmiş Poisson Regresyon modelinin daha iyi uyum sağladığını tespit etmişlerdir. Zeileis ve ark. (2008), sayma regresyon modellerinin kavramsal ve hesaplamalı özelliklerini inceledikten sonra, pscl paketindeki hurdle() ve zeroinfl() işlevlerinde hurdle ve sıfır yığılmalı regresyon modellerinin yeni bir uygulamasını tanıtmışlardır. Sumathi ve Rao (2009), çalışmalarında sıfır yığılmalı modeller üzerine literatürün kısa bir incelemesini sunmuşlardır. Ayrıca en çok olabilirlik tahmin edicisinin elde edilmesi sırasında oluşan denklemlerin çözümündeki hesaplama güçlükleri ortaya koymuşlardır. Çalışmalarının sonunda dental bir çalışmadan elde edilen bir veri seti kullanarak elde ettikleri sonuçları göstermişlerdir. Flynn ve Francis (2009), sayma verilerinde fazla sayıda sıfır olması durumunu ve bu verilerde doğrusallık

varsayımının sağlanmaması sorunlarını incelemişlerdir. Örnek bir uygulama olarak sigortacılık alanında hasar sayısının modellenmesinde sıfır yığılmalı modelleri kullanmışlardır. Yang ve arkadaşları (2009), veride aykırı değer olması durumunda aykırı değerleri belirleme yöntemleri ve buna ilişkin uygulamaları sonuçları ile öne sürmüşlerdir. Bu değerlerin etkisini azaltmak veya ortadan kaldırmak için kesilmiş veya Winsorized ortalamaya dayalı parametre tahminleri önermişlerdir. Simülasyon çalışması yaparak sıfır yığılmalı Poisson modelde sonuçları göstermişlerdir. Yeşilova ve ark. (2010), bitki koruma çalışmalarında sıfır yığılmalı gözlemlerin sıklıkla görüldüğünü ve bu nedenle sıfır yığılmalı modellere bu alanda ihtiyaç duyulduğunu vurgulamışlardır. Yaptıkları çalışmada bu alandan sağlanan bir veri için uygun regresyon modeli olarak sıfır yığılmalı negatif binom regresyon olduğunu göstermişlerdir. Javali ve Pandit (2010), diş çürüğü ile ilişkili faktörleri analiz etmeyi ve belirlemeyi amaçlayan çalışmalarında sıfır yığılmalı modelleri kullanmışlardır. Mart-Ağustos 2007'de Hindistan, Karnataka, Dharwad'da 18-40 yaşları arasındaki 1760 bireyden oluşan sistematik rastgele örneklem üzerinde bir çalışmalarını yapmışlardır ve elde ettikleri verileri sıfır yığılmalı negatif binom regresyon modelleri kullanarak modellemişlerdir. Hu ve ark. (2011), Poisson, NB, ZIP, Poisson hurdle (PH), ZINB, negatif binom hurdle (NBH) modellerinin özeliklerini incelemişlerdir. Bu modelleri toplum temelli kötü madde kullanımı tedavisine katılan kadın hastalar arasında daha önce tamamlanmış bir klinik araştırmadan elde edilen verileri kullanarak, HIV risk azaltma müdahalesinin korunmasız cinsel olayların sayısı üzerindeki etkisini modellemek için kullanmışlardır. Her modelden elde edilen uyum iyiliği ve tedavi etkisinin tahminleri karşılaştırmışlardır. Çalışma sonucunda ZINB modelinin en uygun model olduğunu görmüşlerdir. Kaya ve Yeşilova (2012), Yüzüncü Yıl Üniversitesi (YYÜ) e-posta sunucusundan personelin 2009 bahar eğitim öğretim döneminde yaptıkları e-posta trafiğini inceledikleri veride sıfır yığılma olduğunu görmüşler ve veriye ZIP, ZINB, PH ve NBH regresyon yöntemleri uygulamışlardır. Çalışma sonucunda ZINB ve NBH regresyon modellerinin veri için daha uygun olduğu göstermişlerdir. Garay ve ark. (2015), çalışmalarında ZINB regresyon modelinde kullanılan Bayes yaklaşımı önermişlerdir. Önerdikleri yöntemin performansını simülasyon çalışması ve gerçek veri seti analizi ile göstermişlerdir. Wang ve arkadaşları (2015), çalışmalarında Almanya'da sağlık hizmetleri çalışmalarını sıfır yığılmalı negatif binom model kullanarak incelemişlerdir. Model parametrelerini tahmin etmek için EM algoritmasını önermişlerdir. Martin ve Hall (2016), sayma verileri için sapma oranına dayalı R^2 benzeri bir hesaplama geliştirmişlerdir. Bu hesaplamayı kullanarak ZIP ve ZINB regresyonları için R^2 önermişlerdir. Ayrıca, modele alakasız bağımsız değişkenlerin dahil

edilmesi nedeniyle bu istatistiklerin şişmesini önlemek için bu miktarların düzeltilmiş R^2 benzeri versiyonlarını da çalışmalarına eklemişlerdir. Önerilerini simülasyon ve iki farklı gerçek veri seti üzerinde göstermişlerdir. Fagundes ve ark. (2016), Yazılım alanında çok sayıda sıfır içeren yazılım hatası verilerinin olduğunu söylemişlerdir. Çalışmalarında bu tür verileri tahmin etmede ve tahmin doğruluğunu iyileştirmede sıfır yığılmanın dikkate alınması gerektiğini vurgulamışlardır. Ayrıca yaptıkları çalışmayı makine öğrenmesi ve istatistiksel literatürdeki diğer yöntemlerle karşılaştırılmıştır. Yöntemlerin performansı, göreceli hataların ortalama büyüklüğüne dayalı olarak değerlendirilen tahmin doğruluğu ile ölçmüşlerdir. Xu ve ark. (2017), depresyon belirtilerinin sayısına ilişkin bir risk faktörü çalışması için en uygun modeli belirlemek amacıyla geniş bir anket örneklemiyle farklı sayma regresyon modellerini uyum iyiliği kullanarak karşılaştırmayı amaçlamışlardır. En büyük log olasılığına ve en küçük AIC' ye sahip ZINB modelinin bu veri seti için en uygun model olduğunu belirtmişlerdir. Avcı (2018), çalışmasında şizofren hastalara ilişkin verilere sayma verisi regresyon modellerini uygulamıştır. Kullandığı regresyon modelleri Poisson, Negatif Binom ve Conway-Maxwell-Poisson (COM-Poisson) Regresyondur. Analiz sonuçlarına göre; veri setine en iyi uyan modelin Negatif Binom Regresyon modeli olduğunu göstermiştir. Altun (2018), yapmış olduğu çalışmada sıfır yığılmalı modellere yeni bir yaklaşım olarak sıfır yığılmalı Poisson- Lindley modelini önermiştir. Poisson dağılımının aksine, Poisson-Lindley dağılımı aşırı yayılıma izin vermesi nedeniyle, bu modelin aşırı yayılımlı ve sıfır yığılmalı verileri modellemek adına iyi bir seçenek olduğunu öne sürmüştür. Önerdiği modelin gerçek veri üzerinde uygulama çalışmasını yapmıştır ve Poisson ve ZIP regresyon modelleriyle karşılaştırmıştır. Sıfır yığılmalı Poisson-Lindley regresyon modelinin, sıfır yığılmalı ve aşırı yayılım gösteren veri seti için, sıfır yığılmalı Poisson regresyon modelinden daha iyi uyum sağladığını göstermiştir. Castellares ve ark. (2018), olasılık fonksiyonu basit bir forma sahip olan tek parametreli Bell dağılımını tanıtmışlardır. Ayrıca çalışmada incelenen bu dağılımı dikkate alarak, yanıt değişkeninin bir sayma verisi olduğu durumda kullanılmak üzere yeni bir regresyon modeli önermişlerdir. Bell regresyon modelinin kullanışlılığını göstermek için bir uygulama yapmışlardır. Lemonte ve ark. (2019), sıfır yığılmalı Bell dağılımını tanıtmışlardır. Sıfır yığılmalı Bell regresyon parametrelerini tahmin etmek için en çok olabilirlik yöntemi kullanmışlardır. Sıfır yığılmalı Bell regresyon modelinin pratikte kullanışlılığını göstermek için enfekte kan hücrelerinin sayımına yönelik bir uygulama çalışması yapmışlardır. Kim ve arkadaşları (2019), Güney Kore'de iklim değişikliği sonucu aşırı sıcağa maruz kalan kişilerde yaşanan ölüm sayısını tahmin etmek için sıfır yığılmalı regresyon modellerini kullanmışlardır.

Çalışma sonucunda ZIP regresyon modelinin ele alınan sorun için kullanılabilecek en uygun istatistiksel yaklaşım olduğunu göstermişlerdir. Kılıç ve Bayrak (2020), yapmış oldukları çalışmada Türkiye İstatistik Kurumu (TÜİK) hanehalkı bilişim araştırması verilerine en uygun modelin ZINB regresyon modeli olduğunu öne sürmüştür. Feng ve ark. (2020), sayma regresyon modellerinde uyum iyiliğinde kullanılmak üzere farklı artık türleri olan Pearson, sapma ve rasgele kantil artıkları (Randomize Quantile Residual- RQR) karşılaştırmışlardır. RQR artıkların diğer artıklardan daha iyi sonuç verdiğini simülasyon ve gerçek veri uygulaması üzerinde göstermişlerdir. Amin ve ark. (2020), yanıt değişkeni aşırı yayılıma sahip sayma verisi olduğu durumda Bell regresyonu kullanmayı önermişlerdir. Bell regresyon katsayılarının genellikle en çok olabilirlik tahmin edicisi (MLE) kullanılarak tahmin edildiğini vurgulamışlardır. Klasik MLE' nin çoklu bağlantıya duyarlı olduğundan çoklu bağlantı problemlerine çözüm olarak Bell ridge regresyonu (BRR) önermişlerdir. BRR'nin değerlendirilmesi için, hata kareler ortalamasını (MSE) bir değerlendirme kriteri olarak kabul etmişlerdir. Tahmin edicinin performansını ölçmek için Monte Carlo simülasyon çalışması yapmışlardır. Ayrıca, BRR tahmin edicisinin üstünlüğünü göstermek için iki gerçek veri seti kullanarak yorumlamışlardır. Majid ve ark. (2022), Bell regresyon modelinde ortaya çıkabilecek sorunlardan biri olan çoklu bağlantı sorununu ortadan kaldırmak için bir Bell Liu regresyon (BLR) tahmin edicisini önermiştir. Önerilen tahmin edicilerin performansını değerlendirmek için, bir Monte Carlo simülasyon çalışması yapmışlardır. Akdur (2022), çalışmasında aşırı yayılım ve bağımlılık yapısını ele almak için Bell dağılımına dayalı boylamsal (longitudinal) sayma verileri için yeni bir marjinal model önermiştir. Bell marjinal modeli, 5 yıllık zaman diliminde Yunanistan'da yaralanmalara neden olan kazaların sayısını analiz etmek için kullanmıştır. Ayrıca bu model için tanı ölçütleri ve simüle edilmiş zarf algoritması kullanmıştır.

3. SAYMA REGRESYON MODELLERİ

Sayma regresyon modeli, bağımlı değişkeninin sayı olduğu istatistiksel modellere verilen isimdir (Hilbe, 2014). Sayma verilerinin modellenmesi anlatılmadan önce "sayma verileri" ile tam olarak ne kastedildiğinin netleştirilmesi önemlidir. Adından da anlaşılacağı gibi, sayma verileri, sayımdan kaynaklanan verilerdir. "Sayma" sözcüğü genellikle birimleri, öğeleri veya olayları sıralamak için fiil anlamında kullanılır. Örneğin, bir otoyolda gözlemlenen ölümlü kazaların sayısı, belirli bir hastanede yatan kişi sayısı veya spor yapılan gün sayısı gibi.

Sayma verileri ise, sayma yoluyla ifade edilebilen olaylar veya öğeler hakkında yapılan gözlemleri ifade etmek için kullanılır. İstatistikte sayma verileri, sıfır ve sıfırdan daha büyük değerler alabilen negatif olmayan tamsayı değerlerine sahip gözlemleri ifade etmektedir. Bu veriler kesikli bir dağılım özelliği gösterirler. Teorik olarak, sayma verileri sıfırdan sonsuza kadar değişebilir, ancak çoğunlukla belli bir değerle sınırlıdır. Sayma verileri aktüerya, ekonometri, biyoistatistik, eğitim, sağlık gibi birçok farklı alanda kullanılır.

Sayma verileri söz konusu olduğunda dağılım genellikle sağa çarpıktır. Bu nedenle normallik varsayımı sağlanamadığından klasik regresyon yöntemleri kullanılmaz (Pittman ve ark., 2020). Eğer sayma regresyon modelleri yerine klasik lineer regresyon modeli kullanılırsa yanlı parametre tahminleri elde edilir.

Bu tez kapsamında sayma regresyon modelleri içerisinde Poisson regresyon, Negatif Binom (NB) regresyon, sıfır yığılmalı regresyon modelleri (sıfır yığılmalı Poisson, sıfır yığılmalı negatif binom) ve yeni bir regresyon modeli olan Bell regresyon modeli ile sıfır yığılmalı Bell regresyon modeli anlatılacaktır.

3.1. Klasik Sayma Regresyon Modelleri

Sayma verilerinin modellenmesinde bağımlı değişkenin geldiği dağılım önem arz etmektedir. Literatürde sayma regresyon modelleri üzerine yapılan çalışmalar incelediğinde çoğunlukla kullanılan modellerin Poisson ve Negatif Binom regresyon modelleri olduğu görülmektedir. Bağımlı değişken kesikli dağılım ailesinden olan Poisson dağılımından ya da negatif binom dağılımından gelmekte ise regresyon modelinin adı bu dağılımdan gelir.

3.1.1. Poisson regresyon modeli

Poisson regresyon modeli, sayma verilerinin modellenmesinde en çok kullanılan yöntemlerin başında gelmektedir. Sayma verilerinin regresyon ile analiz edilmesinde kullanılan diğer tüm modeller Poisson regresyon modeli temel alınarak ortaya çıkarılmıştır. Ancak, Poisson dağılımının temel varsayımı olan ortalama ve varyansın birbirine eşit olması nedeniyle, gerçek dünyada çok uygulanabilir değildir.

Poisson Dağılımına İlişkin Varsayımlar

1) Poisson modelinde bağımlı değişken bir sayı olmalıdır.
2) Dağılım tek parametrelidir ve ortalama parametresi genellikle λ veya μ ile gösterilir. Ortalama aynı zamanda bir oran parametresidir. Birim zaman, alan veya hacim başına bir öğenin veya olayın gerçekleşmesinin beklenen sayısıdır.
3) Bağımlı değişken, negatif olmayan tam sayılardır. Yani, $Y \geq 0$'dır.
4) Gözlem değerleri birbirinden bağımsızdır.
5) Gözlemlenen sayıların hiçbir hücresi, ampirik dağılımın ortalamasına dayalı olarak beklenenden önemli ölçüde farklılaşmamaktadır. Örneğin, veriler, belirli bir ortalamaya sahip bir Poisson dağılımına dayalı olarak beklenenden daha fazla sıfır sayısına sahip olmamalıdır.
6) Pearson Ki-kare dağılım istatistiği 1'e oldukça yakın bir değere sahiptir. 1 değeri, yanıt değerindeki gözlenen ve tahmin edilen varyansların aynı olduğu durumda ortaya çıkar.

İncelenen bir model bahsedilen varsayımları sağlarsa veriler Poisson regresyon modeli kullanarak modellenebilir.

Poisson regresyon modeli, dağılım varsayımı ve dağılım parametresinin açıklayıcı değişkenlerin bir fonksiyonu temeline dayanır. Poisson dağılımı yukarıda da bahsedildiği gibi tek parametreli bir dağlımdır. Bu parametrenin λ ile gösterildiği varsayılsın. λ açıklayıcı değişkenler olan Xi'lerin log-lineer fonksiyonu olarak yazılabilir. Yani,

$\lambda_i = \exp(\beta_0 + \sum_j x_{ij}\beta_j)$ ya da
$\lambda_i = \exp(x_i'\beta)$ şeklindedir.

Burada β parametre vektörüdür.

Bağımsız örneklem varsayımı altında, modele ilişkin parametreler en çok olabilirlik tahmin edicisi kullanılarak tahmin edilebilir. Bu tahmin edici doğrusal olmamasına ve kapalı biçimde çözülemez olmasına rağmen, log olabilirlik fonksiyonunun içbükey olmasından kaynaklı olarak bu değeri maksimum yapan nokta yinelemeli algoritmalar kullanılarak bulunabilir. Doğru tanımlama altında, tahmin edici, özellikle asimptotik etkinlik ve normallik olmak üzere, en çok olabilirlik tahmin edicilerinin tüm istenilen özelliklerine sahiptir.

Poisson dağılımı için ortalamadan bağımsız bir varyans hesaplanamamaktadır. Bu özellik çok kısıtlayıcıdır. Bununla birlikte, Poisson regresyonu sağlamdır (robust). Çünkü varyans ortalamaya eşit olmasa bile, β için tahmin edici, ortalama fonksiyon olan λ için doğru bir şekilde belirtildiği sürece tutarlı kalır (Winkelmann, 2000).

Normal lineer modelin sürekli veriler için farklı modellerle karşılaştırmalarda kullanıldığı gibi Poisson regresyon modeli de sayma verileri için kullanılan kıyaslama modelidir. Yani sayma veri modelleri arasında karşılaştırma yapılmak istendiğinde Poisson regresyon modeli kullanılan modellerin başında gelmektedir. Birçok farklı disiplindeki veriyi modellemek için bu regresyon modeli kullanılabilir. Örneğin; kaza, sağlık, demografi, eğitim, pazarlama ve işgücü hareketliliği gibi çeşitli alanlarda uygulamalarını görmek mümkündür.

Poisson regresyon modeli için üç temel varsayım sağlanmalıdır. Bunlar aşağıda verilmiştir:

- $y_i|x_i \sim Poisson[\lambda(x_i)]$. Burada $y_i|x_i$, x_i biliniyorken y_i'nin koşullu dağılımını temsil etmektedir. $\lambda(x_i) = E(y_i|x_i)$'dir.
- $\lambda(x_i;\beta) = \exp(x_i'\beta)\,, i = 1,2,\dots,n$. Burada β, $(k \times 1)$ boyutlu parametre vektörüdür.
- $(y_i|x_i), i = 1,2,\dots,n$ bağımsız dağılımlıdır.

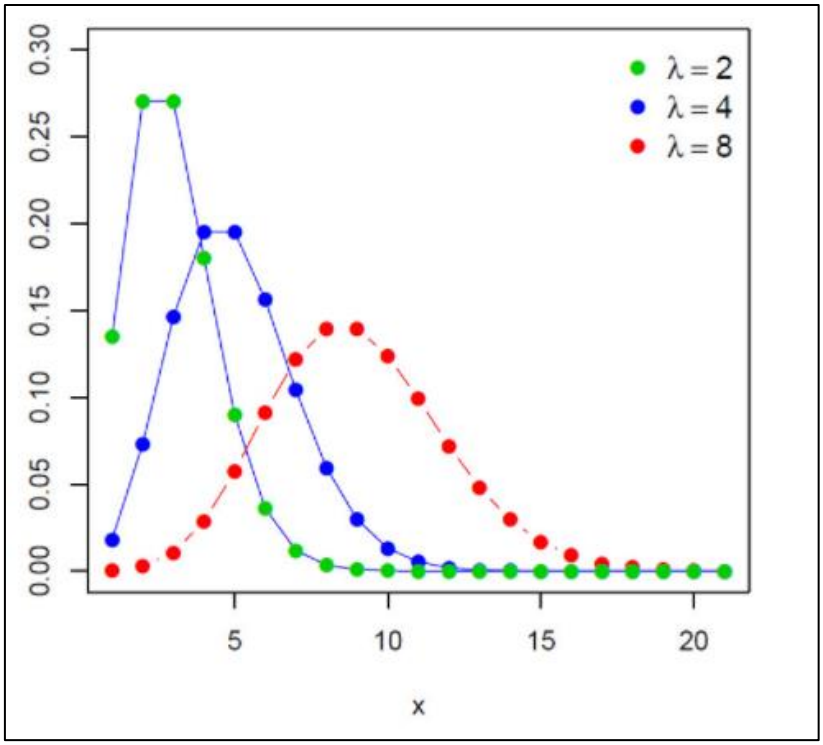
Poisson regresyon, sayma verilerini ve çapraz tablolarını modellemek için kullanılan genelleştirilmiş bir doğrusal modeldir. Poisson regresyonda Y yanıt değişkeni Poisson dağılımına sahiptir ve beklenen değerinin logaritmasının bilinmeyen parametrelerin doğrusal bir kombinasyonu ile modellenebileceği varsayılır. Bir Poisson regresyon modeli, log-doğrusal model olarak da bilinmektedir.

Poisson regresyon, doğrusal modellerin genelleştirilmiş ailesine aittir (Agresti, 2003). Bu modeller, sıradan doğrusal regresyonun kapsamını iki şekilde genişletir. İlk olarak, açıklayıcı değişkenlerin doğrusal fonksiyonları olarak, ortalamanın kendisinden ziyade bağımlı değişkenin koşullu ortalamasının dönüşümlerini tanımlarlar; ikincisi, bağımlı değişkenin normalden farklı koşullu dağılımlara sahip olmasını sağlarlar (Durmuş ve Güneri, 2020).

x_i ve y_i bir veri setinden gözlemler olsun. Burada x_i ve y_i sırasıyla bağımsız değişkenlerin ve bağımlı değişkenlerin bir vektörüdür. Poisson regresyon analizi, bağımlı değişken y_i' nin Poisson dağılımını gösterdiğini varsayar. λ_i parametresi ile Poisson dağılımı için olasılık fonksiyonu Eş. 3.1' de verilmektedir.

$$f(y_i|x_i) = \frac{{\lambda_i}^{y_i} e^{-\lambda_i}}{y_i!}, y_i = 0,1,2, \ldots. \quad (3.1)$$

Bu ifadede y_i bağımlı değişken ve λ_i dağılımın ortalamasıdır. Poisson olasılık dağılımı sağa çarpıktır. Ancak λ_i büyüdükçe dağılım normal dağılıma yaklaşır. Şekil 3.1 de farklı λ parametreleri altında Poisson dağılımı gösterilmektedir.



Şekil 3.1. Farklı λ parametreleri altında Poisson dağılımı

Poisson regresyon modelinin en önemli özelliği, ortalama ve varyansın eşit olmasıdır. Aşırı veya az yayılımlı veri kümeleri Poisson dağılımı ile modellenemez çünkü koşullu beklenen değerin varyansa eşit olduğu varsayımında bozulmalar görülür ve varsayım karşılanmaz. Bu durumda veri setini güncellemek veya farklı yöntemlerle analize başlamak çözüm olabilir.

y_i' nin beklenen değeri ve varyansı Eş. 3.2 de verilmiştir.

$$\lambda_i = E(y_i|x_i) = Var(y_i|x_i) \tag{3.2}$$

y_i' nin beklenen değerinin negatif değerler almamasını sağlamak için, beklenen değer ile bağımsız değişkenler arasındaki ilişkiyi gösteren link fonksiyonu Eş 3.3 ve Eş 3.4 de verilmiştir (Cameron ve Trivedi, 2013).

$$\log(\lambda_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m \tag{3.3}$$

Veya

$$\lambda_i = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m) = e^{x_i{}'\beta} \tag{3.4}$$

Burada β_0, β_1, …, β_m bilinmeyen parametre vektörünü temsil eder. Poisson regresyon analizinde, y_i bağımlı değişkeninin dağılımına dayalı olarak β tahmin edicileri hesaplamak için birçok yöntem vardır. Bu yöntemlerden en yaygın olarak kullanılan en çok olabilirlik yöntemidir. Olabilirlik yönteminde genellikle Newton Raphson yineleme tekniği kullanılmaktadır. Poisson regresyon modelinin log olabilirlik fonksiyonu Eş 3.5 de verilmiştir.

$$L\big(\beta\big|(y,x)\big) = \sum_{i=1}^{n} P(y_i|\lambda_i) = \prod_{i=1}^{n} \frac{\lambda_i{}^{y_i} e^{-\lambda_i}}{y_i!} \tag{3.5}$$

Bu fonksiyonun logaritması alındığında Eş 3.6’ da verilen şekle gelmektedir.

$$lnL(\beta) = \sum_{i=1}^{n} \{y_i \ln(\lambda_i) - \lambda_i - lny_i!\} \tag{3.6}$$

Poisson regresyon modelinde ilgilenilen olayın gözlemlenen sayısı bağımlı değişkeninin Poisson dağılımı gösterdiği varsayılmaktadır. λ_i nin logaritmasının, bağımsız değişkenlerin doğrusal fonksiyonu olduğu varsayılmaktadır.

Link (Bağlantı) Fonksiyonu:

Genelleştirilmiş doğrusal modellerde link fonksiyonu, doğrusal olmayan bir ilişkiyi doğrusal hale getirmek için kullanılır. Link fonksiyonu, bir GLM'nin rastgele ve sistematik (rastgele olmayan) bileşenlerini birbirine bağlar. Rastgele bileşen, X|Y için bir olasılık dağılımını belirtirken, sistematik bileşen, bir parametre ile tahmin ediciler arasında ilişki kurar (internet1). Yani genelleştirilmiş doğrusal modeller, yanıtın beklenen değerini modeldeki doğrusal tahmin edicilerle ilişkilendiren bir fonksiyondur. Örneğin, Kategorik bir bağımlı değişkeninin düzeylerinin olasılıklarını sürekli bir ölçeğe dönüştürür. Dönüşüm tamamlandıktan sonra, tahmin ediciler ve bağımlı değişken arasındaki ilişki doğrusal regresyon ile modellenebilir.

Link fonksiyonunun genel formu Eş 3.7 de verilmiştir.

$$g(\mu_i) = X_i'\beta \tag{3.7}$$

Burada $g(.)$, link fonksiyonu, μ_i ortalama, X_i bağımsız değişkenlere ilişkin vektör ve β ise katsayılar matrisini göstermektedir. Ayrıca link fonksiyonu diferensiyellenebilir ve monoton bir fonksiyondur. Link fonksiyonunun seçimi üstel dağılımlar ailesindeki dağılımın tipine bağlıdır. Üstel dağılımlar ailesinin her bir üyesi için birçok uygun link fonksiyonu vardır.

3.1.2. Negatif binom (NB) regresyon modeli

Poisson regresyonda aşırı yayılım (overdispersion) ile yani varyansın ortalamadan büyük olması durumu ile karşılaşılabilir. Böyle bir durumda yayılım parametresi içeren farklı bir regresyon yöntemi kullanılmalıdır. Bunlardan bir tanesi ve en çok kullanılanı ise negatif binom regresyon modelidir. Eğer dağılımda aşırı yayılım durumu varsa, Poisson regresyon modeli sonucu elde edilen tahminler tutarlı olacaktır ancak tahmin edicilerin standart hataları yanlı olacaktır. Sonuç olarak araştırmacı hatalı istatistiksel yorumlar yapacaktır. Bu nedenle aşırı yayılım olduğu durumda NB regresyon modeli bir seçenek olarak kullanılır.

NB regresyonda bağımlı değişken negatif binom dağılımından gelmektedir. Kurulan regresyon modeli bağımlı değişkenin gözlenen bir sayı olması dışında klasik çoklu regresyona benzemektedir. Bağımlı değişkenin alabileceği değerler negatif olmayan tam sayılardır: 0, 1, 2, 3, vb.

NB regresyon, Poisson modelindeki varyansın ortalamaya eşit olduğu şeklindeki kısıtlayıcı varsayımı genişleten Poisson regresyonunun bir genellemesidir. Bu model, Poisson-Gamma karma dağılımına dayanmaktadır (Cameron ve Trivedi, 2013). Poisson ve Gamma dağılımlarının ilişkisinden elde edilen formülasyon popülerdir. Bunun nedeni Gamma dağılımı kullanılarak Poisson heterojenliğinin modellenmesine izin verilmesidir. Burada heterojenlik, Poisson regresyon modelindeki ortalamadır.

Poisson ve Gamma ilişkisinden negatif binom dağılımının elde edilmesine ilişkin çıkarım Eş 3.8 de verilmiştir:

$$f(k) = \int_0^\infty Poisson(k|\lambda) Gamma\left(\lambda \middle| r, \frac{(1-p)}{p}\right) d\lambda$$

$$= \int_0^\infty \frac{\lambda^k}{k!} \exp(-\lambda) \frac{\lambda^{r-1} \exp\left(-\frac{\lambda p}{1-p}\right)}{\Gamma(r)\left(\frac{1-p}{p}\right)^r} \, d\lambda$$

$$= \frac{1}{k!\,\Gamma(r)} p^r \left(\frac{1-p}{p}\right)^r \int_0^\infty \lambda^{(r+k)-1} \exp\left(-\frac{\lambda}{1-p}\right) d\lambda$$

$$= \frac{1}{k!\,\Gamma(r)} p^r \left(\frac{1-p}{p}\right)^r (1-p)^{r+k} \Gamma(r+k)$$

$$= \frac{\Gamma(r+k)}{k!\Gamma(r)} p^r (1-p)^k \qquad (3.8)$$

Yukarıdaki çıkarımda Poisson parametresi λ ve Gamma parametreleri ise p ve r dir.

Poisson dağılımı, ortalaması 1 ve yayılım parametresi ν olan bir gamma değişkeni dahil edilerek Eş 3.9' da verildiği gibi genelleştirilebilir.

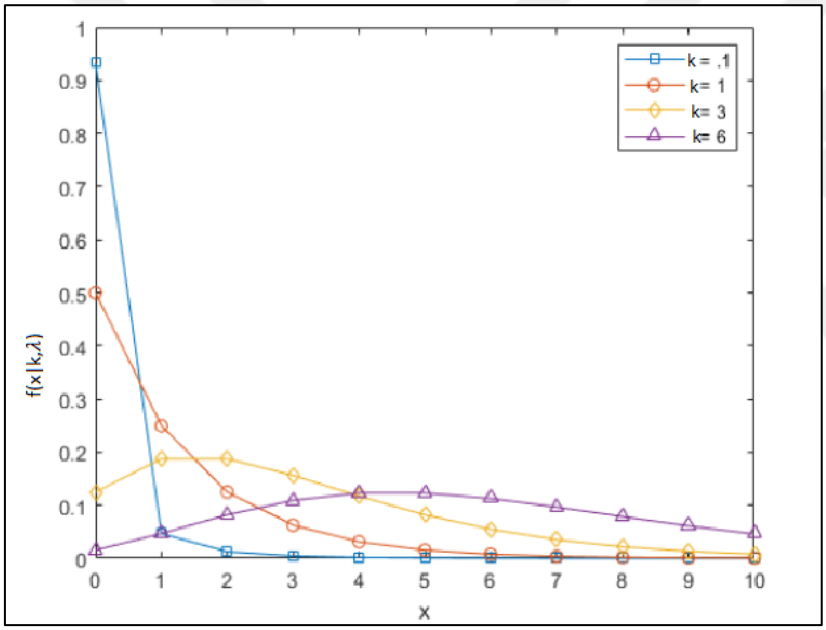
$$f(y; \lambda, v) = \frac{(\lambda_i v_i)^{y_i} e^{-\lambda_i v_i}}{y_i!} \qquad (3.9)$$

Burada v, gamma ölçek parametresidir. Bu değer, negatif binom heterojenliğine ya da diğer bir ifadeyle aşırı yayılım parametresinin tersine eşittir.

Negatif binom dağılımına ilişkin olasılık fonksiyonu Eş 3.10 da verilmiştir.

$$P(y_i; \lambda_i, k|x_i) = \frac{\Gamma(y_i+k)}{\Gamma(k)\Gamma(y_i+1)} \left(\frac{\lambda_i}{\lambda_i+k}\right)^{y_i} \left(\frac{k}{\lambda_i+k}\right)^{k} \quad (3.10)$$

Bu ifade de k = 0 durumunda fonksiyon poisson dağılımına indirgenir. Şekil 3.2 de farklı k değerleri altında negatif binom dağılımını gösterilmektedir.



Şekil 3.2. Farklı k değerleri altında negatif binom dağılımı

Negatif binom dağılımına ait beklenen değer ve varyans Eş 3.11 de verilmiştir:

$$E(y_i) = \lambda_i$$
$$Var(y_i) = \lambda_i + \frac{\lambda_i^2}{k} \quad (3.11)$$

Negatif binom regresyonunda da Poisson regresyonda verilen Eş. 3.3 ve Eş. 3.4 de verilen link fonksiyonu kullanılarak beklenen değer ile bağımsız değişkenler arasındaki ilişki ifade edilebilir. Ancak negatif binom regresyon modelinde Poisson regresyon modelinden farklı olarak β tahmin edicisine ait log olabilirlik fonksiyonu Eş 3.12 de verilmiştir.

$$ln(y_i|\lambda_i, k) = \sum_{i=1}^{n}(\log\left(\frac{\Gamma(y_i+k)}{\Gamma(y_i+1)\Gamma(k)}\right) - (y_i + k)\log(1 + k\lambda_i) + y_i \log(k\lambda_i)) \quad (3.12)$$

Negatif binom regresyon modeli ile Poisson regresyon modeli arasında doğrusallık, gözlemlerin bağımsızlığı gibi birçok ortak nokta vardır. Ancak Negatif binom modelinde bağımlı değişkene ilişkin koşullu varyansı, koşullu ortalamadan daha büyük olabilir (Hilbe, 2014).

Negatif binom regresyon, Poisson regresyonuna alternatif bir yöntem olarak kullanılmaktadır. Negatif binom regresyon modelinde, bağımlı değişkeninin varyansı, beklenen değerinden daha büyük değerler almaktadır. Geleneksel negatif binom regresyon modelinin parametreleri, GLM'lerin bir üyesi olarak maksimum olabilirlik yöntemi kullanılarak tahmin edilebilir (Hilbe, 2014).

Negatif binom regresyonu kullanmanın en önemli amacı varyans değerinin ortalamayı aştığı veya gözlemlenen varyansın beklenen varyansı aştığı verileri modellemektir.

3.2. Sıfır Yığılmalı Regresyon Modelleri (Zero Inflated Regression Models)

Sayma verileri söz konusu olduğunda yanıt değişkeninde sıfır sayısının fazla olması durumuna "sıfır yığılma" denilmektedir. Sıfır yığılmalı veri setleri ile uygulamada sıklıkla karşılaşılmaktadır. Min ve Agresti (2005) sıfır yığılmalı verileri "veride çok sayıda sıfır olmasından dolayı klasik genelleştirilmiş doğrusal modellerin uyumsuz olduğu veriler" olarak tanımlamaktadır.

Veri setindeki sıfırlar, gerçek sıfırlar veya örnekleme sıfırları olarak sınıflandırılabilir. Gerçek sıfırlar, gerçekten boş olan sıfır yanıtlarını temsil eder. Örneğin, eğitim araştırmasına ilişkin bir ankette "Üniversiteye hazırlık için kaç tane kursa katıldınız?" sorusu olduğu varsayılsın. Örneklemdeki bazı katılımcıların üniversiteye başvurma düşüncesi olmayabilir. Böyle bireyler için bu sayı hiçbir zaman sıfırdan büyük olamaz. Örnekleme sıfırları ise bir olasılık olarak ortaya çıkar. Örneğin, bazı öğrenciler kendilerini hazır hissedebilir ve hazırlık kursuna katılmak için hiçbir sebepleri olmayabilir.

Sıfır yığılma gösteren veri setlerinde sıfır yığılma durumu gözardı edilip klasik yöntemlerle analiz yapılırsa hatalı ve anlamsız sonuçlar ortaya çıkacaktır. Ayrıca parametre tahminlerinde yanlılık, tutarsızlık ve artan standart hatalar olduğu görülecektir (Miller, 2007).

Veride sıfır yığılma durumu olması durumunda;

- sıfırlar veriden silinerek,
- sıfır yığılma durumu yok sayılarak asimptotik normallik varsayımı ile ve en küçük kareler regresyonu gibi standart teknikleri kullanarak veriler analiz edilerek,
- sıfırlar üzerine dönüşüm uygulanarak ve
- sıfır yığılmalı regresyon modelleri kullanılarak

Veri analiz edilebilir.

Sıfır yığılma içeren veriler özellikle sağlık araştırmalarında, alkol kullanım sıklığı, içilen sigara sayısı, hastaneye yatış sayısı gibi belirli davranışsal olayların oluşumunu sayarken öne çıkmaktadır.

Sıfır yığılmalı modelleri kullanmanın en büyük avantajı, yalnızca aşırı yayılımdan kaynaklanan yanlılığı azaltmak değil aynı zamanda modelleme konusunda da oldukça iyi olmasıdır.

Sıfır yığılmalı verilerin analizinde kullanılan çeşitli teknikleri değerlendirmek için hem simüle edilmiş verilerden hem de ampirik verilerden yapılan çalışmalar farklı sonuçlar ortaya koymuştur. Bu çalışmalar sonucunda farklı model önerileri üretilmiştir. Modeller arasındaki farklar farklı derecede sıfır yığılma ve aşırı yayılımın anlaşılmamasından kaynaklanabilir. Bu nedenle, çok fazla sıfır içeren verilerle nasıl başa çıkılacağını anlamak için farklı koşullar altında bu yöntemlerin kapsamlı bir incelemesini ve karşılaştırmasını yapmak gerekir.

Sıfır yığılmalı veriler ile kurulan regresyon modelleri bağımlı değişkenin geldiği dağılıma göre adlandırılır. Örneğin bağımlı değişken Poisson dağılımdan geliyorsa ve veride sıfır

yığılma durumu söz konusu ise kullanılan regresyon modeli sıfır yığılmalı Poisson regresyon modeli olacaktır. Bu tez çalışması kapsamında sıfır yığılmalı Poisson regresyon, sıfır yığılmalı negatif binom regresyon ve sıfır yığılmalı Bell regresyon anlatılacaktır.

3.2.1. Sıfır yığılmalı Poisson regresyon modeli (zero inflated Poisson regression-ZIP)

Sıfır yığılmalı Poisson regresyon modeli (ZIP), ilk olarak Lambert (1992) tarafından önerilmiştir. Lachenbruch (2002), "ZIP regresyon modelini, 0'da yığılma gösteren değerlerin Poisson dağılımı ile karıştırılması (mixed)" olarak ifade etmektedir. ZIP, bir veri setinde bağımlı değişkenin Poisson dağılımından geldiği ve veri seti içerisinde beklenenden daha fazla sıfır olduğu durumlarda tercih edilir. ZIP regresyon modelinde yanıt değişkeninin iki bileşeni vardır. İlk bileşen, p_i olasılığı ile y_i'nin sıfıra eşit olmasını gösterir. İkinci bileşen ise, y_i'nin (1-p_i) olasılığı ile Poisson dağılımına sahip olmasını gösterir. Ayrıca bu bileşenlerden biri pozitif sayma değişkenini modellemektedir. Diğeri ise Logit bağlantı fonksiyonu ile veri setinde bulunan sıfırları modellemede kullanılır (Peng, 2013).

ZIP regresyonunun olasılık fonksiyonu Eş 3.13 de verilmiştir.

$$P(y_i|x_i) = \begin{cases} p_i + (1-p_i)e^{-\lambda_i}, y_i = 0 \\ (1-p_i)\frac{e^{-\lambda_i - \lambda_i y_i}}{y_i!}, y_i \neq 0 \end{cases} \quad (3.13)$$

Burada λ_i, Poisson dağılımına ilişkin ortalama olmak üzere $\lambda_i > 0$ ve $0 < p_i < 1$'dir. p_i değeri sıfıra eşit olduğunda ZIP regresyon modeli, Poisson regresyon modeline dönüşmektedir.

Ayrıca λ_i Eş. 3.14 de verildiği gibi ifade edilebilir.

$$\lambda_i = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m) = e^{x_i'\beta} \quad (3.14)$$

p_i, Eş. 3.15 de verildiği gibi ifade edilebilir.

$$p_i = \frac{\exp(\tau_0 + \tau_1 z_{1i} + \cdots + \tau_q z_{qi})}{1 + \exp(\tau_0 + \tau_1 z_{1i} + \cdots + \tau_q z_{qi})} \quad (3.15)$$

ZIP modeline ait bu gösterimde $(\beta_0, \beta_1, \dots, \beta_m)^T$ ve $\tau = (\tau_0, \tau_1, \dots, \tau_q)^T$ bağımsız kabul edilen bilinmeyen regresyon katsayılarının vektörleridir. p_i, logit bağlantı(link) fonksiyonudur. x_i ve z_i ise bilinen ortak değişkenler üzerindeki gözlemlerdir.

Dağılıma ilişkin beklenen değer ve varyans Eş. 3.16 da verilmiştir

$$E(Y_i) = (1 - p_i)\lambda_i$$
$$Var(Y_i) = (1 - p_i)\lambda_i(1 - p_i\lambda_i) \quad (3.16)$$

Dağılıma ilişkin log olabilirlik fonksiyonu ise Eş 3.17 de verilmiştir.

$$\log(L) = \sum_{i=1}^{n}[I(y_i = 0)\log(p_i + (1 - p_i)e^{-\lambda_i} + I(y_i > 0)(\log(1 - p_i) - \lambda_i +$$
$$y_i \log(\lambda_i) - \log(y_i!))] \quad (3.17)$$

3.2.2. Sıfır yığılmalı negatif binom regresyon modeli (zero inflated negative binom-ZINB)

Sıfır yığılmalı modellerde sayma verisine dayalı bileşeni modellemek için Poisson regresyonu, bazı durumlarda yeterli olmayabilir. ZIP regresyon modeli koşullu ortalamanın koşullu varyansa eşit olduğunu varsayar. Ancak veriler ZIP regresyon modelinde varsayıldığından daha büyük koşullu varyansa sahipse aykırı değerlerin varlığından kaynaklanabilecek aşırı yayılım meydana gelir (Feng ve ark, 2020). Böyle durumlarda sıfır yığılmalı negatif binom regresyon modeli uygulanması daha doğru olduğu bilinmektedir. ZINB regresyon modeli Greene (1994) tarafından negatif binom regresyonundan yararlanılarak geliştirilmiştir.

ZINB regresyon modeli, veri setinde çok sayıda sıfır olduğunda ve verilerde aşırı dağılım durumu olduğunda kullanılır. Bu regresyon modeli, Poisson regresyonunun açıklayamadığı aşırı dağılım ve sıfır değer ağırlığını modellemek için geliştirilmiş ve bağımlı değişkendeki heterojenliği açıklamaya uygun bir regresyon yöntemidir (Zhuo ve diğerleri, 2008). ZIP modelinde olduğu gibi, sıfır değerli gözlemler ve sıfır olmayan gözlemler ayrı ayrı modellenir. Ancak, ZIP'den farklı olarak, bu regresyon modelindeki sıfır olmayan gözlemler, negatif binom dağılımı ile modellenmiştir. ZINB regresyon modeli Eş 3.18 de gösterilmiştir.

$$P(y_i|x_i) = \begin{cases} p_i + (1-p_i)\left(\frac{k}{\lambda_i+k}\right), y_i = 0 \\ (1-p_i)\left(\frac{\Gamma(y_i+k)}{\Gamma(k)\Gamma(y_i+1)}\left(\frac{\lambda_i}{\lambda_i+k}\right)^{y_i}\left(\frac{k}{\lambda_i+k}\right)^{k}\right), y_i \neq 0 \end{cases} \quad (3.18)$$

Burada $\lambda_i > 0$ ve $0 < p_i < 1$'dir. k, sıfırdan küçük olmayan ve ortak değişkenlerden bağımsız aşırı yayılım parametresidir. k'nın 0 değerine yakınsaması durumunda ZINB regresyon modeli ZIP regresyon modeline dönüşür.

Ayrıca ZIP regresyon modelinde olduğu gibi λ_i Eş 3.14 de ve p_i 'de Eş 3.15 de verildiği gibi ifade edilebilir.

Dağılıma ilişkin beklenen değer ve varyans Eş 3.19 da verilmiştir.

$$E(Y_i) = (1-p_i)\lambda_i$$
$$Var(Y_i) = (1-p_i)\lambda_i(1 + k\lambda_i + p_i\lambda_i) \quad (3.19)$$

Dağılıma ilişkin log olabilirlik fonksiyonu ise Eş 3.20 de verilmiştir.

$$\log(L) = \sum_{i=1}^{n}[I(y_i = 0)\log\left(p_i + (1-p_i)\left(\frac{1}{1+k\lambda_i}\right)^{\frac{1}{k}} + I(y_i > 0)\right(\log(1-p_i) +$$
$$\log\left(\frac{\Gamma\left(y_i+\frac{1}{k}\right)}{\Gamma(y_i+1)\Gamma\left(\frac{1}{k}\right)}\right) - \left(y_i + \frac{1}{k}\right)\log(1+k\lambda_i) + y_i\log(k\lambda_i)\Big)] \quad (3.20)$$

3.3. Bell Dağılımlı Regresyon Modelleri

Sayma verilerine, bir takvim yılındaki gök gürültülü fırtınaların meydana gelme sayısı, kaza sayısı, devamsızlık sayısı, iş kazası nedeniyle kaybedilen gün sayısı, sigorta tazminatı sayısı gibi birçok günlük hayat probleminde rastlamak mümkündür. Poisson dağılımı, basitliği nedeniyle pratikte kullanılan sayma verileri için en popüler modeldir. Bu dağılımın en büyük dezavantajı, varyansın ortalamaya eşit olacak şekilde sınırlandırılmasıdır. Bu nedenle, sayma verilerinin modellenmesinde alternatif olarak kesikli dağılımların kullanılabilirliği literatürde önerilmiştir. Yeni kesikli dağılımların çoğu bilinen bir sürekli dağılımın farklılaştırılmasıyla elde edilmektedir.

Yapılan tez çalışmasının bu bölümünde, Bell dağılımı adı verilen ve bir dizi açılım temeline dayanan oldukça yeni tek parametreli kesikli dağılım ve bu dağılımın sıfır sayısı fazla olduğunda kullanılmak üzere önerilen sıfır yığılmalı Bell dağılımı tanıtılacaktır.

3.3.1. Bell regresyon

Bell regresyon yanıt değişkenin sayılabilir ve tek parametreli Bell dağılımından geldiği durumda kullanılır. Bell dağılımına ilişkin bazı özellikler şu şekildedir:

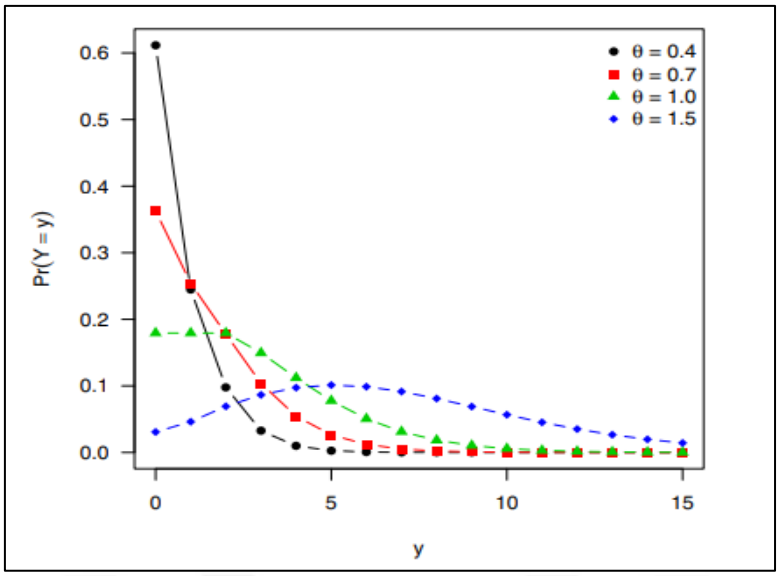
- Tek parametreli bir dağılımdır.
- Tek parametreli üstel dağılım ailesine aittir.
- Parametrenin küçük değerleri için Bell dağılımı Poisson dağılımına yaklaşır.
- Sonsuz bölünebilir.
- Çoklu Poisson sürecinin özel bir durumudur.
- Tüm $\theta > 0$ için tanımlanabilir (burada θ, dağılıma ilişkin parametredir).
- Tek modludur.

Bell dağılımı aslında çoklu Poisson sürecinin (multiple Poisson process) özel bir durumudur. Bu yöntem karmaşık çıkarımlar içermediğinden pratik ve uygulanabilirdir.

Tek parametreli Bell dağılımına ilişkin olasılık yoğunluk fonksiyonu Eş 3.21 de verilmiştir.

$$P(Y = y) = \frac{\theta^y e^{-e^\theta + 1} B_y}{y!}, y = 0,1,2, \ldots. \quad (3.21)$$

Burada $\theta > 0$ ve B_y Bell sayılarını ifade etmektedir. Farklı θ değerleri için Bell dağılımına ait grafik Şekil 3.3 de verilmiştir.



Şekil 3.3. Farklı θ değerleri altında Bell dağılımı

Bell sayılarının elde edilmesinde Eş 3.22 de verilen Bell tarafından önerilen eşitlik kullanılmaktadır (Bell,1934).

$$\exp(e^x - 1) = \sum_{n=0}^{\infty} \frac{B_n}{n!} x^n, x \in R \tag{3.22}$$

Eş 3.22 de yer alan B_n, Bell sayılarıdır. Bu sayılar Eş 3.23 de gösterilen formül ile hesaplanır.

$$B_n = \frac{1}{e} \sum_{k=0}^{\infty} \frac{k^n}{k!} \tag{3.23}$$

İlk Bell sayısı $B_0 = B_1 = 1$' dir. Diğer Bell sayılarına örnek olarak $B_2 = 2, B_3 = 5, B_4 = 15, B_5 = 52$ verilebilir. B_n, Bell sayısı parametresi 1'e eşit olan Poisson dağılımının n'inci momentidir.

Y parametresi $\theta > 0$ olan bir Bell dağılımından geldiği bilindiğinde kullanılan gösterim Y~Bell(θ) şeklindedir. Eş 3.22 ile gösterilen eşitlik basittir ve herhangi bir karmaşık fonksiyon içermez. Bu nedenle önerilen bu yeni dağılım araştırmacıya kolaylık sağlayacaktır.

Bell dağılımına ilişkin olasılık yoğunluk fonksiyonu Lambert fonksiyonu ilişkisi kullanarak gösterilebilir.

$$xe^{x} = a \tag{3.24}$$

Eş 3.24 de verilen eşitlik Euler ve Lambert tarafından incelenmiştir. Bu eşitliğin sol tarafta verilen fonksiyonun tersi Lambert fonksiyonu olarak ifade edilmektedir. Bu fonksiyon *W* ile gösterilmektedir.

$\mu = \theta e^{\theta}$ olmak üzere burada $\theta = W_0(\mu)$ şeklinde gösterilebilirken W_0 Lambert fonksiyonunu ifade etmektedir. (Castellares,2018).

Bell dağılımına ilişkin olasılık yoğunluk fonksiyonu Lambert fonksiyonu kullanılarak ifade edilirse yeni oluşan olasılık fonksiyonu Eş 3.25 de verilmiştir.

$$P(Y) = \exp\left(1 - e^{W_0(\mu)}\right)\frac{W_0(\mu)^{y}B_y}{y!}, y = 0,1,2, \ldots \tag{3.25}$$

Beklenen değer ve varyans ise Eş 3.26 da verilmiştir.

$$E(Y) = \theta e^{\theta}$$
$$Var(Y) = \theta(1+\theta)e^{\theta} \tag{3.26}$$

Buradan açıkça görüldüğü üzere varyans beklenen değerden büyüktür yani aşırı yayılım durumunda kullanıma uygundur.

Bell dağılımına ilişkin detaylı çıkarımlar için Castellares ve ark.(2018) yaptıkları çalışma okunabilir.

Bell dağılımı için parametre tahmini incelendiğinde gözlemlenen $y_1, y_2, \ldots, y_n$ değerlerine ilişkin log olabilirlik fonksiyonu Eş 3.27 de verilmiştir.

$$l(\theta) = -ne^{\theta} + \log(\theta)\sum_{i=1}^{n} y_i \tag{3.27}$$

En çok olabilirlik tahmin edicisi ise:

$$-e^{\hat{\theta}} + \frac{\bar{Y}}{\hat{\theta}} = 0$$

şeklindedir. Bu ifade $\bar{Y} = \hat{\theta}\exp(\hat{\theta})$ ifadesine eşittir.

Bell dağılımı tek parametreli ve üstel ailenin özelliklerine sahip olduğundan skor fonksiyonu, Fisher bilgi matrisi ve log olabilirlik fonksiyonlarını elde etmek daha basittir. Bu nedenle Poisson ve negatif binom regresyon modellerine alternatif bir modeldir. Yanıt değişkeni sayılardan oluştuğunda ve fazla dağınık olduğunda, Poisson modellemesi yerine Bell dağılımına dayalı modelleme tercih edilir (Amin ve ark., 2020). NB dağılımına göre Bell dağılımının bir avantajı, aşırı dağılıma uyum sağlamak için ek (dağılım) parametresinin gerekli olmamasıdır.

Bir dağılımın regresyon modelinin oluşturulması için, yanıt değişkeninin ortalamasını modellemek uygundur (Akdur, 2022). Bell dağılımının ortalamasına ilişkin bir regresyon yapısı elde etmek için, Bell olasılık yoğunluk fonksiyonu üzerinde yeniden parametrelendirme işlemi yapılır. Bu yeniden parametrelendirme işlemi yukarıda da verildiği gibi $\mu = \theta e^{\theta}$ eşitliğinde θ, $\theta = W_0(\mu)$şeklinde yazılması ile yapılır. Regresyon modeli için beklenen değer ve varyansı Eş 3.28 de verilmiştir.

$$E(Y) = \mu$$
$$Var(Y) = \mu[1 + W_0(\mu)] \quad (3.28)$$

Burada μ >0 olmak üzere yanıt değişkeni olan Y'nin ortalamasıdır.

Bell olasılık fonksiyonu, bahsedilen yeni parametreleme elde edilmiş hali Eş 3.25 de verilmiştir.

$Y_1, Y_2, \dots, Y_n$ n bağımsız rastgele değişken ve ortalama μ_i olmak üzere ($Y_i \sim Bell(W_0\ (\mu_i))$ ortalama için bağlantı (link) fonksiyonu:

$$g(\mu_i) = x_i^T \beta, i = 1,2,, \dots, n.$$

Şeklindedir. Burada $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T \in R^p$, regresyon katsayılarının ($p<n$) p boyutlu bir vektörü ve $x_i^T = (x_{i1}, x_{i2}, \ldots, x_{ip})$ bilinen p bağımsız değişkenler üzerindeki gözlemleri belirtir.

3.3.2. Sıfır yığılmalı Bell regresyon (Zero inflated Bell regression-ZIBell)

ZIBell regresyon modeli, Lemonte ve ark. (2020) tarafından istatistiksel çıkarımlarda kolaylık sağlayan ve sayma verilerinde klasik sıfır yığılmalı regresyon modellerine bir alternatif olarak önerilmiştir. Oldukça yeni bir regresyon modelidir.

Başlık 3.2 de bahsedilen sıfır yığılmalı regresyon modelleri ile aynı özellikler ZIBell regresyon modeli içinde geçerlidir. ZIBell regresyon modeli de benzer şekilde tıp, halk sağlığı, çevre bilimleri, tarım ve imalat uygulamaları dahil olmak üzere birçok disiplinde kullanılabilir.

ZIBell dağılımı, sıfırdaki dejenere bir dağılımı yukarıda bahsedilen Bell dağılımıyla karıştırarak oluşturulmuştur ve Eş 3.29 da verilmiştir.

$$P(Y=y) = \begin{cases} p_i + (1-p_i)\exp\left(1-e^{W(\mu)}\right), & y=0 \\ (1-p_i)\exp\left(1-e^{W(\mu)}\right)\frac{W(\mu)^y B_y}{y!}, & y>0 \end{cases} \quad (3.29)$$

Burada μ >0 ve $p_i \in (0, 1)$' dir. Y ~ ZIBell(μ, p_i)'nin ortalaması ve varyansı Eş 3.30 da gösterilmiştir.

$$E(Y) = (1-p_i)\mu,$$
$$Var(Y) = (1-p_i)\mu[1 + W(\mu) + \mu p_i] \quad (3.30)$$

ZIBell dağılımına dayalı olarak önerilen regresyon modeli daha öncede bahsedildiği gibi oldukça yenidir. Sıfır yığılma durumu görüldükten sonra eğer verilerde aşırı yayılım da söz konusu ise ZIBell modeli kullanılabilir. Diğer sıfır yığılmalı regresyon modellerinin kurulumuna benzer şekilde μ ve p_i parametreleri bağımsız değişkenler ile ilişkilendirilir. Ayrıca, ZIBell regresyonu ile ilgili bazı niceliklerin (örneğin, skor fonksiyonu, Fisher bilgi matrisi, vb.) elde edilmesi basittir, bu da uygulamada kolaylık sağlamaktadır.

$Y_1, Y_2, \dots, Y_n$ bağımlı değişkeni n bağımsız rastgele değişken olsun. Her bir Y_i'ye ait olasılık dağılımı ZIBell dağılımından gelsin. Regresyon modeli için kullanılan link fonksiyonları Eş 3.31 de verilmiştir.

$$\log(\mu) = x_i^T \beta = \eta_{1i}, \log\left(\frac{p_i}{1-p_i}\right) = \eta_{2i} = s_i^T \tau \tag{3.31}$$

burada $\beta = (\beta_1, \beta_2, \dots, \beta_p)^T \in R^p$ ve $\tau = (\tau_1, \tau_2, \dots, \tau_q)^T$ bilinmeyen regresyon katsayılarının vektörleri ve $x_i^T = (x_{i1}, \dots, x_{ip})$ ve $s_i^T = (s_{i1}, \dots, s_{iq})$ bağımsız değişkenlerdir.

Modele ilişkin log olabilirlik fonksiyonu Eş 3.32 dedir.

$$l(\theta) = \sum_{y_i=0} log\,[e^{\eta_{2i}} + \exp(1 - e^{W(\mu_i)})] - \sum_{i=1}^{n} \log(1 - e^{\eta_{2i}}) + \sum_{y_i>0} y_i \log[W(\mu_i)] - \sum_{y_i>0} e^{W(\mu_i)} \tag{3.32}$$

Burada $\mu_i = e^{\eta_{1i}} = \exp\,(x_i^T \beta)'$dir.

4. ARTIKLARIN İNCELENMESİ

Artıkların incelenmesi, modeller ve veriler arasındaki tutarsızlıkları belirlemek ve model uyum iyiliğini değerlendirmek için istatistiksel analizlerde kullanılan önemli bir adımdır. Artıklar modelin yanlış tanımlamasını tespit etmek için, aykırı değerleri veya etkin gözlemleri belirlemek için kullanılabilirler. 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, yanıt değişkeni kesikli olduğunda, bu artıklar normallikten uzaklaşır ve farklı kesikli yanıt değerlerine göre neredeyse paralel eğrilere sahiptir, bu da görsel inceleme için büyük zorluklar doğurur.

Artıkları incelemek, modeller ve veriler (örneğin, doğrusal olmayan etkiler, aşırı dağılım, sıfır yığılma) arasındaki genel tutarsızlıkları ve modellere uymayan gözlemleri (örneğin, aykırı değerler) belirlemek için kullanılan bir yaklaşımdır. Artıklar, bir modelin genel uyum iyiliği hakkında bilgi verir.

Artıkları kullanmanın belki de en verimli yolu, artıkları diğer ilgili değişkenlere karşı çizmektir. Bu tür grafiklere örnek olarak bağımlı değişkenin değerleri ile tahmin edilen değerleri arasındaki uyumu gösteren grafikler verilebilir.

Ayrıca, 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 tanımı yapmak mümkün değildir. Poisson ve diğer genelleştirilmiş lineer modeller (generalised linear model- GLM) için birkaç artık önerilmiştir. Bu tez çalışmasında Pearson artıklar, sapma (deviance) artıklar ve rastgele kantil artıklar (randomize quantile residuals-RQR) dan bahsedilecektir. RQR kullanımı bilindiği kadarıyla yeni bir yöntemdir ve Bell ile sıfır yığılmalı Bell regresyon modellerinde henüz hiç kullanılmamıştır. Bu nedenle bu tez çalışmasının bu alanda ilk olacağı düşünülmektedir.

4.1. Pearson Artıklar

Model uyum iyiliğinin incelenmesi, uygun regresyon modelinin geçerli olduğundan emin olmak için bir regresyon modeli oluşturmada önemli bir adımdır. Pearson Artıklar, gözlenen ve tahmin edilen sıklık arasındaki farkın standart hatasına oranı olarak tanımlanmıştır. Gözlenen ve tahmin edilen değerler arasındaki göreli sapmayı ölçer. Model uyumluluğunu ölçmede sıklıkla kullanılır. Bu artık türünü yanıt değişkeninin tahmini standart sapması ile ölçeklenen klasik artıklar olarak tanımlamak mümkündür. Genel gösterimi Eş 4.1 de verilmiştir.

$$r_i^p = \frac{y_i - \widehat{\mu_i}}{\sqrt{\widehat{V(y_i)}}} \tag{4.1}$$

Burada $\widehat{\mu_i}$ ler y_i nin uydurulmuş (fit edilmiş) değerleridir. $\widehat{Var(y_i)}$ ise y_i nin varyansının tahminidir.

Çizelge 4.1 de bu tez çalışmasında kullanılan regresyon modellerine ilişkin artıklar verilmiştir.

Çizelge 4.1. Kullanılan regresyon modellerine göre Pearson artıklar

Model	Pearson Artıklar
Poisson	$r_i^p = \frac{y_i - \hat{\lambda}}{\sqrt{\widehat{\lambda_i}}}$
NB	$r_i^p = \frac{y_i - \hat{\lambda}}{\sqrt{\widehat{\lambda_i} + \widehat{\lambda_i^2}/k}}$
Bell	$r_i^p = \frac{y_i - \hat{\mu}}{\sqrt{\hat{\mu}[1 + W_0(\hat{\mu})}}$
ZIP	$r_i^p = \frac{y_i - (1 - \widehat{p_i})\widehat{\lambda_i}}{\sqrt{(1 - \widehat{p_i})\widehat{\lambda_i}[1 + \widehat{p_i}\widehat{\lambda_i}]}}$
ZINB	$r_i^p = \frac{y_i - (1 - \widehat{p_i})\widehat{\lambda_i}}{\sqrt{(1 - \widehat{p_i})\left(\widehat{\lambda_i} + \frac{\widehat{\lambda_i^2}}{k}\right) + \widehat{\lambda_i}^2(\widehat{p_i^2} + \widehat{p_i})}}$
ZIBell	$r_i^p = \frac{y_i - (1 - \widehat{p_i})\hat{\mu}}{\sqrt{(1 - \widehat{p_i})\hat{\mu}[1 + W(\hat{\mu}) + \hat{\mu}\widehat{p_i}]}}$

4.2. Sapma (Deviance) Artıklar

Sapma artıklar, model sapmasına bireysel katkıların işaretli karekökü olarak tanımlanır (yani, doymuş ve uygun modellerin log-olasılıkları arasındaki farktır). Sapmaların farkları üzerine oluşturulmuştur. Sapma artıkları, doymuş (saturated) ve uygun modeller için log-olabilirlik arasındaki farkın iki katı olarak tanımlanan ve verilen sapmaya dayalı olarak türetilen bir artık türüdür (Feng ve ark., 2020). Tanım olarak doymuş bir model, verilere mükemmel bir uyum sağlar ve tüm modeller arasında en yüksek log olasılığına sahiptir.

Bu artık türü için genel gösterim Eş 4.2 de verilmiştir.

$$D(y,\hat{\mu}) = 2\sum_{i=1}^{n}\{\log\left[p\left(y_i|\widehat{\theta_s}\right)\right] - \log\left[p\left(y_i|\hat{\theta}\right)\right]\} \quad (4.2)$$

Burada $\log\left[p\left(y_i|\widehat{\theta_s}\right)\right]$ ifadesi doymuş model için log olabilirlik fonksiyonudur. $D(y,\hat{\mu})$ ifadesi için farklı bir gösterim Eş 4.3 de verilmiştir.

$$r_i^D = sign(y_i - \hat{\mu}_i)\sqrt{d_i} \quad (4.3)$$

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

Bir ZIP modeli için, Poisson(y_i)'nin doymuş model olduğu gösterilebilir; bu nedenle, ZIP modeli için sapma artığı, uygun model (sıfır yığılmalı Poisson modeli) ve doymuş model (Poisson modeli) arasındaki olabilirlik oranının karekökü olarak tanımlanır. ZINB ve ZIBELL modelleri için de sapma artıkları benzer şekilde tanımlanabilir.

Çizelge 4.2 de bu tez çalışmasında kullanılan regresyon modellerine ilişkin sapma artıklar verilmiştir.

Çizelge 4.2. Kullanılan regresyon modellerine göre sapma artıklar

Model	Sapma Artıklar
Poisson	$r_i^D = sign(y_i - \widehat{\lambda_\iota})\{2\left[y_i log\frac{y_i}{\widehat{\lambda_\iota}} - (y_i - \widehat{\lambda_\iota})\right]\}^{1/2}$
NB	$r_i^D = sign(y_i - \widehat{\lambda_\iota})\{2\left[y_i log\frac{y_i}{\widehat{\lambda_\iota}} - (y_i + k)log\frac{y_i + k}{\widehat{\lambda_\iota} + k}\right]\}^{1/2}$
Bell	$r_i^D = sign(y_i - \widehat{\mu_\iota})\{2\begin{cases}\exp(1 - e^{W_0(\widehat{\mu_\iota})}) & , y_i = 0\\ e^{W_0(\widehat{\mu_\iota})} - e^{W_0(y_i)} + y_i \log\left(\frac{W_0(y_i)}{W_0(\widehat{\mu_\iota})}\right), & y_i > 0\end{cases}$
ZIP	$r_i^D = sign(y_i - \widehat{\lambda_\iota})(2\{-y_i + y_i logy_i - \log y_i! - I(y_i = 0)\log[\widehat{p_\iota} + (1 - \widehat{p_\iota})e^{-\widehat{\lambda_\iota}}]$ $- I(y_i > 0)[\log(1 - \widehat{p_\iota}) - \widehat{\lambda_\iota} + y_i log\widehat{\lambda_\iota} - logy_i!]\})^{1/2}$
ZINB	$r_i^D = sign(y_i - \widehat{\lambda_\iota})(2\left\{\log\frac{\Gamma(y_i + k)}{\Gamma(k)\Gamma(y_i + 1)} + y_i \log\left(\frac{y_i}{y_i + k}\right) + klog(\frac{k}{y_i + k})\right.$ $- I(y_i = 0)\log\left[\widehat{p_\iota} + (1 - \widehat{p_\iota})(\frac{k}{\widehat{\lambda_\iota} + k})^k\right]$ $- I(y_i > 0)\left[\log(1 - \widehat{p_\iota}) + \log\frac{\Gamma(y_i + k)}{\Gamma(k)\Gamma(y_i + 1)} + y_i \log\left(\frac{y_i}{y_i + k}\right)\right.$ $\left.\left. + k\log\left(\frac{y_i}{y_i + k}\right)\right]\right\})^{1/2}$
ZIBell	$r_i^D = sign(y_i - \widehat{\mu_\iota})\{2\begin{cases}e^{W_0(\widehat{\mu_\iota})} - e^{W_0(\widehat{y_\iota})} & , y_i = 0\\ e^{W_0(\widehat{\mu_\iota})} + y_i \log\left(\frac{W_0(y_i)}{W_0(\widehat{\mu_\iota})}\right) - e^{W_0(\widehat{y_\iota})}, & y_i > 0\end{cases}$

Çizelge 4.2 de verilen sapma artıkların elde edilişleri ile ilgili olarak yapılan ispatlar aşağıda verilmiştir.

İspatlar:

Poisson regresyon modeli için Deviance Artıkların Elde edilmesi:

Klasik Poisson model için olabilirlik fonksiyonunun elde edilmesi:

Eş 4.4 de Poisson modeline ait log olabilirlik fonksiyonu verilmiştir.

$$\ln L\,(\lambda; Y_1, \dots, Y_n) = \frac{e^{-\sum_{i=1}^n \lambda}\lambda^{\sum_{i=1}^n Y_i}}{\prod_{i=1}^n \log\,(y_i!)} = -\sum_{i=1}^{n}\lambda + \sum_{i=1}^{n} Y_i\, log\lambda - log\prod_{i=1}^{n}\log\,(y_i!)$$

$$\frac{d}{d\lambda}\ln L(\lambda; Y_1, \dots, Y_n) = -n + \frac{\sum Y_i}{\lambda} = 0 \qquad (4.4)$$

Buradan $\hat{\lambda} = \bar{y}$ olarak bulunur.

Doymuş (saturated) model için olabilirlik fonksiyonunun elde edilmesi:

Eş 4.5 de Poisson için saturated modele ait log olabilirlik fonksiyonu verilmiştir.

$$l(\lambda; Y) = -\sum_{i=1}^{n} \lambda_i + \sum_{i=1}^{n} Y_i \, log\lambda_i - log \prod_{i=1}^{n} \log (y_i!)$$

$$\frac{d}{d\lambda} \ln L(\lambda; Y_1, \dots, Y_n) = -1 + \frac{y_i}{\lambda_i} = 0 \qquad (4.5)$$

Buradan $\hat{\lambda} = \overline{y_i}$ olarak bulunur. Doymuş modelde log olabilirlik elde etmek için tahmin edici yerine konur ve

$$l(\hat{\lambda}; Y) = -\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} Y_i \, logy_i - log \prod_{i=1}^{n} \log (y!)$$

Elde edilir.

Buradan sapma artıkların Eş.4.2 de verilen formülü kullanılarak Eş 4.6 elde edilir.

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} Y_i \, logy_i - log \prod_{i=1}^{n} \log (y!) - \; y_i log\hat{\lambda} + \sum_{i=1}^{n} \hat{\lambda} + log \prod_{i=1}^{n} \log (y!) \quad (4.6)$$

Eş 4.6 da gerekli matematiksel işlemler yapıldığında

$$r_i^D = sign(y_i - \widehat{\lambda_i})\{2\left[y_i log\frac{y_i}{\widehat{\lambda_i}} - (y_i - \widehat{\lambda_i})\right]\}^{1/2}$$

Elde edilmiş olur.

NB regresyon modeli için Deviance Artıkların Elde edilmesi:

Negatif binom dağılımına ilişkin olasılık fonksiyonu:

$$P(Y_i, \lambda_i, \theta) = \frac{\Gamma(y_i + \theta)}{\Gamma(\theta) y_i!}\left(\frac{\theta}{y_i + \theta}\right)^{\theta} (\frac{\lambda_i}{\lambda_i + \theta})^{y_i}$$

Verildiği gibidir. Bu olasılık fonksiyonuna ilişkin Log olabilirlik fonksiyonu Eş 4.7 de verilmiştir.

$$L = \prod_{i=1}^{n}[\frac{\Gamma(y_i+\theta)}{\Gamma(\theta)y_i!}\left(\frac{\theta}{y_i+\theta}\right)^{\theta}(\frac{\lambda_i}{\lambda_i+\theta})^{y_i}] \quad 4.7.$$

$$\log L = \sum_{i=1}^{n}\{y_i \log\left(\frac{\lambda_i}{\lambda_i+\theta}\right) + \theta \log\left(\frac{\theta}{\lambda_i+\theta}\right) + \log\big(\Gamma(y_i+\theta)\big) - log\Gamma(\theta) - \log\Gamma(y_i+1)\}$$

$\frac{dlogL}{d_\lambda} = 0$ işlemi yapıldığında $\hat{\lambda} = y_i$ elde edilir.

Doymuş (saturated) model için olabilirlik fonksiyonunun elde edilmesi:

Eş 4.8 de doymuş modele ait log olabilirlik fonksiyonu verilmiştir.

$$y_i \log\left(\frac{y_i}{y_i+\theta}\right) + \theta \log\left(\frac{\theta}{y_i+\theta}\right) + \log\big(\Gamma(y_i+\theta)\big) - log\Gamma(\theta) - log\Gamma(y_i+1) \quad (4.8)$$

Deviance artıklar için Eş.4.2 de verilen formül kullanıldığında Eş 4.9 elde edilmiş olur.

$$r_i^D = sign\big(y_i - \widehat{\lambda_i}\big)\{2\left[y_i log\frac{y_i}{\widehat{\lambda_i}} - (y_i+k)log\frac{y_i+k}{\widehat{\lambda_i}+k}\right]\}^{1/2} \quad (4.9)$$

Bell regresyon modeli için Deviance Artıkların Elde edilmesi:

Bell dağılımına ilişkin olasılık fonksiyonu:

$$P(Y) = \exp\big(1 - e^{W_0(\mu)}\big)\frac{W_0(\mu)^y B_y}{y!}, y = 0,1,2,\ldots$$

Olarak ele alınsın.

Olabilirlik fonksiyonunun elde edilmesi:

$\log(y_i, \mu_i) = Y \cdot \log\big(W_0(\mu)\big) - e^{W_0(\mu)}$

$\frac{d(Y_i,\mu_i)}{d\mu_i} = 0$ işlemi sonucu $Y = W_0(\hat{\mu})e^{W_0(\hat{\mu})}$ elde edilir.

Doymuş (saturated) model için olabilirlik fonksiyonunun elde edilmesi:

$$l(\hat{\mu}; y) = W_0(\mu)e^{W_0(\mu)} \log\big(W_0(\mu)\big) - \sum_{i=1}^{n} e^{W_0(\mu_i)}$$

$$= e^{W_0(\mu)}[W_0(\mu) \log\big(W_0(\mu)\big) - 1]$$

Elde edilir.

Deviance artıklar için Eş.4.2 de verilen formül kullanıldığında Eş 4.10 elde edilmiş olur.

$$r_i^D = sign(y_i - \widehat{\mu_i})\{2 \begin{cases} \exp\big(1 - e^{W_0(\widehat{\mu_i})}\big) & , y_i = 0 \\ e^{W_0(\widehat{\mu_i})} - e^{W_0(y_i)} + y_i \log\left(\frac{W_0(y_i)}{W_0(\widehat{\mu_i})}\right), & y_i > 0 \end{cases} \quad (4.10)$$

Sıfır Yığılmalı Poisson regresyon modeli için Deviance Artıkların Elde edilmesi:

Sıfır yığılmalı modellerde doymuş model bölümü klasik modelin doymuş modeli ile aynıdır. Örneğin ZIP model için doymuş model kısmı Poisson dağılımından gelmektedir.

Sıfır yığılmalı Poisson için dağılım fonksiyonu:

$$P(y_i|x_i) = \begin{cases} p_i + (1 - p_i)e^{-\lambda_i}, y_i = 0 \\ (1 - p_i)\dfrac{e^{-\lambda_i} - \lambda_i{}^{y_i}}{y_i!}, y_i \neq 0 \end{cases}$$

ZIP modele ilişkin log olabilirlik fonksiyonu Eş 4.11 de verilmiştir.

$\log(L) = \sum_{i=1}^{n}[I_{y_i=0}\log\,(p_i + (1 - p_i)e^{-\lambda_i} + I_{y_i>0}\log(1 - p_i) - \lambda_i + y_i log\lambda_i - logy_i!]$
(4.11)

Deviance artıklar için Eş.4.2 de verilen formül kullanıldığında Eş 4.12 elde edilir.

$$r_i^D = sign(y_i - \hat{\mu}_i)(2\{-y_i + y_i log y_i - \log y_i! - I(y_i = 0)\log[\hat{p}_i + (1-\hat{p}_i)e^{-\hat{\lambda}_i}] - I(y_i > 0)[\log(1-\hat{p}_i) - \hat{\lambda}_i + y_i log\hat{\lambda}_i - log y_i!]\})^{1/2} \quad (4.12)$$

Sıfır yığılmalı negatif binom regresyon modeli için deviance artıkların elde edilmesi:

ZINB dağılımına ilişkin olasılık fonksiyonu:

$$P(Y=y) = \begin{cases} p_i + (1-p_i)\left(\frac{\theta}{\lambda_i+\theta}\right), y_i = 0 \\ (1-p_i)\left(\frac{\Gamma(y_i+\theta)}{\Gamma(\theta)\Gamma(y_i+1)}\left(\frac{\lambda_i}{\lambda_i+\theta}\right)^{y_i}\left(\frac{\theta}{\lambda_i+\theta}\right)^{\theta}\right), y_i \neq 0 \end{cases}$$

Şeklindedir.

ZINB Model için Olabilirlik fonksiyonunun elde edilmesi:

$$\log L = \sum_{i=1}^{n}[\log\{(1-p_i)\frac{\Gamma(y_i+\theta)}{\Gamma(\theta)\Gamma(y_i+1)}\left(\frac{\lambda_i}{\lambda_i+\theta}\right)^{y_i}\left(\frac{\theta}{\lambda_i+\theta}\right)^{\theta}\}$$

$$= \sum_{i=1}^{n}\{I_{y_i=0}\log(p_i + (1-p_i)(\frac{\theta}{\lambda_i+\theta})^{\theta} + I_{y_i>0}\log\{(1-p_i) + y_i\log\left(\frac{\lambda_i}{\lambda_i+\theta}\right) + \theta log\left(\frac{\theta}{\lambda_i+\theta}\right) + log\Gamma(y_i+\theta) - log\Gamma(\theta) - log\Gamma(y_i+1)\}$$

Deviance artıklar için Eş.4.2 de verilen formül kullanıldığında Eş 4.13 elde edilir.

$$r_i^D = sign(y_i - \hat{\lambda}_i)(2\left\{\log\frac{\Gamma(y_i+k)}{\Gamma(k)\Gamma(y_i+1)} + y_i\log\left(\frac{y_i}{y_i+k}\right) + klog(\frac{k}{y_i+k}) - I(y_i=0)\log\left[\hat{p}_i + (1-\hat{p}_i)(\frac{k}{\hat{\lambda}_i+k})^k\right] - I(y_i>0)\left[\log(1-\hat{p}_i) + \log\frac{\Gamma(y_i+k)}{\Gamma(k)\Gamma(y_i+1)} + y_i\log\left(\frac{y_i}{y_i+k}\right) + k\log\left(\frac{y_i}{y_i+k}\right)\right]\right\})^{1/2} \quad (4.13)$$

Sıfır yığılmalı bell regresyon modeli için deviance artıkların elde edilmesi:

Sıfır yığılmalı Bell dağılımına ilişkin olasılık fonksiyonu:

$$P(Y=y)=\begin{cases} p_i+(1-p_i)\exp\left(1-e^{W(\mu)}\right), & y=0 \\ (1-p_i)\exp\left(1-e^{W(\mu)}\right)\dfrac{W(\mu)^y B_y}{y!}, & y>0 \end{cases}$$

ZIBell modele ilişkin log olabilirlik fonksiyonu:

$$logL(\mu)=\sum_{i=1}^{n}[I_{y_i=0}\log\left(p_i+(1-p_i)\exp\left(1-e^{W_0(\mu_i)}\right)+I_{y_i>0}\log(1-p_i)\right.$$
$$\left.+\left(1-e^{W_0(\mu_i)}\right)+y_i\log\left(W_0(\mu_i)\right)\right]$$

Şeklindedir. Deviance artıklar için Eş.4.2 de verilen formül kullanıldığında Eş 4.14 elde edilir.

$$r_i^D=sign(y_i-\hat{\mu}_i)\{2\begin{cases} e^{W_0(\hat{\mu}_i)}-e^{W_0(\hat{y}_i)} & ,y_i=0 \\ e^{W_0(\hat{\mu}_i)}+y_i\log\left(\frac{W_0(y_i)}{W_0(\hat{\mu}_i)}\right)-e^{W_0(\hat{y}_i)} & ,y_i>0 \end{cases} \tag{4.14}$$

4.3. Pearson ve Sapma (Deviance) Artıklar ile İlgili Sorunlar

Normal doğrusal regresyon modeli için, Pearson ve sapma artıkları özdeştir ve yaklaşık olarak normal dağılıma sahiptir. Bununla birlikte, bu artıklar sayma regresyon modelleri için normal dağılım göstermez ve dağılımları genellikle çarpıktır.

Sapma artıklarının, tipik olarak, Pearson artıklarına göre daha yakın bir normal dağılım izlediği ileri sürülmektedir (Feng ve ark 2020). Yanıt değişkeninin normale yakınsamasından dolayı üstel ailelerde hem Pearson hem de sapma artıkları normal dağılıma yaklaşır. Bununla birlikte, asimptotik normal dağılım, yalnızca yanıt değişkeninin ortalaması nispeten büyük olduğunda geçerlidir. Ayrıca, artık grafikler genellikle farklı yanıt değerlerine göre paralel eğriler sergileyerek görsel inceleme için büyük zorluklar ortaya çıkarır. Bu nedenle, sayma regresyon modellerinin uyum iyiliğinin grafiksel olarak değerlendirmesi için Pearson ve sapma artıklarının kullanılması zordur.

Artıkları incelemek, model uyumunu kontrol etmenin anahtar yoludur. Ancak, artıklar bazı durumlarda normal olarak dağılmamaktadır. Teoride, sapma artıkların Pearson artıklardan

normal dağılıma daha fazla yakınsadığı varsayılır. Bununla birlikte, yayılımın ortalamaya bölünmesiyle elde edilen (φ/μi) değeri büyük olduğunda artıkların hiçbiri normal dağılmaz. Yani, normal bir lineer model için artıklar normal olarak dağılır ve eşit varyanslara sahiptir. Ancak normal olmayan modellerde Pearson ve sapma artıkları gibi artıklar normal dağılmaz. Bunun nedeni, yanıt değişkeninin kesikli olması ve genellikle birkaç farklı değere sahip olmasıdır. Örneğin, Poisson verilerinin ortalaması sıfıra yakın olduğunda, Pearson ve sapma artıkları, farklı yanıt değerlerine karşılık gelen neredeyse paralel eğriler oluşturabilir. Bu nedenle grafikler istenildiği kadar belirleyici olmamaktadır (Bai, W., 2018).

Pearson ve sapma artıkları genellikle genelleştirilmiş doğrusal modellerin belirlenmesinde kullanılmıştır. Pearson artıkları, gözlemlenen ve beklenen yanıtlar arasındaki standartlaştırılmış mesafeler olarak tanımlanır ve sapma artıkları, model sapmasına bireysel katkıların işaretli karekökü olarak tanımlanır (yani, doymuş ve uygun modellerin log-olasılıkları arasındaki fark). Lineer regresyonda, Pearson ve sapma artıkları aynıdır ve normal dağılır. Model uyumunu değerlendirmek için, bu artıklar, standart normal dağılımla karşılaştırılmasının yanı sıra, genellikle uygun değerlere ve her bir ortak değişkene karşı çizilir. $\chi 2$ test istatistiği, genellikle normal bir regresyon modelinin genel model uyum iyiliğini ölçmek için kullanılır. $\chi 2$ istatistiği $(n - p)$ serbestlik dereceli asimptotik $\chi 2$ dağılımına sahiptir. Ancak, normal olmayan regresyonda, özellikle yanıt değişkeninin az sayıda farklı değere dağıtıldığı verilerde, Pearson ve sapma artıkları tipik olarak (marjinal ve koşullu olarak) normal dağılım izlemez. Artıkların grafikleri neredeyse paralel eğriler içerir, bu da karşılaştırma için birleşik bir referans dağılımının olmaması nedeniyle yorumlanmasını zorlaştırır.

4.4. Rastgele Kantil Artıklar (RQR)

Dunn ve Smyth (1996) tarafından önerilen RQR, özellikle sayma regresyon modelleri için uygun bir yöntemdir. RQR, yaklaşık olarak normal dağılmış artıklar elde etmek için her gözlem için tahmin edilen dağılım fonksiyonunu tersine çevirme fikrine dayanmaktadır.

Normal bir regresyon modeli durumunda, kantil artıklar Pearson ve sapma artıklarına eşittir. Verilerin çarpık ve sürekli bir yanıt değişkenini modellemek için normal olmayan bir regresyon modeli durumunda, Scudilio ve Pereira (2020), geleneksel artık analizi yöntemine göre bu yöntemin daha iyi olduğunu önermiştir. Lemonte ve Arenas (2019) sırasıyla beta

regresyon ve genelleştirilmiş Johnson SB regresyon modellerinde kuantil artıkların özelliklerini araştırmışlardır. Elde ettikleri sonuçlar, kuantil artıkların standart normal dağılıma daha iyi bir şekilde yaklaştığını göstermiştir.

RQR, geleneksel artıklarda yukarıda bahsedilen sorunları aşmak için literatürde önerilmiştir. Literatürde sıfır yığılmalı modellerin model seçiminde RQR uygulanması oldukça yeni bir konudur. Bu tez çalışmasında, Bell ve sıfır yığılmalı Bell regresyon modelleri için RQR'leri hesaplayabilen R fonksiyonları oluşturulmuştur.

RQR'lerin artan farkındalığına rağmen, RQR'lerin kullanımına ilişkin literatür öncelikle, Poisson, Negatif binom, ters Gauss, Gamma veya sürekli olmayan üstel aileye ait yanıt değişkeninin dağılımı ile genelleştirilmiş doğrusal modellere odaklanmıştır.

RQR yöntemi, yanıt değerinin sayma verileri olduğu regresyon modellerinden elde edilen sonuçların değerlendirilmesinde geleneksel artıkların kullanılmasındaki zorlukların üstesinden gelmek için önerilen bir yöntemdir (Dunn ve Smyth, 1996). RQR'yi hesaplamak için gereken tek bilgi, yanıt değişkeninin kümülatif dağılım fonksiyonunu bilmektir. Tanımdan da anlaşılacağı gibi RQR yöntemi diğer yöntemlere göre daha basittir. RQR yöntemi temelde bir dönüşüm yapmaktadır. $F(y_i, \hat{\mu}_i, \hat{\Phi})$, regresyon modelinde $\hat{\mu}_i$ ile parametrelenen rastgele değişken y_i için kümülatif dağılım fonksiyonunu göstersin. $\hat{\Phi}$ dağılım parametresi bağımsız değişkenler kümesine bağlı değildir. Dağılım fonksiyonu sürekli ise, $F(y_i, \hat{\mu}_i, \hat{\Phi})$fonksiyonu (0,1) üzerinde düzgün bir dağılım gösterir.

RQR ayrıca $q_i = \phi^{-1}\{F(y_i, \hat{\mu}_i, \hat{\Phi})\}$ şeklinde tanımlanabilir, burada ϕ^{-1} (•) standart normal dağılımın nicel (quantile) bir fonksiyonudur.

Kümülatif dağılım fonksiyonu kesikli ise, fonksiyonu sürekli hale getirmek için rasgeleleştirme terimi eklenir. Kesikli durum için p$(y_i, \hat{\mu}_i, \hat{\Phi})$, $F(y_i, \hat{\mu}_i, \hat{\Phi})$'nin olasılık yoğunluk fonksiyonu olsun. ui, [0, 1]'deki uniform dağılımdan gelen rastgele bir sayı olmak üzere dağılım fonksiyonu Eş 4.15 de verilmiştir.

$$F^*(y_i, u_i, \hat{\mu}_i, \hat{\Phi}) = F^{-1}(y_i{}^- y_i, \hat{\mu}_i, \hat{\Phi}) + u_i \cdot p(y_i, \hat{\mu}_i, \hat{\Phi}) \quad (4.15)$$

Burada $F^{-1}\left(y_i^{-}y_i, \hat{\mu}_i, \widehat{\Phi}\right)$ alt sınırdır.

RQR, $q_i = q\{F\left(y_i; \hat{\mu}_i, \widehat{\Phi}, u_i\right)\} = \phi^{-1}(F^{*}\left(y_i; \hat{\mu}_i, \widehat{\Phi}, u_i\right))$ şeklinde de ifade edilebilir (Feng ve ark, 2020).

Çizelge 4.3 de bu tez çalışmasında kullanılan regresyon modellerine ilişkin RQR'lar verilmiştir.

Çizelge 4.3. Kullanılan regresyon modellerine göre RQR

Model	RQR
Poisson	$z_i^Q = \Phi^{-1}(ppois(y_i - 1; \widehat{\mu_i}) + u_i\ dpois(y_i; \widehat{\mu_i}))$
NB	$z_i^Q = \Phi^{-1}(pnb\left(y_i - 1; \widehat{\mu_i}; \hat{k}\right) + u_i\ dnb\left(y_i; \widehat{\mu_i}, \hat{k}\right))$
Bell	$z_i^Q = \Phi^{-1}(pbell\left(y_i - 1; \widehat{W_0}\right) + u_i\ dbell\left(y_i; \widehat{W_0}\right))$
ZIP	$z_i^Q = \Phi^{-1}(pzip(y_i - 1; \widehat{\mu_i}, \widehat{p_i}) + u_i\ dzip(y_i; \widehat{\mu_i}, \widehat{p_i}))$
ZINB	$z_i^Q = \Phi^{-1}(pzinb\left(y_i - 1; \widehat{\mu_i}; \hat{k}, \widehat{p_i}\right) + u_i\ dzinb\left(y_i; \widehat{\mu_i}, \hat{k}, \widehat{p_i}\right)$
ZIBell	$z_i^Q = \Phi^{-1}(pzibell\left(y_i - 1; \widehat{W_0}, \widehat{p_i}\right) + u_i\ dzibell\left(y_i; \widehat{W_0}, \widehat{p_i}\right))$

5. UYUM İYİLİĞİ ÖLÇÜTLERİ

Model seçim kriterleri, bir dizi aday model arasından en iyi istatistiksel modeli seçmek için kullanılmaktadır. Sayma verisi içeren veri setlerinin modellenmesinde farklı regresyon modelleri olduğu önceki bölümlerde anlatılmıştır. Literatürde veriler ile modeller arasındaki uygunluğun test edilmesi için farklı istatistiksel yöntemler mevcuttur. En çok kullanılan yöntemler Akaike bilgi kriteri(AIC), Olabilirlik Oran testi (LL) ve Bayesci bilgi kriteri (BIC)dir. Bunların dışında rootogram çizimleri ve Vuong Testi gibi farklı yöntemlerde kullanılabilir. Bu tez çalışmasında bahsedilen yöntemler tanıtılmıştır. Ancak uygulama bölümünde literatürde en çok kullanılan yöntemler olan AIC ve LL değerleri kullanılmıştır.

5.1. Akaike Bilgi Kriteri (Akaike Information Criteria- AIC)

Akaike Bilgi Kriteri (AIC), Japon istatistikçi Hirotsugu Akaike tarafından 1973 yılında önerilmiştir (Akaike, 1973). Ancak, tarihi geçmişe dayansa da AIC istatistiği yirmi birinci yüzyıla kadar yaygın olarak kullanılmaya başlamamıştır. Günümüzde ise AIC, en yaygın olarak kullanılan genel uyum istatistiğidir.

Literatürde AIC istatistiğinin iki farklı gösterimini görmek mümkündür. Geleneksel gösterim Eş 5.1 de verilmiştir.

$$AIC = -2L + 2K = -2(L - K) \tag{5.1}$$

Bir diğer gösterim ise AIC terimlerinin n 'ye bölünmesiyle ile elde edilir ve Eş 5.2 de verilmiştir.

$$AIC = \frac{-2L+2K}{n} = \frac{-2(L-K)}{n} \tag{5.2}$$

Bu iki formülde de n gözlem sayısı, L modele ait log olabilirlik değeri ve K ise tahmin edici (predictor) sayısıdır. $2K$, modelin boyutunu(dimension) ayarlayan(adjust) bir ceza terimi olarak adlandırılır. Bir modele daha fazla parametre eklemenin verileri daha olası hale getirdiği göz önüne alındığında, tahmin edicilerin sayısını artırdıkça $-2L$ küçülür. $2K$ ceza terimi, bu olası yanlılığı ayarlamak için eklenir (Hilbe,2014).

R ve Stata gibi yazılımlarda AIC'nin birinci eşitliğini kullanarak hesaplamalar yapılır. Her iki denklemin aynı yapısal bilgiye sahip olduğunu varsayar. AIC için yukarıda bahsedilen iki gösterim dışında farklı gösterimler de mevcuttur. Bu gösterimler farklı isimler almıştır. Örneğin bunlar dışında popüler olan diğer bir AIC gösterimi sonlu örnekler için AIC (finite sample AIC)'dir. Bunlar dışında tutarlı AIC (consistent AIC) ve düzeltilmiş AIC (Corrected AIC) gibi kullanımları mevcuttur. Diğer gösterimlerde ceza terimi değişse de büyük örneklerde sonuçlar ilk verilen AIC gösteriminden elde edilen sonuçlar yakınsamaktadır.

Modelleme sonrası hesaplanan AIC sonuçları arasında en küçük değere sahip olan model diğer modellere göre daha iyi sayılmaktadır. Ancak bazı durumlarda iki ya da daha fazla model için elde edilen AIC değerleri arasında küçük farklara rastlanabilmektedir. Böyle bir durumda karar verebilmek için Hilbe (2009) yılında bir önerme yapmıştır. Yaptığı önerme Çizelge 5.1 de verilmiştir.

Çizelge 5.1. Hilbe tarafından yapılan önerme

İki model arası fark	Eğer İlk model<İkinci model ise
>0,0 ve ≤2.5	Modeller arasında bir fark yoktur
>2.5 ve ≤6	n>256 ise ilk model seçilir
>6 ve ≤9	n>64 ise ilk model seçilir
>9	İlk model seçilir

5.2. Bayesci Bilgi Kriteri (Bayesian Information Criteria- BIC)

Bayesci Bilgi Kriteri (BIC) ilk olarak 1978'de Gideon Schwarz tarafından önerilmiştir. Hem ticari istatistiksel paketlerde hem de R gibi ücretsiz yazılımlarda bulunan en çok olabilirlik modelleri için ikinci en sık kullanılan uyum istatistiğidir. Gösterimi Eş 5.3 de verilmiştir.

$$BIC = -2L + klog(n) \tag{5.3}$$

Burada k tahmin edici (predictors) sayısı ve n gözlem sayısıdır. Düzeltme terimi olan $klog(n)$ değeri AIC' de yer alan düzeltme teriminden daha yüksek sonuç verdiğinden çoğu araştırmacı tarafından AIC tercih edilir.

5.3. Olabilirlik Oran Testi

Olabilirlik oran testi, farklı modelleri karşılaştırmak için kullanılır. Örneğin, sıfır yığılmalı modellerin klasik sayma regresyon modelleriyle karşılaştırılmasında Poisson modeli negatif binom modeline karşı test etmek için kullanılabilir.

Genelde olabilirlik oran testi Eş 5.4 de verilmiştir.

$$LL = -2(L_R - L_F) \tag{5.4}$$

Burada L_F, tam(full) model için log olasılığı ve L_R indirgenmiş (reduced) bir model için log olabilirliğidir. Bu gösterim birçok araştırmacı tarafından tercih edilmektedir. Sayıma dayalı verileri modelleme amacıyla, testin en önemli kullanımı, sınır testi olarak kullanılmasıdır (boundary likelihood test).

Sınır olabilirlik oranı (BLR) testi, yayılım parametresinin değerinin önemli ölçüde farklı olup olmadığını belirlemek için kullanılan bir testtir.

Ortaya çıkan değer, bir (1) serbestlik derecesine sahip Ki-kare dağılımı ile ölçülür. Test edilen dağılım 0'dan daha küçük bir değer alamadığından sınır budur.

Log olabilirlik değerlerindeki farkın -2 katı değeri verildiğinde, Poisson ve negatif binom logaritmik olasılıklardaki fark 1.352'den küçükse, model Poisson'dur; eğer daha büyükse, verilerin Poisson'dan farklı olarak modellenmesi gerekir.

Diğer modellerle karşılaştırma yapıldığında Log olabilirlik değeri büyük olan model tercih edilir.

5.4. Rootogram

Modelleri karşılaştırmada bir diğer yöntem olarak Rootogram kullanılabilir. Tukey (1977) çalışmasında ilişkili bir grafik aracı olan rootogram'ı kullanmış ve başlangıçta tek değişkenli dağılımların uyumluluğunu değerlendirmiştir. Kleiber ve Zeileis (2016), Rootogram'ı regresyon modelleri için genişleterek bu grafiğin özellikle sayma veri modellerinde aşırı

yayılım ve/veya sıfır yığılma gibi sorunların anlaşılmasında yararlı olduğunu öne sürmüşlerdir. Rootogram, gözlenen frekanslar için histogram benzeri dikdörtgenler veya çubuklar ve frekanslar için bir karekök ölçeğinde eğri çizerek gözlenen ve beklenen değerleri grafiksel olarak karşılaştırır (Kılıç ve Bayrak, 2020). Üç tür Rootogram vardır: sarkık (suspended), asılı (hanging) ve askıya alınmış (standing).

5.5. Vuong Testi

Vuong testi, iç içe geçmeyen (nonnested) modellerin karşılaştırılmasında kullanılan hipotez testlerinden biridir (Tüzel, 2011). İç içe model karşılaştırmalarının haricinde ikili modellerde Vuong testiyle karşılaştırılabilmektedir.

Vuong testi için test istatistiği Eş 5.5 de ifade edilmiştir.

$$V = \frac{\bar{m}\sqrt{n}}{sd(m)} \tag{5.5}$$

Burada $\bar{m}$, m_i nin ortalamasını, $sd(m)$ standart sapmayı ve n ise örnek çapıdır. m_i ise Eş 5.6 da verilmiştir.

$$m_i = \ln\left(\frac{p_{1i}(y_i)}{p_{2i}(y_i)}\right) \tag{5.6}$$

Burada $p_1(.)$ ve $p_2(.)$ olasılık yoğunluk fonksiyonlarıdır.

Vuong test istatistiği standart normal dağılımlıdır. Bu değerin yorumlanması için örnek verilecek olursa 0.05 anlamlılık düzeyi için, V değeri 1.96'dan büyükse, ilk model gerçek modele "daha yakındır" yorumu yapılabilirken eğer V değeri -1.96'dan küçükse ikinci model gerçek modele "daha yakındır" denilebilir. Eğer hesaplanan değer (-1.96; 1.96) arasında değil ise hiçbir model gerçek modele "daha yakın" değildir veya birinci veya ikinci modeli kullanma arasında fark yoktur yorumu yapılabilir (İsmail ve Zamani, 2013).

6. UYGULAMA

Yapılan tez çalışmasının bu bölümünde Türkiye İstatistik Kurumu'ndan alınan gerçek veriler kullanılarak Poisson, Negatif Binom, Bell, sıfır yığılmalı Poisson, sıfır yığılmalı Negatif Binom ve sıfır yığılmalı Bell regresyon modellerinden hangisinin veri setini daha iyi temsil ettiği araştırılmıştır. Daha sonra veri setinin en iyi temsil ettiği düşünülen modele ilişkin Pearson, sapma ve RQR grafikleri çizilmiştir. Yapılan analizler için R Studio paket programı kullanılmıştır. Regresyon analizleri R Studio içinde var olan "MASS", "pscl" ve "bellreg" kütüphaneleri kullanılarak yapılırken çizdirilen artıklar için kodlar tez kapsamında ilk kez yazılmış olup herhangi bir kütüphane altında yer almamaktadır.

6.1. Veri Setinin Tanıtılması ve Betimsel İstatistikler

TÜİK tarafından yapılan "Türkiye Sağlık Araştırması (TSA)" verileri kullanmıştır. 2019 Türkiye Sağlık Araştırması ile bireylerin genel sağlık profilinin ortaya çıkarılması ve ülkelerin gelişmişlik düzeylerini gösteren kalkınma göstergeleri içerisinde önemli bir paya sahip olan sağlık göstergesine yönelik bilgilerin elde edilmesi amaçlanmaktadır. TSA Türkiye toplam bazında tahminler üretecek şekilde tasarlanmıştır. Bu araştırma ile bebek, çocuk ve yetişkinlerin sağlık durumunun yanı sıra 15 ve daha yukarı yaştaki bireylerin sağlık hizmetinden faydalanma durumları, günlük aktivitelerini gerçekleştirirken yaşadıkları zorlukların derecesi ile sigara ve alkol kullanma alışkanlıkları başta olmak üzere sağlık alanındaki birçok gösterge elde edilmektedir (TÜİK,2020).

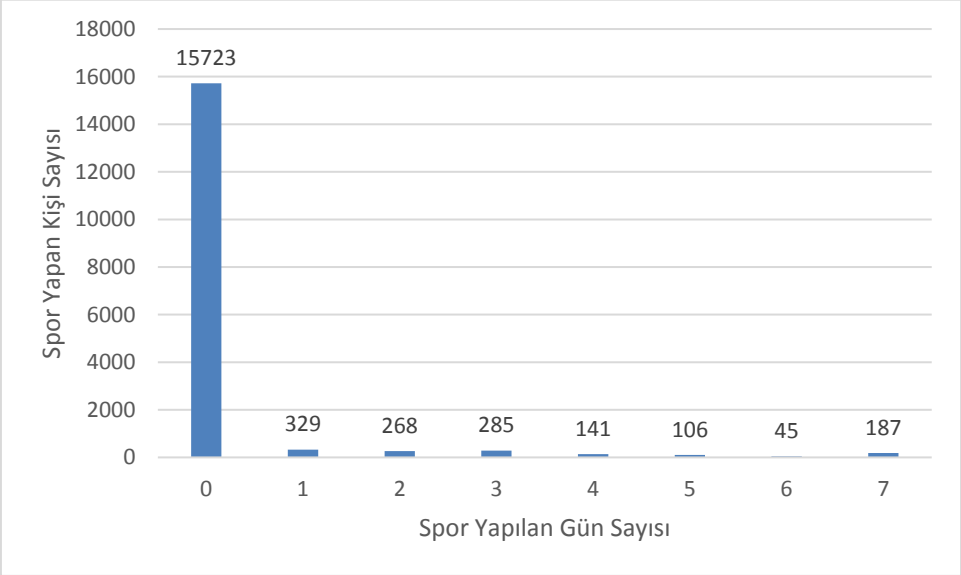
Bu tez çalışmasında anket kapsamında belirtilen konulardan fiziksel aktivite/egzersiz bölümü ile ilgilenilmiştir. Ayrıca yapılan analizlerde kullanılan değişkenlerin seçiminde literatürde yapılan çalışmalar göz önüne alınmıştır.

Yapılan çalışmada kişinin spor yaptığı gün sayısını etkileyen faktörler araştırılmak istenmiştir. Spor yapılan gün sayısı bağımlı değişken iken yaş, kilo, cinsiyet ve genel sağlık durumu değişkenleri bağımsız değişken olarak belirlenmiştir. Seçilen değişkenlere ilişkin temel bilgiler Çizelge 6.1 de verilmiştir.

Çizelge 6.1. Kullanılan değişkenlere ait bilgiler

DEĞİŞKEN ADI	DEĞİŞKEN AÇIKLAMASI	Ortalaması ya da sıklığı	Min-Max
SPOR_GUN (bağımlı değişken)	Bireyin normal bir hafta içerisinde, en az 10 dakika spor, fitness veya bos zaman aktivitesi gerçekleştirmesi (gün olarak)	0.26 (1.04)	0-7
YAS	Kişinin yaşı	43.95(17.67)	15-98
Kilo	Kişinin kilosu	73.7 (14.94)	20-150
Cinsiyet	Kişinin cinsiyeti	Kadın 9300 (%54) Erkek 7784 (%46)	
Genel sağlık durumu	Kişinin genel sağlık durumunu ne şekilde ifade ettiğidir. Kategorileri: Çok iyi, iyi, orta, kötü, çok kötü	Çok iyi 1247 (%7.30) İyi 8741 (51.16%) Orta 5214 (30.52%) Kötü 1655 (9.69%) Çok kötü 227 (1.33%)	

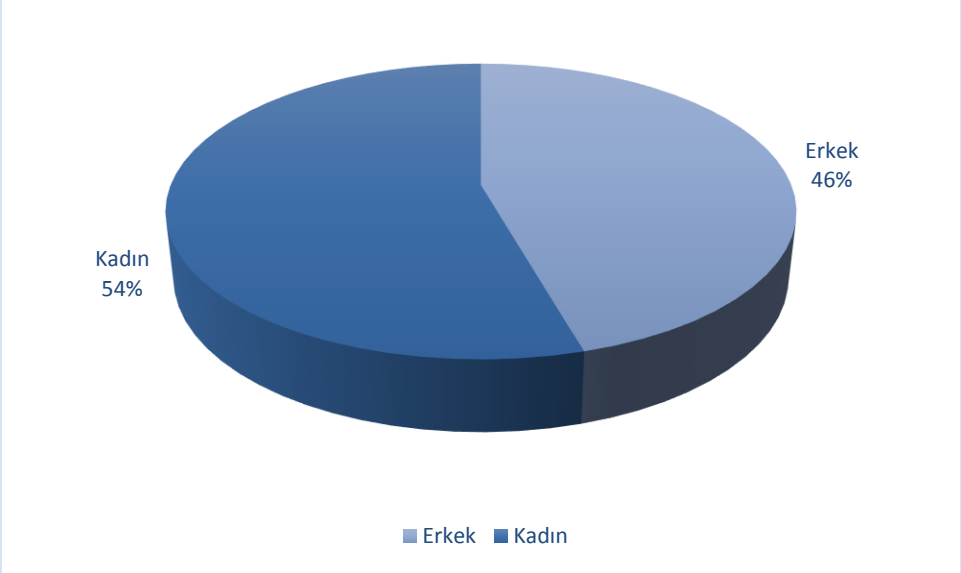
Bağımlı değişken olarak seçilen "spor_gun" sayısının minimum ve maksimum değerleri incelendiğinde değerlerin 0 ile 7 arasında değiştiği görülmektedir. Şekil 6.1 de "spor_gun" sayısı değişkenin dağılımı verilmiştir.



Şekil 6.1. Spor_gun değişkenine ait dağılım

Şekil 6.1 incelendiğinde bağımlı değişkendeki gözlem değerlerinin %92 sinin sıfır değerine sahip olduğu görülmektedir. Yani veride sıfır yığılma durumu söz konusudur. Diğer taraftan bağımlı değişkenin ortalaması 0.26 ve varyansı 1.04 olarak hesaplanmıştır. Varyans ortalamadan büyük olduğundan veride aşırı yayılım durumu söz konusudur. Grafik incelenmeye devam edildiğinde 45 gözlemin 6 gün spor yaptığı ve 187 gözlemin ise haftanın

7 günü de spor yaptığı görülmektedir.



Şekil 6.2. Cinsiyet değişkenine ait dağılım

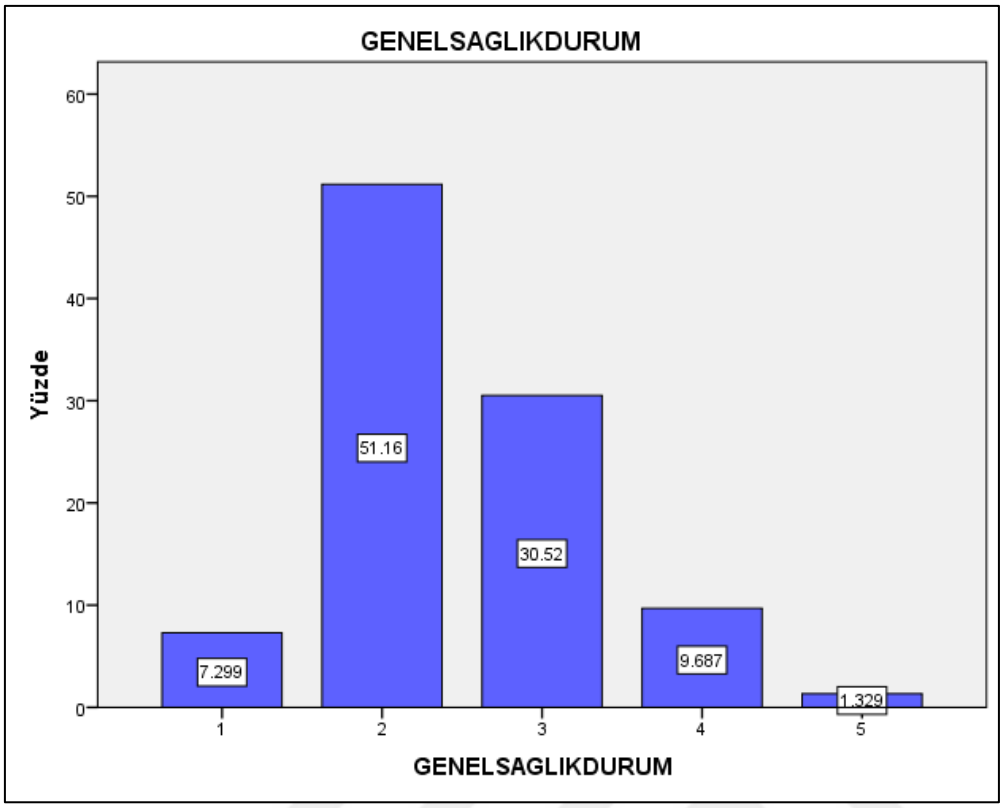
Şekil 6.2 incelendiğinde araştırmaya katılanların %54 ünün kadın %46 sının ise erkek olduğu görülmektedir. Cinsiyetin spor yapılan gün içindeki dağılımı Çizelge 6.2 de verilmiştir.

Çizelge 6.2. Cinsiyetin spor yapılan günlere dağılımı

		CINSIYET		Toplam
		Erkek	Kadın	
SPOR_GUN	0	6929	8794	15723
	1	235	94	329
	2	144	124	268
	3	160	125	285
	4	100	41	141
	5	68	38	106
	6	33	12	45
	7	115	72	187
Toplam		7784	9300	17084

Çizelge 6.2 incelendiğinde erkeklerin 6929'u hiç spor yapmazken kadınların 8794 ünün hiç spor yapmadığı görülmektedir. Haftanın her günü spor yapan erkek sayısı 115 iken haftanın hergünü spor yapan kadın sayısı ise 72 dir.

Genel sağlık durumu incelendiğinde bu değişkenin dağılımın Şekil 6.3 de verildiği gibidir. Grafikte sırasıyla 1: genel sağlık durumu çok iyi, 2:iyi, 3:orta, 4:kötü ve 5: çok kötü olarak belirlenmiştir.



Şekil 6.3. Genel sağlık durumu değişkenine ait dağılım

Şekil 6.3 incelendiğinde kitlede genel sağlık durumu iyi olan gözlemlerin %51 ile en yüksek orana sahip olduğu genel sağlık durumu çok kötü olanların ise %1 ile en düşük orana sahip olduğu görülmektedir. Genel sağlık durumu değişkeni ile spor yapılan gün sayısına ait inceleme Çizelge 6.3 de verilmiştir.

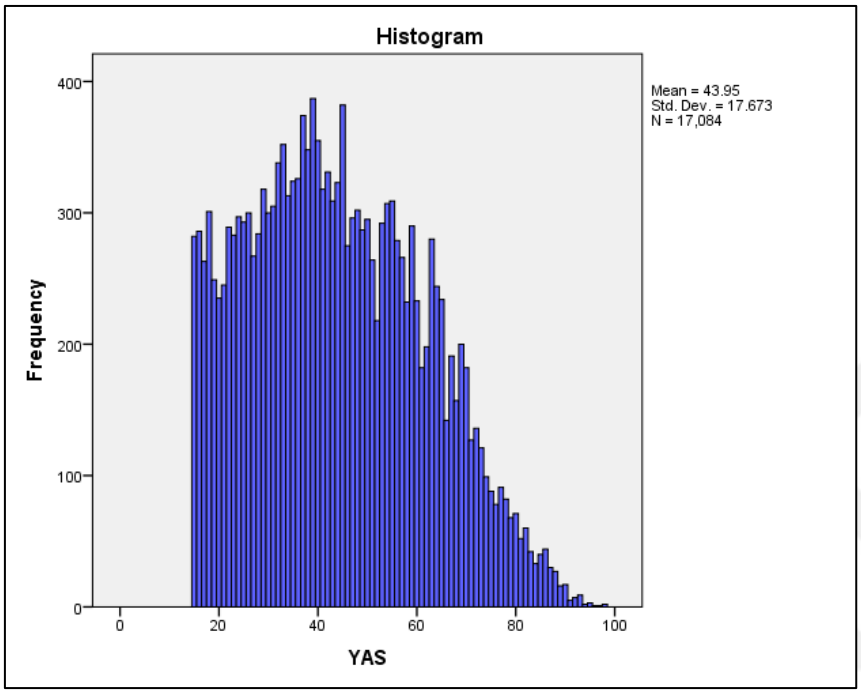
Çizelge 6.3. Genel sağlık durumu spor yapılan gün dağılımı

		GENELSAGLIKDURUM					Toplam
		Çok iyi	İyi	orta	Kötü	Çok kötü	
SPOR_GUN	0	1046	7885	4935	1631	226	15723
	1	46	219	57	7	0	329
	2	32	165	69	2	0	268
	3	39	194	51	1	0	285
	4	34	87	20	0	0	141
	5	18	65	17	5	1	106
	6	5	27	11	2	0	45
	7	27	99	54	7	0	187
Toplam		1247	8741	5214	1655	227	17084

Çizelge 6.3 incelendiğinde genel sağlık durumu "çok kötü" olan bireylerin spor yapmadığı görülürken genel sağlık durumu iyi olan bireylerin ise diğer bireylere göre spora daha fazla

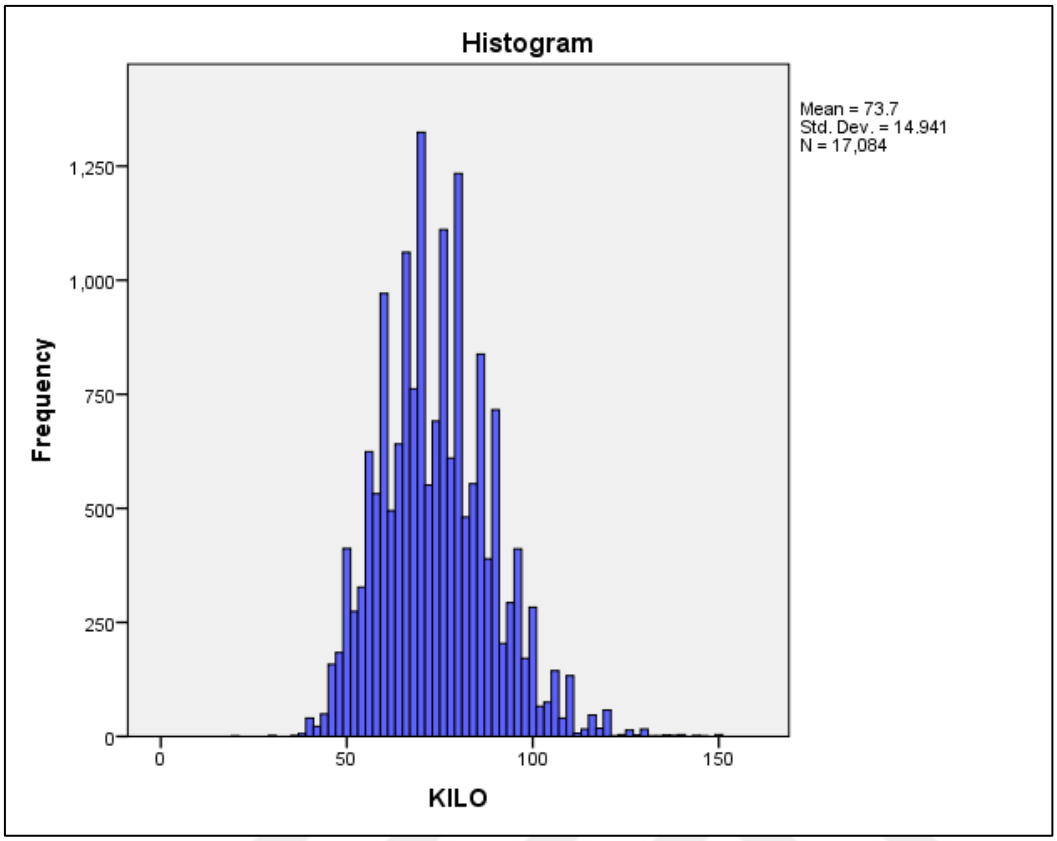
zaman ayırdığı görülmektedir.

Analizde yer alan yaş değişkeni incelendiğinde ise yaşın sağa çarpık bir dağılımı olduğu görülmektedir. Buna ilişkin dağılım Şekil 6.4 de verilmiştir.



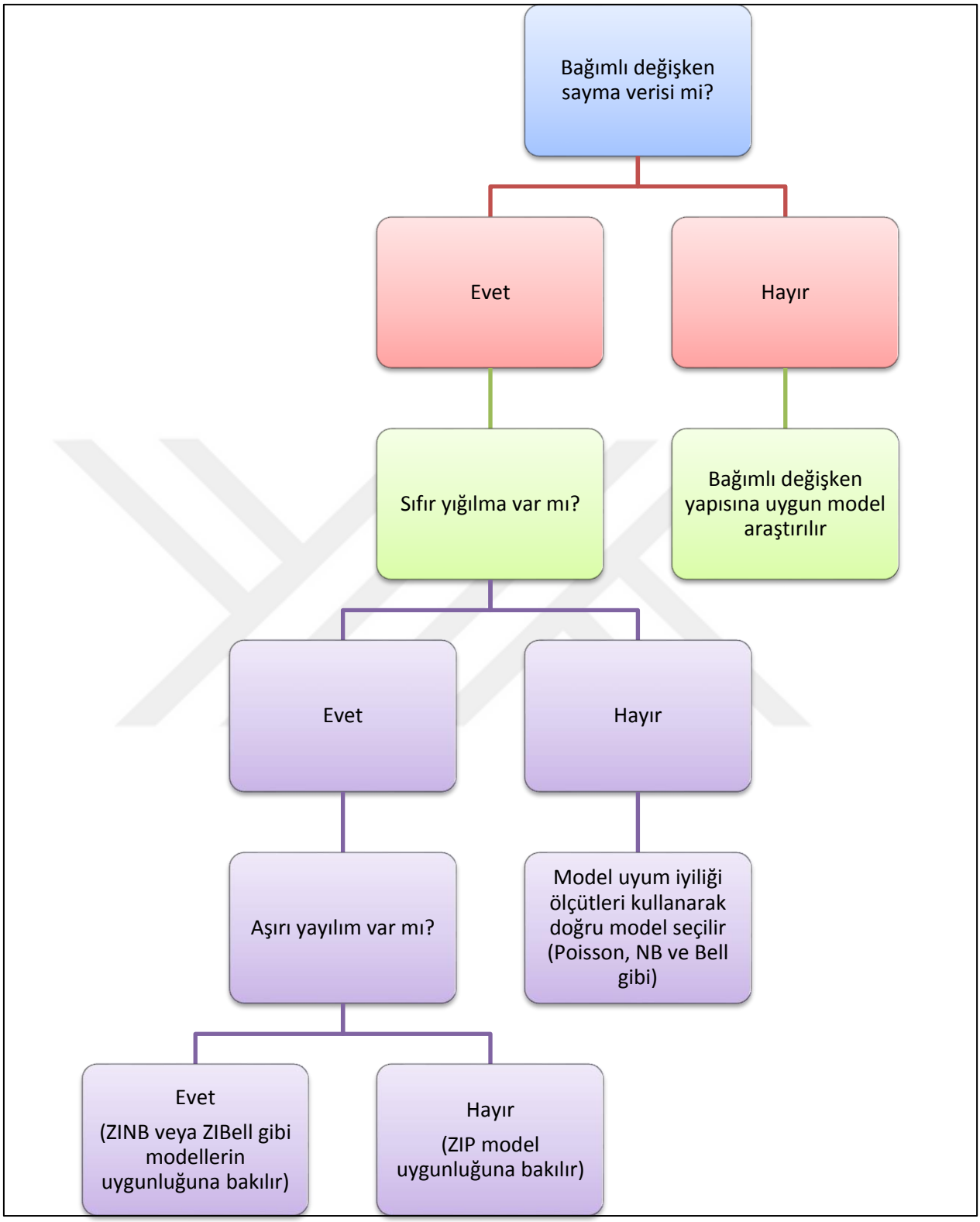
Şekil 6.4. Yaş değişkenine ait dağılım

Kilo değişkeni incelendiğinde bu değişkeninde normal dağılmadığı Şekil 6.5 de görülmektedir.



Şekil 6.5. Kilo değişkenine ait dağılım

Sayma verisi kullanılarak yapılan analizler için uygulanabilecek bir akış diyagramı örneği Şekil 6.6 da verilmiştir.



Şekil 6.6. Sayma verilerinin modellenmesinde uygulanabilecek adımlar

6.2. Uygulama Sonuçları

Regresyon analizine geçilmeden önce yukarıda bağımlı değişkenin ve bağımsız değişkenlerin yapısıyla ilgili bilgi edinmek için bazı ön incelemeler yapılmıştır. Ardından bu değişkenler kullanılarak Poisson, NB, Bell, ZIP, ZINB ve ZIBell regresyon modelleri üzerinden tahminler elde edilmiştir. Bahsedilen modellerden uygun olan modele uygunluğu bazı uygunluk kriterleri kullanılarak karar verilmiştir. Karar verilen modele ait katsayılar yorumlanmış ve Pearson, sapma ve RQR grafikleri çizdirilmiştir.

Uygun modelin seçilmesi için bu tez çalışmasının 5.bölümünde anlatılan AIC ve LL kriterlerinden yararlanılmıştır. Tahmin edilen modellere ilişkin model uyum sonuçları Çizelge 6.4 de verilmiştir.

Çizelge 6.4. Model uygunluğunun belirlenmesi

Modeller	AIC	LL
Poisson	26003.28	-12993.64
Bell	54387.7	-27185.85
NB	14972.25	-7477.126
ZIP	14497.29	-7238.647
ZINB	14373.95	-7175.977
ZIBell	14121.23	-7050.613

Bölüm 5'de de belirtildiği gibi AIC değeri en küçük olan, LL değeri ise en büyük olan model tercih edilir. Karşılaştırma sonucunda Çizelge 6.4 incelendiğinde en düşük AIC ve en büyük LL değerine göre, ZIBell modelinin verilere diğer modellerden daha iyi uyduğu görülmüştür. En büyük AIC ve en küçük LL değerinin gözlendiği Bell modeli ise diğer modeller arasında en az uyum gösteren model olmuştur.

Model seçiminde betimsel istatistiklerin kullanımı araştırmacıya önsel bilgi sağlamaktadır. Şekil 6.1 incelendiğinde ilgilenilen bağımlı değişken de sıfır yığılma olduğu ve Çizelge 6.1 de de aşırı yayılım durumu olduğu görülmüştür. Herhangi bir test yapılmadan veri hakkında bilgi edinilmiştir. Veri setinde aşırı yayılımın varlığı istatistiksel olarak test edilmek istenirse Cameron ve Trivedi (1990) tarafından önerilen aşırı yayılım testi kullanılabilir. Bu teste ait hipotez:

H_0: ϕ=0

H_1: ϕ>0

Şeklindedir. Burada ϕ aşırı yayılımı ifade etmektedir. Cameron ve Trivedi'nin önerdiği bu test için Freitas Souza ve ark. (2022) overdisp() isimli bir R paketi geliştirmişlerdir. Bu paket kullanılarak elde edilen sonuçlar Çizelge 6.5 de verilmiştir.

Çizelge 6.5. Aşırı yayılım testi sonucu

Lambda t test değeri	P değeri
12.475	< 2.2e-16

Buna göre p değeri<0.05 olduğundan H_0 yokluk hipotezi reddedilir. Yani veride aşırı yayılım vardır. Betimsel istatistiklerden elde edilen bu önsel sonuç yapılan istatistiksel model seçim kriterleri sonucunda ZIBell regresyon modelin uygunluğunu göstermiştir.

ZIBell regresyon modeline ait parametre tahminleri Çizelge 6.6 ve Çizelge 6.7 de verilmiştir.

Çizelge 6.6. ZIBell model için parametre tahmini (log kısım)

Log Kısım	Tahmin Değeri	Standart Hata	Z değeri	P değeri	e^{β}
Sabit Terim	0.2839254	0.1861184	1.5255	0.127132	1.3283
Yaş	0.0336471	0.0022990	14.6354	2.2e-16*	1.0342
Kilo	-0.0020685	0.0022316	-0.9269	0.353967	0.9979
Cinsiyet(kadın)	0.6981057	0.0667936	10.4517	2.2e-16*	2.0099
Genel sağlık (iyi)	0.3386502	0.0917969	3.6891	0.000225*	1.4030
Genel sağlık (orta)	0.5773164	0.1101375	5.2418	1.59e-07*	1.7812
Genel sağlık (kötü)	1.6331135	0.2291174	7.1278	1.02e-12*	5.1197
Genel sağlık (çok kötü)	2.6270592	10.093.939	26.026	0.009252*	13.832

Çizelge 6.7. ZIBell model için parametre tahmini (logit kısım)

Logit Kısım	Tahmin Değeri	Standart Hata	Z değeri	P değeri	e^{β}
Sabit Terim	0.7705307	0.0665642	11.5758	2.2e-16*	2.1608
Yaş	0.0054633	0.0017282	3.1613	0.001571*	1.00547

Modelden elde edilen katsayıların yorumlanmasında regresyon katsayılarının üstel formu (e^katsayı)kullanılır. Katsayı sıfır ya da anlamsız ise e^{β} = 1 olacaktır. Pozitif bir etki söz konusu ise e^{β}>1 ve son olarak negatif bir etki söz konusu ise e^{β}<1 olacaktır.

Çizelge 6.6 incelendiğinde Log kısım için kullanılan kilo hariç bütün değişkenlerin p değeri 0.05'ten küçük olduğu için %5 hata payı ile istatistiksel olarak anlamlı olduğu görülmüştür. Elde edilen regresyon denklemi,

μ= exp(0.2839 + 0.0336 yaş - 0.0020 kilo + 0.6981 Kadın + 0.3386 Genel sağlık (iyi) + 0.5773 genel sağlık (orta)+ 1.6331 genel sağlık (kötü)+ 2.6270 genel sağlık (çok kötü)) olarak elde edilmiştir.

Çizelge 6.6 'da cinsiyet değişkeni için referans kategori olarak erkekler alınmıştır. Buna göre kadınların erkeklere göre spor yaptıkları gün sayısı 2 kat daha fazladır. Genel sağlık durumu değişkeni için ise referans kategori çok iyi olarak alınmıştır. Genel sağlık durumu iyi olanlar genel sağlık durumu çok iyi olanlara göre 1.4 kat daha fazla gün spor yaparken genel sağlık durumu çok iyi olanların; genel sağlık durumu iyi olanlara göre spor yaptığı gün sayısı 1,4 kat fazla iken, genel sağlık durumu orta olanlara göre 1,78 kat, kötü olanlara göre 5.11 kat çok kötü olanlara göre ise 13,83 kat daha fazladır. Genel sağlık durumu ile spor yapılan gün arasındaki ilişki betimsel istatistiklerde de benzer sonucu ortaya koymuştu.

Çizelge 6.7 incelendiğinde ise Logit kısım için sabit terim ve yaş değişkenleri p değeri 0.05'ten küçük olduğu için %5 hata payı ile istatistiksel olarak anlamlı olduğu görülmüştür. Logit kısım için elde edilen regresyon denklemi,

$\left(\frac{\pi}{1-\pi}\right)$= exp(0.7705 + 0.0054 Yaş) şeklindedir.

Tez kapsamında belirtilen diğer modeller veri setine uygulandığında elde edilen sonuçlar sırasıyla aşağıdaki tablolarda verilmiştir.

Poisson regresyon modeline ait tahmin sonuçları Çizelge 6.8 de verilmiştir.

Çizelge 6.8. Poisson model için parametre tahmini

Değişkenler	Tahmin Değeri	Standart Hata	Z değeri	P değeri	e^{β}
Sabit Terim	0.2798713	0.0905371	3.091	0.00199*	1.32
Yaş	-0.0244623	0.0011175	-21.890	<2,00E-16*	0.97
cinsiyet (Kadın)	-0.6153843	0.0340673	-18.064	<2,00E-16*	0.54
Kilo	0.0003325	0.0011241	0.296	0.76738	1.0003
Genel sağlık (iyi)	-0.3541154	0.0437338	-8.097	5.63e-16*	0.70
Genel sağlık (orta)	-0.5049772	0.0539824	-9.354	<2,00E-16*	0.60
Genel sağlık (kötü)	-1.3685365	0.1109844	-12.331	<2,00E-16*	0.25
Genel saglik (cok kötü)	-2.2106305	0.4501885	-4.910	9.09e-07*	0.11

Çizelge 6.8 incelendiğinde Sabit terim, yaş, cinsiyet ve genel sağlık durumu değişkenlerinin p değeri 0.05'ten küçük olduğu için %5 hata payı ile istatistiksel olarak anlamlı olduğu görülmüştür. Kilo değişkeni ise %5 hata payı ile istatistiksel olarak anlamlı bulunamamıştır. Ancak betimsel istatistiklerden yola çıkılarak veride aşırı yayılım ve sıfır yığılma olduğu görülmüştü. Poisson regresyona ait parametre tahminleri ortalamanın varyansa eşit olduğu durumda tutarlıdır. Dolayısı ile Negatif Binom Regresyon modeline geçilmiş ve bu modele uygunluğuna bakılmıştır.

Negatif binom regresyon modeline ait tahmin sonuçları Çizelge 6.9 da gösterilmiştir.

Çizelge 6.9. NB model için parametre tahmini

Değişkenler	Tahmin Değeri	Standart Hata	Z değeri	P değeri	e^{β}
Sabit Terim	2,86E+02	2,50E+02	1.145	0.25226	17.46
Yaş	-2,33E+01	2,57E+00	-9.053	<2,00E-16*	0.009
cinsiyet (Kadın)	-5,61E+02	8,30E+01	-6.763	1.35e-11*	0.003
Kilo	-2,12E-02	2,82E+00	-0.008	0.99400	0.12
Genel sağlık (iyi)	-4,08E+02	1,43E+02	-2.862	0.00421*	0.016
Genel sağlık (orta)	-5,81E+02	1,58E+02	-3.684	0.00023*	0.002
Genel sağlık (kötü)	-1,35E+03	2,11E+02	-6.393	1.63e-10*	0.259
Genel sağlık (cok kötü)	-2,24E+03	5,58E+02	-4.008	6.13e-05*	0.106

Çizelge 6.9 incelendiğinde yaş, cinsiyet ve genel sağlık durumu değişkenlerinin p değeri 0.05'ten küçük olduğu için %5 hata payı ile istatistiksel olarak anlamlı olduğu görülmüştür. Kilo ve sabit terim değişkenleri ise %5 hata payı ile istatistiksel olarak anlamlı bulunamamıştır. Betimsel istatistiklerden yola çıkılarak veride sıfır yığılma olduğu görülmüştü. Bu nedenle aşırı yayılım olduğunda negatif binom regresyonu kullanılabilmesine rağmen sıfır yığılma durumu gözardı edilirse hatalı tahminler yapılmış olacaktır.

Aşırı yayılım varlığında kullanılabilen bu tez kapsamında tanıtılmış ve oldukça yeni olan diğer bir model olan Bell regresyon modeline ait tahmin sonuçları Çizelge 6.10 da gösterilmiştir.

Çizelge 6.10. Bell model için parametre tahmini

Değişkenler	Tahmin Değeri	Standart Hata	Z değeri	P değeri	e^{β}
Sabit Terim	0.28001208	0.10428405	2.6851	0.007251*	1.32
Yaş	-0.02417107	0.00122670	-19.7042	<2,00E-16*	0.98
cinsiyet (Kadın)	-0.60566776	0.03786010	-15.9975	<2,00E-16*	0.54
Kilo	0.00023259	0.00128127	0.1815	0.855951	1.0002
Genel sağlık(iyi)	-0.36207783	0.05106377	-7.0907	1,33E-09*	0.69
Genel sağlık(orta)	-0.51784316	0.06138538	-8.4359	<2,00E-16*	0.59
Genel sağlık(kötü)	-1.37101322	0.11735483	-11.6826	<2,00E-16*	0.25
Genel sağlık (cok kötü)	-2.22028188	0.45686395	-4.8598	1,18E-03*	0.11

Çizelge 6.10 incelendiğinde sabit terim, yaş, cinsiyet ve genel sağlık durumu değişkenlerinin p değeri 0.05'ten küçük olduğu için %5 hata payı ile istatistiksel olarak anlamlı olduğu görülmüştür. Kilo değişkenin ise %5 hata payı ile istatistiksel olarak anlamlı bulunamamıştır. Veride sıfır yığılma durumu olduğundan her ne kadar Bell regresyon aşırı yayılıma izin verse de bu tahmin sonuçlarını kullanmak hatalı olacaktır.

ZIP regresyon modeline ait tahmin sonuçları Çizelge 6.11 ve Çizelge 6.12 de verilmiştir.

Çizelge 6.11. ZIP model için parametre tahmini (Log kısım)

Log Kısım	Tahmin Değeri	Standart Hata	Z değeri	P değeri	e^{β}
Sabit Terim	1.226809	0.109343	11.220	2e-16*	3.41
Yaş	0.004875	0.001250	3.901	9.59e-05*	1.004
Kilo	-0.074310	0.039229	-1.894	0.0582	0.93
Cinsiyet(kadın)	-0.002422	0.001399	-1.731	0.0834	0.99
Genel sağlık (iyi)	-0.103146	0.045939	-2.245	0.0247*	0.90
Genel sağlık (orta)	-0.032652	0.056292	-0.580	0.5619	0.96
Genel sağlık (kötü)	0.084026	0.129952	0.647	0.5179	1.08
Genel sağlık (çok kötü)	0.195464	0.632677	0.309	0.7574	1.21

Çizelge 6.12. ZIP model için parametre tahmini (Logit kısım)

Logit Kısım	Tahmin Değeri	Standart Hata	Z değeri	P değeri	e^{β}
Sabit Terim	0.886224	0.073847	12.00	<2e-16*	2.42
Yaş	0.038281	0.001912	20.03	<2e-16*	1.04

Çizelge 6.11 ve Çizelge 6.12 incelendiğinde log kısım için sabit terim, yaş ve genel sağlık(iyi) değişkenlerinin p değeri 0.05'ten küçük olduğu için %5 hata payı ile istatistiksel olarak anlamlı olduğu görülmüştür. Cinsiyet, kilo ve genel sağlık orta, kötü, çok kötü değişkenlerinin ise %5 hata payı ile istatistiksel olarak anlamlı olmadığı görülmektedir.

ZINB regresyon modeline ait tahmin sonuçları Çizelge 6.13 ve Çizelge 6.14 de verilmiştir.

Çizelge 6.13. ZINB model için parametre tahmini (Log kısım)

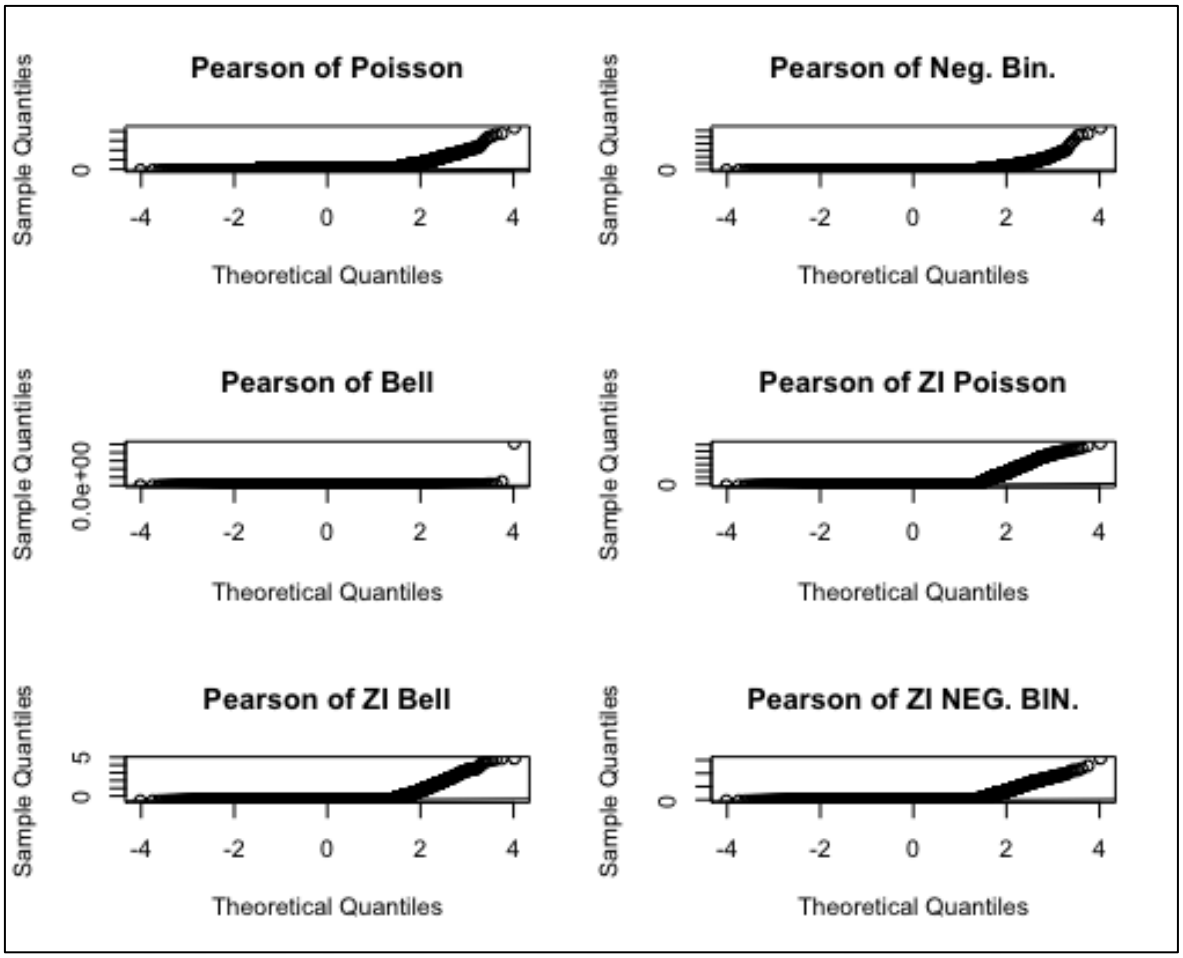
Log Kısım	Tahmin Değeri	Standart Hata	Z değeri	P değeri	e^{β}
Sabit Terim	1.149507	0.149753	7.676	1.64e-14*	3.15
Yaş	0.006008	0.001720	3.493	0.000478*	1.006
Kilo	-0.143784	0.054088	-2.658	0.007853*	0.87
Cinsiyet(kadın)	-0.002292	0.001902	-1.205	0.228350	0.99
Genel sağlık (iyi)	-0.127276	0.062813	-2.026	0.042738*	0.88
Genel sağlık (orta)	-0.079545	0.077957	-1.020	0.307553	0.92
Genel sağlık (kötü)	-0.178778	0.207941	-0.860	0.389925	0.83
Genel sağlık (çok kötü)	-1.293198	0.780963	-1.656	0.097742	0.27

Çizelge 6.14. ZINB model için parametre tahmini (Logit kısım)

Logit Kısım	Tahmin Değeri	Standart Hata	Z değeri	P değeri	e^{β}
Sabit Terim	0.764390	0.079171	9.655	<2e-16*	2.14
Yaş	0.038882	0.001977	19.672	<2e-16*	1.03

Çizelge 6.13 ve Çizelge 6.14 incelendiğinde log kısım için sabit terim, yaş, cinsiyet ve genel sağlık(iyi) değişkenlerinin p değeri 0.05'ten küçük olduğu için %5 hata payı ile istatistiksel olarak anlamlı olduğu görülmüştür. Kilo ve genel sağlık orta, kötü, çok kötü değişkenlerinin ise %5 hata payı ile istatistiksel olarak anlamlı olmadığı görülmektedir.

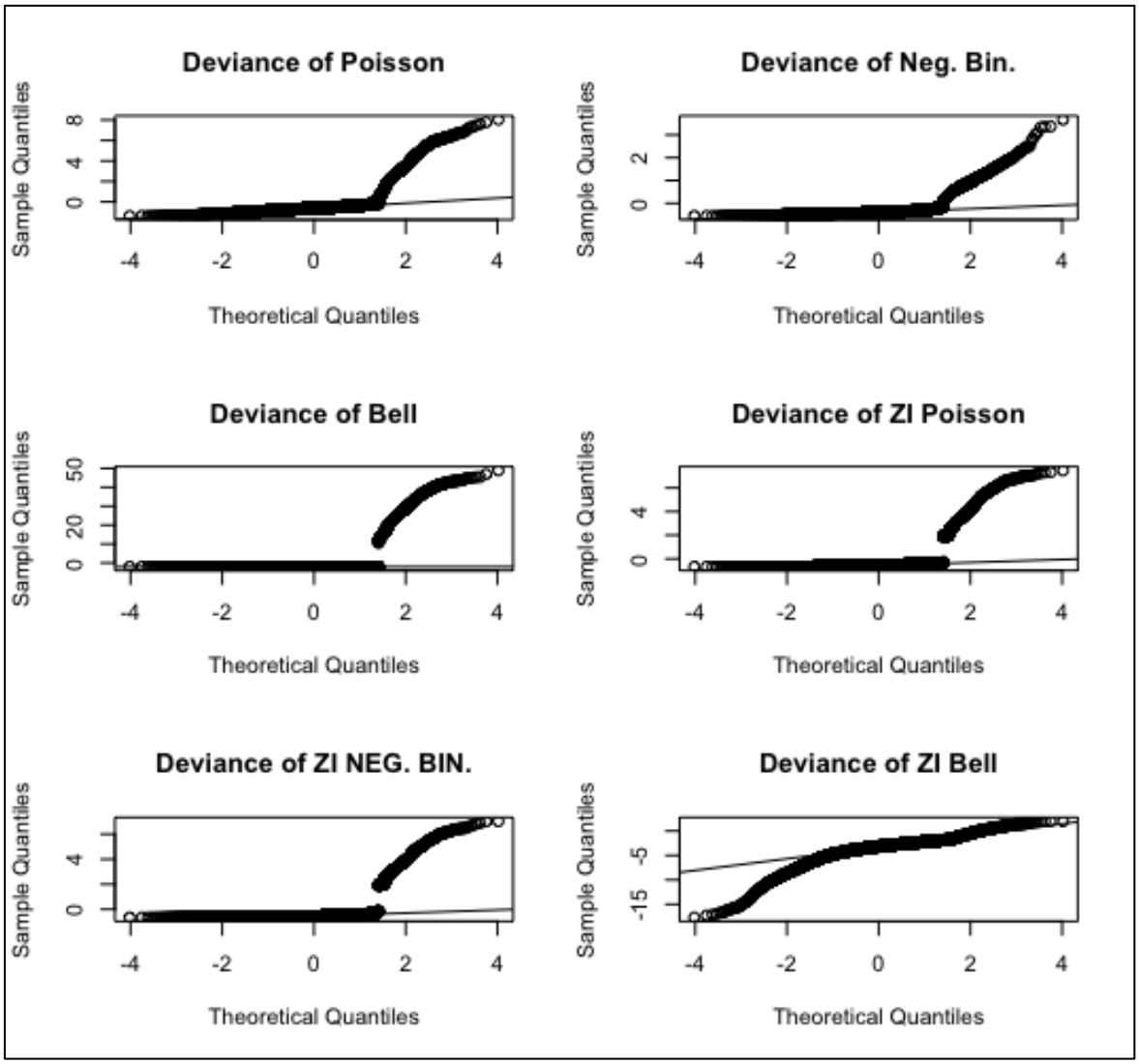
Bu tez çalışmasında seçilen modelin doğruluğunu kontrol etmek çeşitli artık türleri kullanılmıştır. Bu artık türlerinden RQR'ın sayma verilerinde kullanılması son zamanlarda popüler olmaya başlamış bir konudur. Poisson, NB, ZIP ve ZINB modelleri için RQR kullanımı Feng ve diğerleri (2020) önermiştir. Ancak ZIBell ve Bell modelleri için RQR kullanımı ilk kez tanıtılmaktadır. Bu artıklara ilişkin çizilen QQ grafikleri için yazılan kodlar orijinal olup herhangi bir R kütüphanesinde ya da farklı bir paket programda yer almamaktadır. Seçilen modelin doğrulanması kapsamında çizilen ilk artık türü Pearson artıklardır. Şekil 6.7 de elde edilen Pearson artıklar çizdirilmiştir:



Şekil 6.7. Pearson artıkların QQ Grafikleri

Şekil 6.7 incelendiğinde modeller arasında fark olup olmadığı açıkça görülmemektedir. Bütün grafiklerin birbirine oldukça benzer yapıda olduğu görülmektedir. Tez çalışmasının 4.bölümünde de belirtildiği gibi Pearson artıklar sayma verilerinin analizinde uygun değildir.

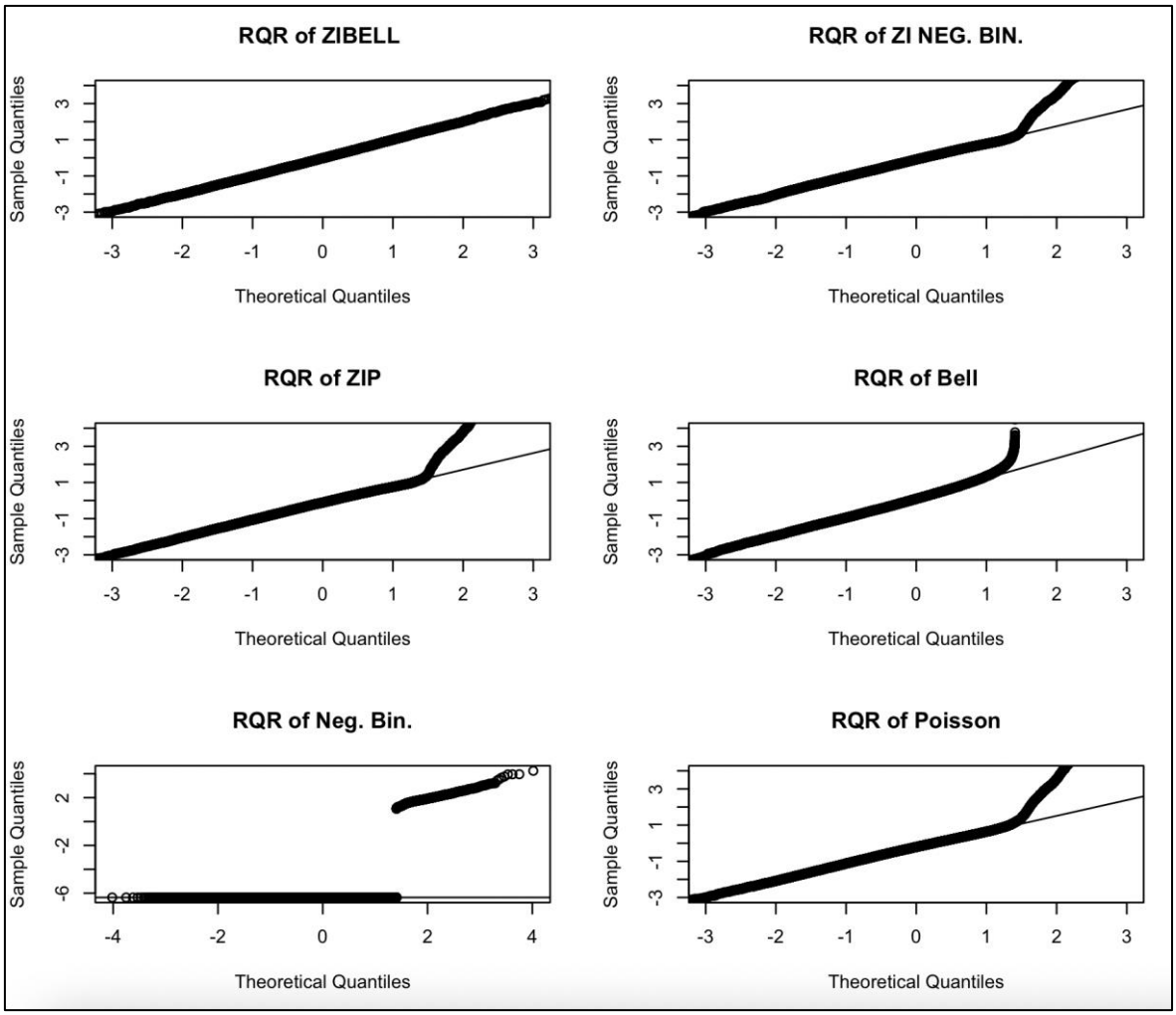
Bir diğer artık türü olan sapma artıklara ait QQ grafikleri Şekil 6.8 de verilmiştir.



Şekil 6.8. Sapma (deviance) artıkların QQ grafikleri

Pearson artıklarda olduğu gibi Sapma artıklarında sayma verilerinde kullanımının önerilmediği daha önce belirtilmiştir. Şekil 6.8 incelendiğinde modeller arasında bir seçim yapmak mümkün değildir.

Bir diğer artık türü olan ve sayma verileri kapsamında önerilen artık türü olan RQR için elde edilen QQ grafiği Şekil 6.9 da verilmiştir.



Şekil 6.9. RQR QQ grafikleri

Şekil 6.9 da verilen QQ grafiklerinde de görüldüğü gibi ZIBell modeli için QQ grafikleri, eğimli düz bir çizgi boyunca düşer ve diyagonal çizgiden sadece birkaç nokta hafifçe sapar, bu da RQR'lerin normal dağıldığını gösterir. Diğer beş model için QQ grafikleri incelendiğinde ise eğrilik ve boşluklar görülmektedir. Yani veri modele iyi bir şekilde uymamaktadır. Model uyum iyiliği sonucunda da görüldüğü gibi artık analizi sonucunda da ZIBell modelin veri seti için uygun bir model olduğunu doğrulanmıştır.

7. SİMÜLASYON ÇALIŞMASI

Bu bölümde, iki farklı simülasyon çalışmasına yer verilecektir. Yapılan çalışmalarda Poisson, NB, Bell, ZIP, ZINB ve ZIBell regresyon modellerinin performanslarını farklı sıfır oranları yönünden değerlendirmek amacıyla yapılan simülasyon çalışması sonuçlarına yer verilmiştir. Ayrıca Bell dağılımına dayalı olarak sıfır yığılma durumunda RQR, sapma ve Pearson artıkların performanslarını araştırılmıştır. Yapılan iki simülasyon çalışması da R programı ile yapılmıştır. Bell, ZIBell regresyon modelleri ve artıkların elde edilmesi için hazır bir kütüphane kullanılmamış olup kodlar çalışma kapsamında yazılmıştır. Poisson, NB, ZIP ve ZINB regresyon modelleri için pscl ve MASS paketleri modelleme bölümünde kullanılmıştır. Bell tipi regresyon modellerinin ve ZINB modellerinin RQR, Pearson ve sapma artıklarının çizdirilmesinde R fonksiyonu geliştirilmiştir.

7.1. Simülasyon Çalışması 1

Yapılan bu simülasyon çalışması tasarlanırken ana amaç veride farklı oranlarda sıfır olması durumunda Poisson, NB, Bell, ZIP, ZINB ve ZIBell regresyon modellerinin sıfır yığılma yüzdesi değiştikçe performanslarını nasıl etkilediğini incelemektir. Bu inceleme için literatürde sıklıkla kullanılan AIC değeri göz önüne alınmıştır. Ardından Bell dağılımı altında sıfır yığılma durumunda tez içerisinden bahsedilen 3 farklı artık yönteminin performanslarını incelemek için simülasyon sonucu üretilip seçilen bir modelin artık grafikleri çizilmiştir. Bahsedilen bu amaçlar için Monte Carlo simülasyonu gerçekleştirilmiştir.

Veride farklı oranlarda sıfır olması durumunda Poisson, NB, Bell, ZIP, ZINB ve ZIBell regresyon modellerinin performanslarını nasıl etkilediğini incelemek amacıyla %30, %50, %70 ve %90 oranlarında sıfır içeren dört farklı durum ele alınmıştır. Simülasyona ilişkin veri üretimi aşağıdaki gibi yapılmıştır:

Bağımlı değişken için Bell dağılımı ve düzgün dağılımdan sayılar üretilerek istenilen sıfır oranlarında sayma verileri elde edilmiştir.

Bağımsız değişken olarak bir sürekli ve bir kategorik değişken üretilmiştir. Kategorik değişken için 0.5 olasılıkla yerine koyarak (0,1) aralığından sayı üretilirken, sürekli değişken için (0,1) aralığında düzgün dağılımdan sayı üretilmiştir.

Regresyon katsayıları β_0= 0.3, β_1=-0.01 ve β_2=-0.1 olarak alınmıştır.

Doğru sonuçların elde edilebilmesi için 5000 tekrar yapılmıştır. Tekrar sayısının belirlenmesi aşamasında farklı sayılar denenmiş olup uygun sayının 5000 olduğuna karar verilmiştir.

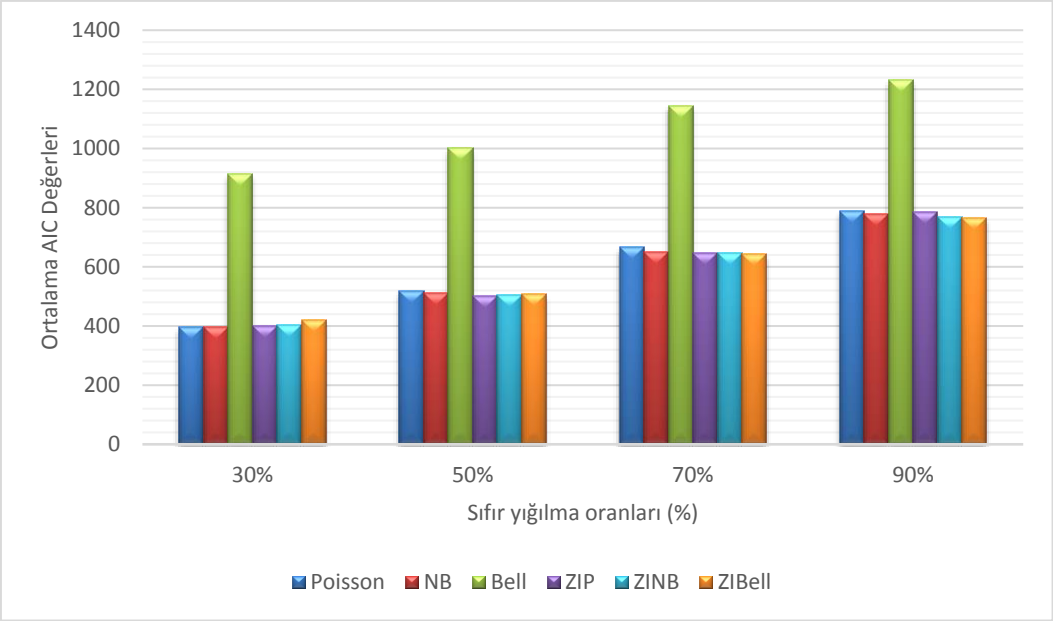
Örneklem büyüklüğü için benzer çalışmalar incelenmiş olup farklı örneklem büyüklüklerinde de benzer sonuçların verildiği görüldüğünden n=250 örneklem büyüklüğü seçilmiş sonuçlar bu örneklem büyüklüğünde incelenmiştir.

Veriler yukarıda anlatılan şekilde üretildikten sonra modellerin performanslarını gözlemlemek amacıyla literatürde sayma veri modellerinin karşılaştırılmasında çoğunlukla tercih edilen AIC değerine bakılmıştır. Üretilen veri seti için en küçük AIC değerine sahip model diğer modellerden daha uygundur yorumu yapılacaktır. Eğer iki AIC değeri yakın sonuçlar vermişse bu modelleri karşılaştırmak için Hilbe (2009)'nin önermiş olduğu kriterden yararlanılacaktır.

Modellerin karşılaştırmasında tekrar sayısının 5000 olarak belirlenmesinden dolayı ortalama AIC değerine bakılarak sonuçlar yorumlanmıştır. Aşağıdaki tabloda farklı sıfır yığılma oranlarına ait AIC değerleri gösterilmiştir.

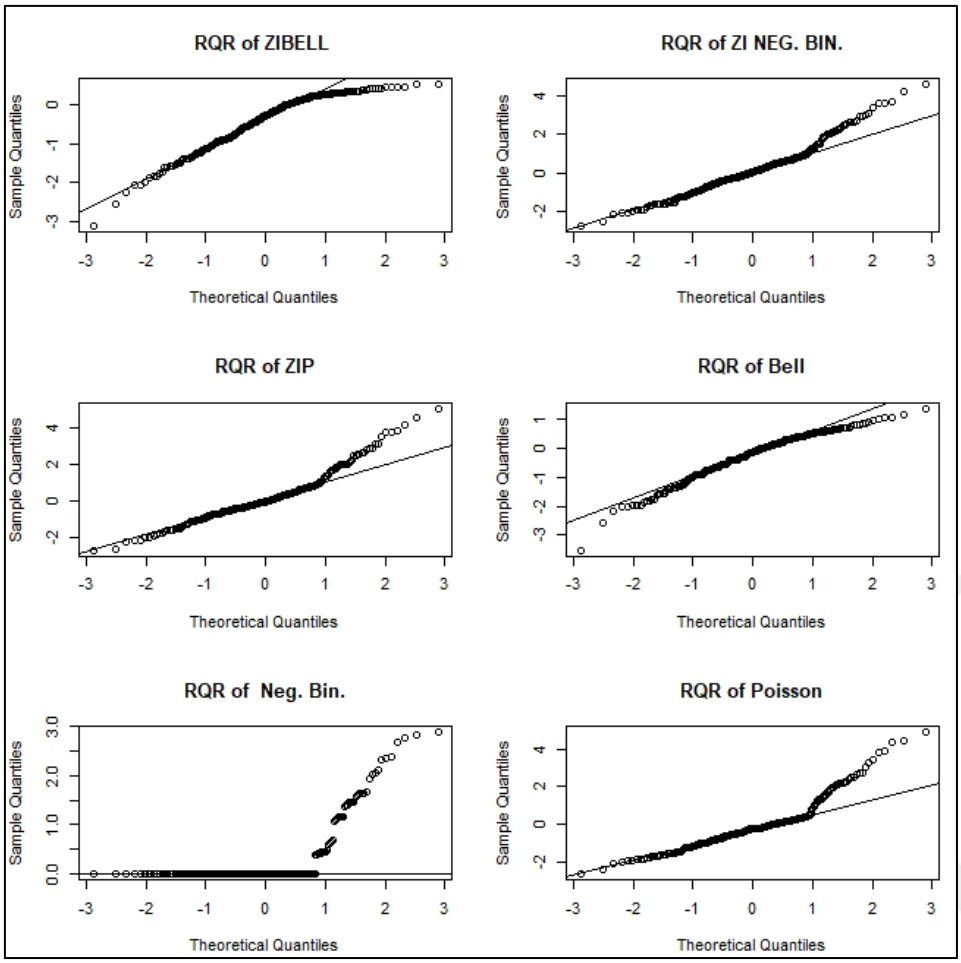
Çizelge 7.1. Farklı sıfır yığılma oranları altın kullanılan regresyon modellerine ait AIC değerleri

Sıfır Yığılma Oranı (%)	Poisson	NB	Bell	ZIP	ZINB	ZIBell
30	396,094	398,096	914,350	401,794	403,800	420,417
50	519,391	511,248	1003,378	503,080	505,079	508,718
70	666,031	651,985	1145,137	647,435	647,587	643,090
90	787,634	777,415	1230,607	786,099	769,306	766,498



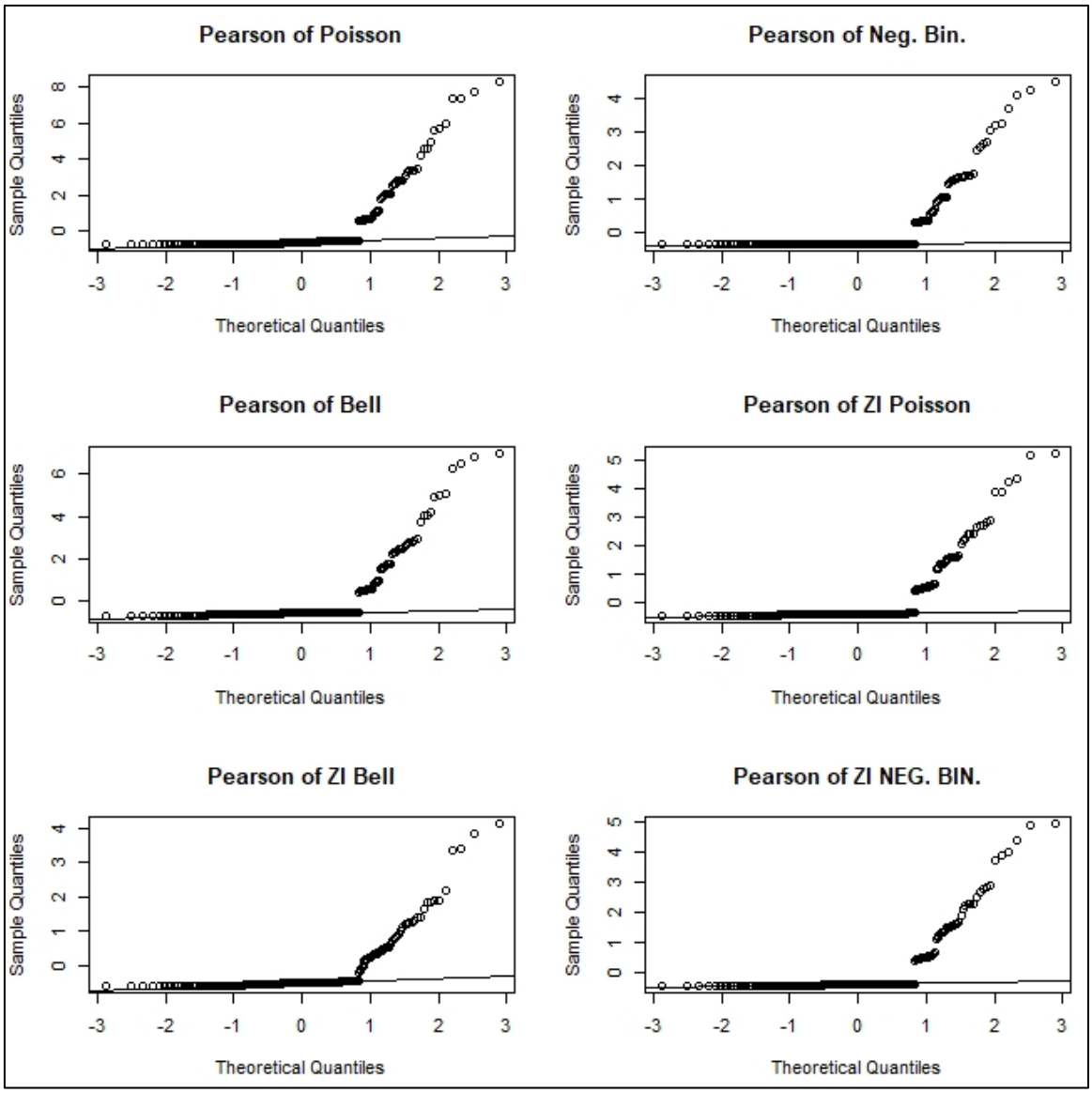
Şekil 7.1. Simülasyon sonucu elde edilen ortalama AIC değerleri

Çizelge 7.1 ve Şekil 7.1 incelendiğinde veride sıfır yığılma durumu arttıkça ZIBell modelin üretilen veri için diğer modellerden daha iyi sonuç verdiği görülmektedir. Ancak sıfır yığılma oranı azaldıkça regresyon modelleri arasında anlamlı bir farklılık olmadığı da gözlenmiştir. Ayrıca %90 sıfır yığılma olması durumunda en başarılı model olan tek bir ZIBell modele ilişkin artıklar çizilerek incelenmiştir.

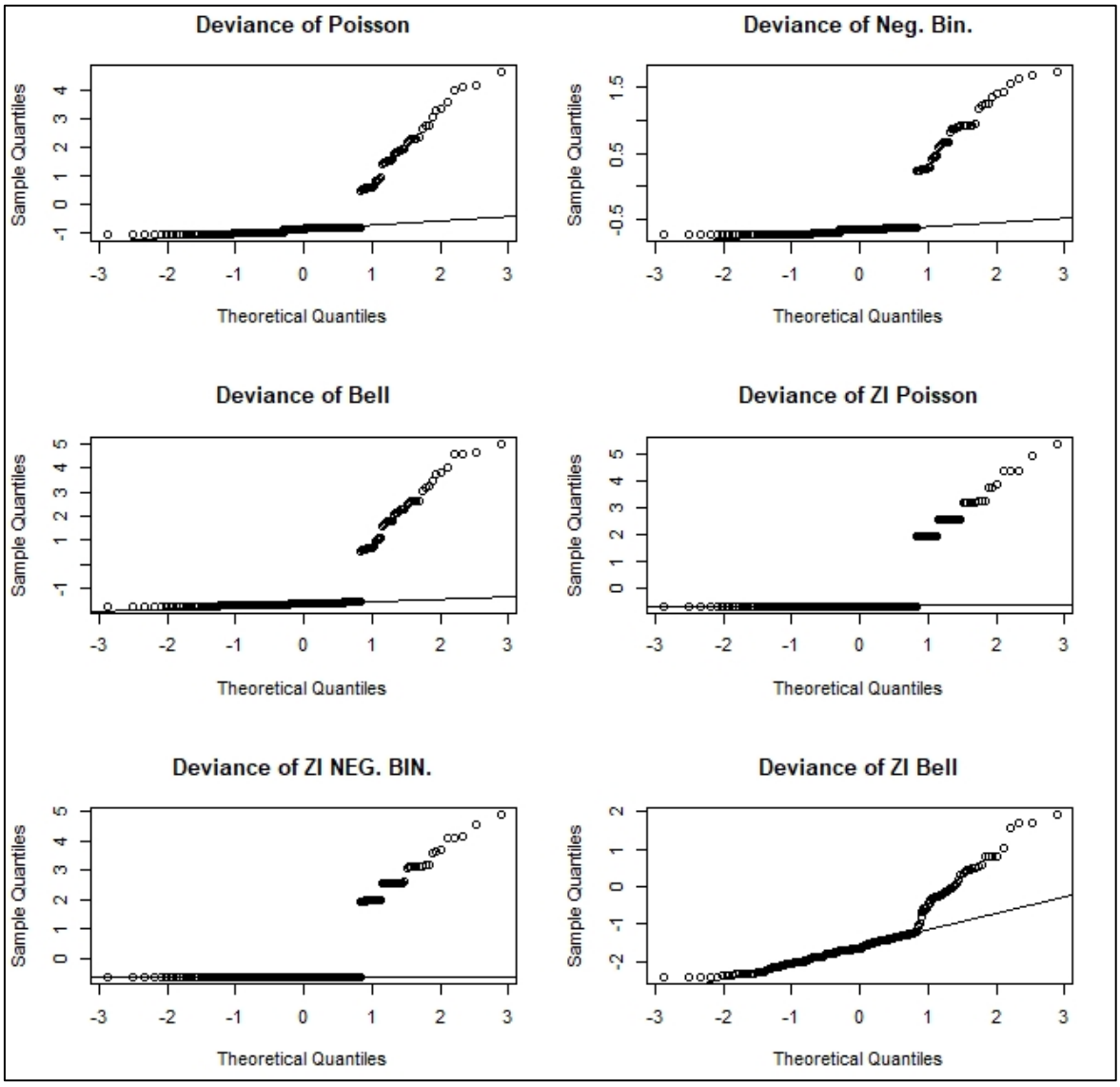


Şekil 7.2. RQR QQ grafikleri

Şekil 7.2 de verilen QQ plot grafiklerinde de görüldüğü gibi ZIBell modeli için QQ grafikleri, eğimli düz bir çizgi boyunca düşer ve diyagonal çizgiden sadece birkaç nokta hafifçe sapmaktadır. Bu durum RQR'lerin normal dağıldığını göstermektedir. Diğer beş model için QQ grafikleri incelendiğinde ise eğrilik ve boşluklar görülmektedir. Yani veri modele iyi bir şekilde uymamaktadır.



Şekil 7.3. Pearson artıkların QQ grafikleri



Şekil 7.4. Sapma artıkların QQ grafikleri

Şekil 7.3 ve Şekil 7.4 de verilen Pearson ve sapma artıklar için çizilen QQ grafikleri incelendiğinde modeller arasında fark olup olmadığı açıkça görülmemektedir. Bütün grafiklerin birbirine oldukça benzer yapıdadır. Modeller arasında eğrilik ve boşluklar görülmektedir. Bölüm 6.2'de verilen gerçek veride yapılan uygulama çalışmasında da görüldüğü gibi simülasyon çalışması sonucunda benzer grafikler elde edilmiştir. Yani Pearson ve deviance artıklar sayma modelleri için bir uygunluk göstermezken RQR daha uyumludur.

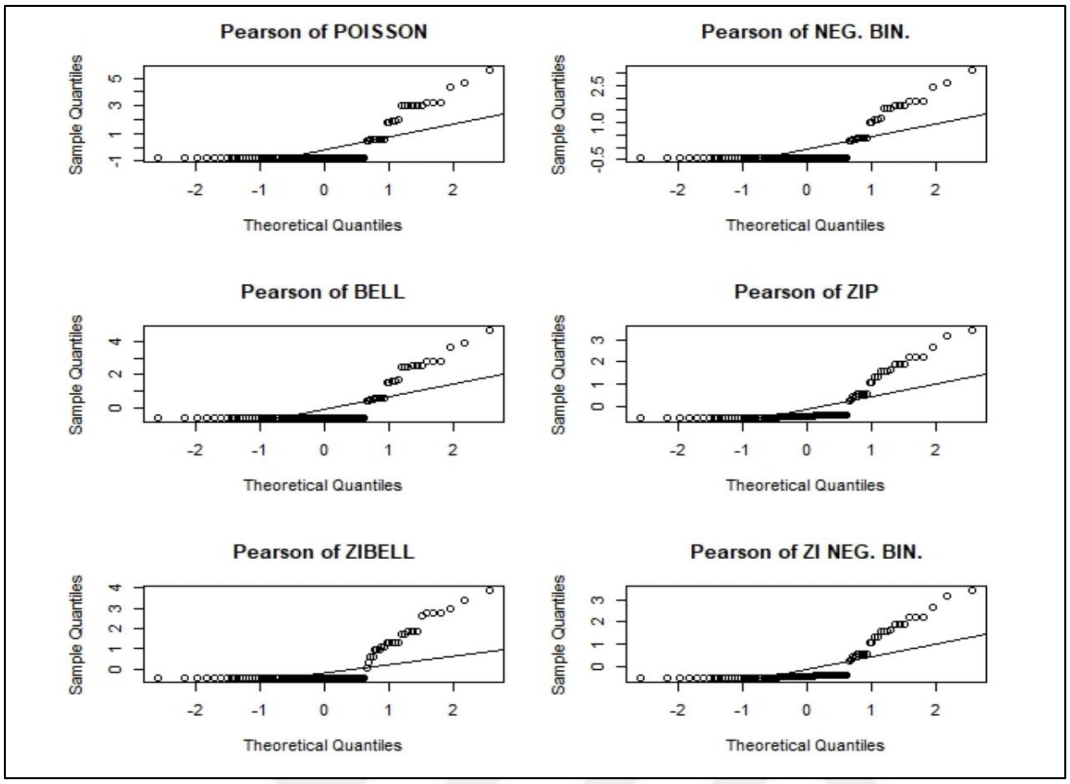
7.2. Simülasyon Çalışması 2

Yapılan bu Monte Carlo simülasyonunda ise Bell dağılımı altında sıfır yığılma durumunda Pearson, sapma ve RQR artık yöntemlerinin performansları araştırılmıştır. Yapılan simülasyonda Eş 7.1 kullanılmıştır:

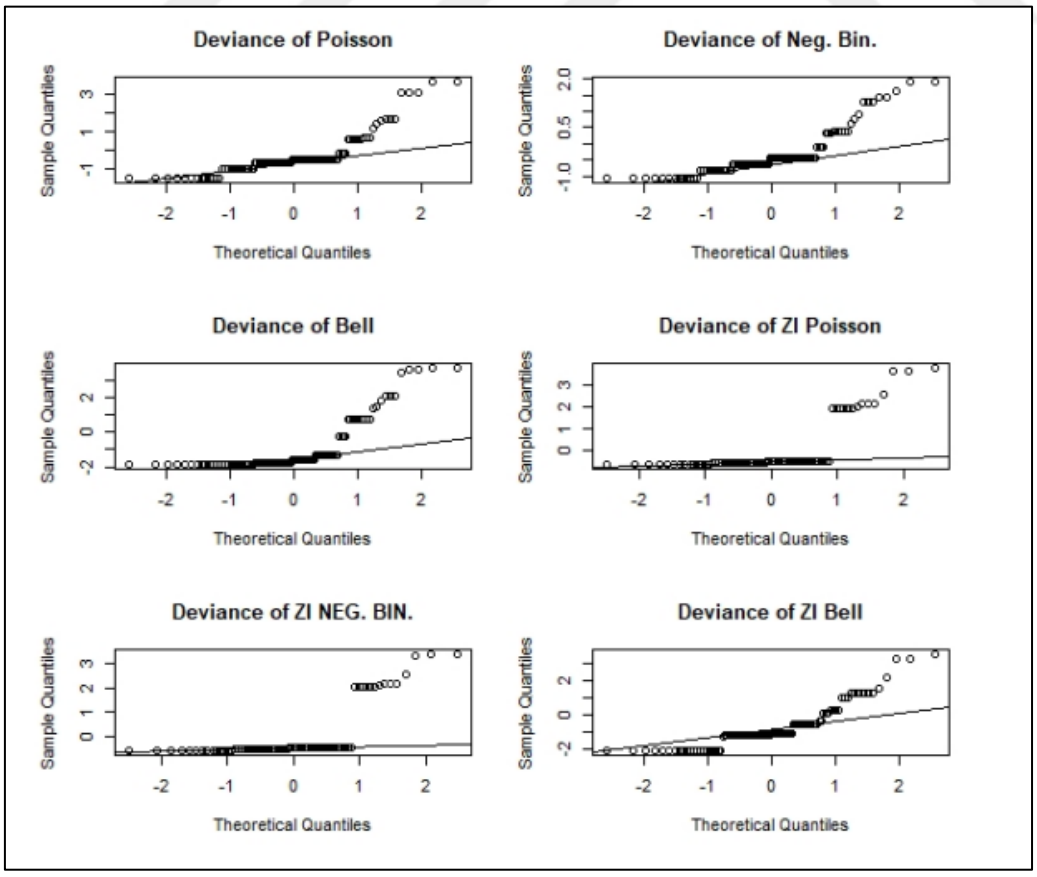
$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2},$$
$$\log\left(\frac{p_i}{1-p_i}\right) = \tau_0 + \tau_1 z_{i1} + \tau_2 z_{i2}, i = 1,2, \dots, n \quad (7.1)$$

Burada $x_{i1} = z_{i1}$ ve $x_{i2} = z_{i2}$ bağımsız değişkenlerinin her ikisi de Bernoulli (p=0.5) dağılımından üretilmiştir. Kullanılacak regresyon modeline ait parametreler Lemonte ve ark. (2019) tarafından yazılan makalede yer alan n = 511 örnek hacimli rastgele seçilmiş denekten elde edilen enfekte kan hücrelerinin sayısının (mm^2 başına) olduğu çalışma ele alınmıştır. Bu veri seti için ZIBell regresyonunun parametre tahminleri olan ve simülasyon tasarımında kullanılan parametrelerin değerleri β_0= 0.716, β_1= −0.611, β_2= 0.036, τ_0=−1.95, τ_1 = 2.17, τ_2 = −0.49 olarak alınmıştır. Ardından bağımlı değişken sıfır yığılma durumu yaklaşık %70 olacak şekilde y_i~ZIBell(μ_i,p_i) olarak üretilmiştir. Model sonuçları elde edilip farklı artık grafikleri çizilmiştir.

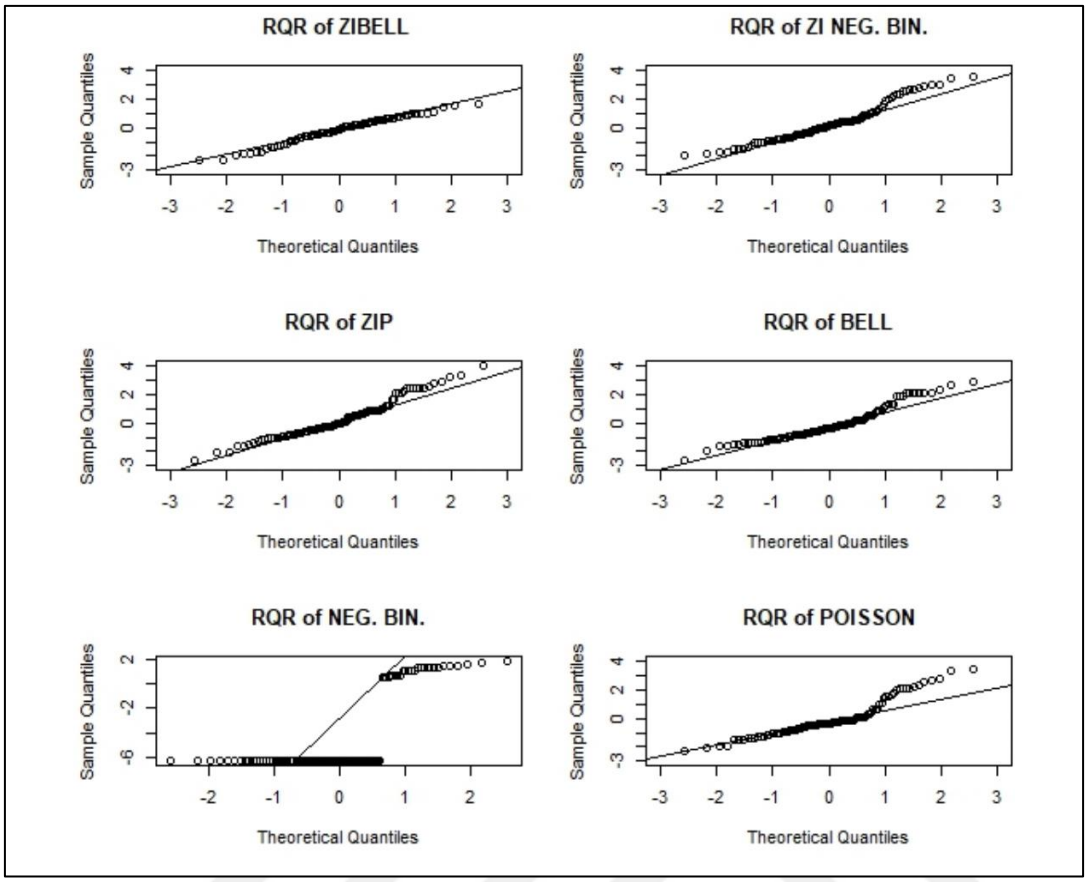
Yapılan simülasyon çalışmasının sonuçları için sıfır yığılma durumunda farklı artık yöntemlerinin performansını incelemek için ZIP, ZINB, Poisson, Bell ve NB regresyon modellerine ait grafikler çizilmiştir. RQR, Pearson ve sapma artıklara ilişkin QQ grafikleri sırasıyla Şekil 7.5, Şekil 7.6 ve Şekil 7.7 da verilmiştir.



Şekil 7.5. Pearson artıklara ait QQ grafiği



Şekil 7.6. Sapma artıklara ait QQ grafiği



Şekil 7.7. RQR ait QQ grafiği

QQ grafikleri, teorik kantiller ile örnek kantiller arasındaki ilişkiyi göstermektedir. QQ plotlar incelendiğinde sıfır yığılmalı modeller için RQR'lerin QQ grafikleri kullanıldığında, verilerin modellerde eğri düz çizgi boyunca düşen noktalarla iyi uyum sağladığı görülmektedir. Özellikle NB modelinde QQ grafiklerinde büyük bir boşluk olması bu modelin üretilen veriler için iyi bir model olmadığını göstermektedir.

Pearson ve sapma artıkları için QQ grafikleri incelendiğinde, tüm grafiklerin gerçek model için bile eğrilik gösterdiği ve diyagonal çizgiden saptığı görülmektedir. Bu nedenle, Pearson ve sapma artıkları, sıfır yığılmalı sayma veri modelleri için uygun değildir. Tek bir bağımlı değişken ele alındığında elde edilen RQR için Shapiro-Wilk testi sonucu Çizelge 7.2 de verilmiştir.

Çizelge 7.2. RQR için Shapiro- Wilk testi sonucu

Model	Shapiro-Wilk testi p değeri
ZIBell	0.5911
ZIP	0.0048
ZINB	0.0131
Bell	0.0027
Poisson	1.3791e-05
NB	8.0203e-16

P değeri>0,05 olan ZIBell model için elde edilen artık değerlerinin normal dağılıma uygun olduğu söylenebilir. RQR yöntemi, sayma veri modellerinde uygun model seçiminin belirlenmesinde kullanılır.

8. SONUÇ VE ÖNERİLER

Bu tez çalışması genel olarak henüz çok yeni bir dağılım olan Bell dağılımının klasik sayma regresyon modellerinde ve sıfır yığılmalı sayma regresyon modellerinde kullanılması problemi ile sayma regresyon modellerinde artık analizinde rasgele kantil artıkların kullanılması problemini ele almayı amaçlamıştır.

Sayma regresyon modellerinde veride aşırı yayılım olduğu durumda Poisson regresyon kullanımı Poisson dağılımının temel varsayımını ihlal ettiğinden kullanılmamalıdır. Literatürde bu soruna çözüm olarak Negatif Binom regresyon modeli yaklaşımı önerilmiştir. Ancak negatif binom dağılımında birden fazla parametre olduğundan Fisher Bilgi matrisi, en çok olabilirlik tahmin edicisi gibi çeşitli istatistiksel çıkarımları elde etmek kolay olmamaktadır. Bu nedenle istatistiksel çıkarımların kolay elde edilmesi ve aşırı yayılım durumunda doğru parametre tahminlerinin elde edilmesi bu çalışmanın önceliklerinden biri olmuştur. Aynı durum veride sıfır yığılma olmasında da geçerlidir. Sıfır yığılmanın gözardı edilmesi parametre tahminlerinde hatalı sonuçlara yol açacaktır. Bu nedenle öncelik olarak alınan diğer bir konu veride sıfır yığılma ve aşırı yayılım olmasıdır. Aşırı yayılım durumunda Negatif Binom regresyon modeline alternatif olarak Bell regresyon kullanımını Türkiye'de öneren ilk çalışma olması nedeniyle sayma verilerinin modellenmesi alanında çalışacak araştırmacılara yeni bir bakış sağlayacağı düşünülmektedir.

Ayrıca sayma verilerinde doğru modelin belirlenmesinde artık analizinin kullanılabilirliği vurgulanmıştır. Ancak sayma verilerinin normal dağılmamasından kaynaklı olarak sıklıkla tercih edilen Pearson ve sapma artıkları kullanmanın model seçiminde etkisi olmayacağı bu nedenle RQR kullanımının gerekliliği vurgulanmıştır. Bell ve sıfır yığılmalı Bell regresyon modelleri için RQR kullanımı ilk kez bu tez çalışmasında önerilmiştir. Sapma ve Pearson artıklar için Bell ve sıfır yığılmalı Bell regresyon modellerinde kullanılabilecek teorik çıkarımlar elde edilmiştir.

Yapılan ilk simülasyon çalışmasında sıfır yığılma durumunda model seçimi üzerinde durularak bu kapsamda karşılaştırmalar yapılmıştır. Sıfır yığılma oranı artıkça veriye uygun olan modelin sıfır yığılmalı modeller olduğu AIC değeri kullanılarak gösterilmiştir.

İkinci simülasyon çalışmasında sayma verilerinde doğru modelin seçiminde doğru artık türünün kullanılması vurgulanmak istenmiştir. Bu amaçla simüle edilen veri kullanılarak parametre tahminleri yapılmış Pearson, sapma ve RQR artıkları için QQ grafikleri çizilerek incelemeler yapılmış ve yorumlanmıştır.

Bu iki simülasyonda elde edilen sonuçlara benzer sonuçlar gerçek bir veri kullanarak uygulama bölümünde de elde edilmiştir. Araştırmanın uygulama bölümünde, Türk insanının spor alışkanlıklarını ortaya çıkarmak amacıyla Türkiye'de bireylerin düzenli bir haftada en az 10 dakika spor, fitness veya boş zaman aktivitelerini gerçekleştirmelerine etki eden faktörlerin belirlenmesi amaçlanmıştır. Bu amaçla sayma veri modelleri dikkate alınmıştır. Bağımlı değişkenin ortalaması 0.26, varyansı 1.08 olarak bulunmuştur. Varyans ortalamadan büyük olduğu için verilerde aşırı yayılım olduğu görülmüştür. Buna ek olarak verilerin %92'si sıfır değerini aldığı için verilerde sıfır yığılma durumu da söz konusudur. Bu bilgi ve model seçim kriterleri ile artık analizi de dikkate alınarak bu veri seti için en uygun modelin ZIBell modeli olduğu görülmüştür. Daha sonra bu modelin parametreleri yorumlanmıştır. Türkiye'deki bireyler için bir hafta içinde yaptıkları sportif faaliyet süresi (gün olarak) incelendiğinde, araştırmaya katılan bireylerin çoğunun spor faaliyetlerine katılmadığı ve bu sayıyı etkileyen başlıca faktörlerin yaş, cinsiyet ve sağlık durumu olduğu görülmüştür. Hem simülasyon çalışmasında hem de gerçek veri uygulamasında kullanılan R kütüphaneleri regresyon modelleri için MASS, pscl ve bellreg'dir. Regresyon analizi sonucu elde edilen artıklar için herhangi bir fonksiyon yada kütüphane kullanılmamış olup ilk kez bu tez kapsamında jenerik kodlar yazılmıştır.

Literatürde artık analizinde kullanılmak üzere önerilmiş student ve Anscombe artıkları gibi başka artık türleri de vardır. Gelecekte sayma veri modellerinde artık analizi hakkında çalışmak isteyen araştırmacılar için yeni artık türlerinin araştırma konusu olabileceği düşünülmektedir.

KAYNAKLAR

Agresti, A. (2003). *Categorical data analysis* (Second edition). New Jersey: John Wiley&Sons Incorporation.

Akaike, H. (1973). *Information theory and extension of the maximum likelihood principle*, Second International Symposium on Information Theory, Budapest, 267–281.

Akdur, H. T. K. (2022). GEE-based Bell model for longitudinal count outcomes. *Communications in Statistics-Theory and Methods*, 51(13), 1-15.

Altun, E. (2018). A new zero-ınflated regression model with application. *İstatistikçiler Dergisi: İstatistik ve Aktüerya*, 11(2), 73-80.

Amin, M., Akram, M. N., and Majid, A. (2020). On the estimation of Bell regression model using ridge estimator. *Communications in Statistics-Simulation and Computation*, 50(5), 1-14.

Avcı, E. (2018). Using count regression models to determine the factors which effects the hospitalization number of people with schizophrenia. *Journal of Data Science*, 16(3), 511-528.

Bai, W. (2018). *Randomized quantile residual for assessing generalized linear mixed models with application to zero-ınated microbiome data*, PhD Thesis, University of Saskatchewan, Canada.

Bell, E. T. (1934). Exponential numbers. *The American Mathematical Monthly*, 41(7), 411-419.

Cameron, A. C. and Trivedi, P. K. (1990). Regression-based tests for overdispersion in the Poisson model. *Journal of Econometrics*, 46(3), 347-364.

Cameron, A.C. and Trivedi, P.K. (2013). *Regression analysis of count data* (Second Edition), New York: Cambridge University Press, 71-140.

Carrivick, P.J.W., Lee, A.H. and Yau, K.K.W. (2003). Zero-inflated Poisson Modeling to Evaluate Occupational Safety Interventions. *Safety Science*, 41(1), 53-63.

Castellares, F., Ferrari, S. L. and Lemonte, A. J. (2018). On the Bell distribution and its associated regression model for count data. *Applied Mathematical Modelling*, 56, 172-185.

Dunn, P. K. and Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5(3), 236-244.

Durmuş, B., and Güneri, Ö. İ. (2020). An application of the generalized poisson model for over dispersion data on the number of strikes between 1984 and 2017. *Alphanumeric Journal*, 8(2), 249-260.

Fagundes, R. A., Souza, R. M. and Cysneiros, F. J. (2016). Zero-inflated prediction model in software-fault data. *IET Software*, 10(1), 1-9.

Famoye, F. and Singh, K.P. (2006). Zero-inflated generalized poisson regression model with an application to domestic violence data. *Journal of Data Science,* 4, 117-130.

Feng, C., Li, L., and Sadeghpour, A. (2020). A comparison of residual diagnosis tools for diagnosing regression models for count data. *BMC Medical Research Methodology*, 20(1), 1-21.

Flynn, M. and Francis, L. A. (2009). More flexible GLMs zero-inflated models and hybrid models. *Casualty Actuarial Society*, 148-224.

Souza, R. D. F., Fávero, L. P., Belfiore, P., and Corrêa, H. L. (2022). overdisp: an R package for direct detection of overdispersion in count data multiple regression analysis. *International Journal of Business Intelligence and Data Mining*, 20(3), 327-344.

Garay, A. M., Lachos, V. H., and Bolfarine, H. (2015). Bayesian estimation and case influence diagnostics for the zero-inflated negative binomial regression model. *Journal of Applied Statistics,* 42(6), 1148-1165.

Greene, W. H., (1994). Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models. *NYU Working Paper*, 94(10), 1-37.

Hilbe, J. M. (2009). *Logistic Regression Models,* New York: Chapman and Hall/CRC Press, 260.

Hilbe, J.M. (2014). *Modelling Count Data*. New York: Cambridge University Press, 20-170.

Hu, M. C., Pavlicova, M. and Nunes, E. V. (2011). Zero-inflated and hurdle models of count data with extra zeros: examples from an HIV-risk reduction intervention trial. *The American journal of drug and alcohol abuse*, 37(5), 367-375.

Ismail N, Zamani H (2013). Estimation of claimcount data using negative binomial, generalized poisson, zero-inflated negative binomial and zero-inflated generalized poisson regression models. *Casualty Actuarial Society E-Forum* 41(20): 1–28.

Javali, S. B., and Pandit, P. V. (2010). Using zero inflated models to analyze dental caries with many zeroes. *Indian Journal of Dental Research*, 21(4), 480.

Jung, B. C., Jhun, M., and Lee, J. W. (2005). Bootstrap Tests for Overdispersion in a Zero-Inflated Poisson Regression Model. *Biometrics,* 61(2), 626-628.

Kaya, Y., ve Yeşilova, A. (2012). E-Posta trafiğinin sıfır değer ağırlıklı regresyon yöntemleri kullanılarak incelenmesi. *Anadolu University of Sciences & Technology-A: Applied Sciences & Engineering,* 13(1), 51-63.

Kılıç, D., and Bayrak, H. (2020). Sayma Verisi Modelleri Üzerine Bir Karşılaştırma: E-Ticarette Yaşanan Sorunlar Türkiye Örneği. *Selçuk Üniversitesi Fen Fakültesi Fen Dergisi*, 46(2), 85-102.

Kim, D. W., Deo, R. C., Park, S. J., Lee, J. S., and Lee, W. S. (2019). Weekly heat wave death prediction model using zero-inflated regression approach. *Theoretical and Applied Climatology*, 137(1-2), 823-838.

Kleiber C, Zeileis A (2016). Visualizing count data regressions using rootograms. *The American Statistician*, 70(3), 296–303.

Lachenbruch, P. A. (2002). Analysis of data with excess zeros. *Statistical Methods in Medical Research*, 11(4), 297-302.

Lambert, D. (1992). Zero-Inflated Poisson Regression: With an Application to Defects in Manufacturing. *Technometrics*, 34(1), 1-14.

Lemonte A.J, and Moreno- Arenas G. (2019). On residuals in generalized johnson sb regressions. *Applied Mathematical Model*, 67, 62–73.

Lemonte, A. J., Moreno-Arenas, G., and Castellares, F. (2020). Zero-inflated Bell regression models for count data. *Journal of Applied Statistics*, 47(2), 265-286.

Lord, D., Washington, S. P., and Ivan, J. N. (2005). Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis & Prevention*, 37(1), 35-46.

Majid, A., Amin, M., and Akram, M. N. (2022). On the Liu estimation of Bell regression model in the presence of multicollinearity. *Journal of Statistical Computation and Simulation*, 92(2), 262-282.

Martin, J., and Hall, D. B. (2016). R^2 measures for zero-inflated regression models for count data with excess zeros. *Journal of Statistical Computation and Simulation*, 86(18), 3777-3790.

Melkersson, M., and Rooth, D. O. (2000). Modeling female fertility using inflated count data models. *Journal of Population Economics*, 13(2), 189-203.

Miller, J., M. (2007). *Comparing poisson, hurdle, and zip model fit under varying degrees of skew and zero-ınflation*, Yayımlanmamış Doktora Tezi, University of Florida-Doctor of Philosophy.

Min, Y., and Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical modelling*, 5(1), 1-19.

Özmen, İ., and Famoye, F. (2007). Count regression models with an application to zoological data containing structural zeros. *Journal of Data Science*, 5, 491- 502.

Peng, J (2013). *Count data models for injury data from the national health interview survey*, Yüksek Lisans Tezi, The Ohio State University Graduate Program in Public Health, Columbus.

Pittman, B., Buta, E., Krishnan-Sarin, S., O'Malley, S. S., Liss, T., and Gueorguieva, R. (2020). Models for analyzing zero-inflated and overdispersed count data: an application to cigarette and marijuana use. *Nicotine and Tobacco Research*, 22(8), 1390-1398.

Ridout, M., Demétrio, C. G. B. and Hinde, J. (1998). *Models for count data with many zeros*. International Biometric Conference. Cape Town, South Africa.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 461-464.

Scudilio J, and Pereira, G.,H., A. (2020). Adjusted quantile residual for generalized linear models. *Computatinal Stat*istics, 35, 399–421.

Sezgin, F.H., ve Deniz, E. (2004). Poisson regresyon modelinde aşırı yayılım durumu ve negatif binomial regresyon analizinin Türkiye grev sayıları üzerine bir uygulaması. *İstanbul Üniversitesi İşletme Fakültesi Yönetim Dergisi*, 15(48), 17-25.

Simons, J.S., Neal, D.J. and Gaher, R.M. (2006). Risk for marijuana-related problems among college students: an application of zero-inflated negative binomial regression. *The American Journal of Drug and Alcohol Abuse*, 32 (1), 41–53.

Sumathi, K., and Rao, A. K. (2009). On estimation and tests for zero inflated regression models. *INTERSTAT*, 4.

Tukey, J., W. (1977). *Exploratory data analysis.* Addison-Wesley, Reading, MA 2,131–160.

İnternet: TUIK (2020). Web: https://data.tuik.gov.tr/Bulten/Index?p=Turkey-Health-Survey-2019-3366. Son Erişim Tarihi: 15.12.2022.

Tüzel, S (2011). *Hasar sıklıkları için sıfır yığılmalı kesikli modeler*, Yüksek Lisans Tezi, Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, Ankara.

Wang, Z., Ma, S., and Wang, C., Y. (2015). Variable selection for zero-inflated and overdispersed data with application to health care demand in Germany. *Biometrical Journal*, 57(5), 867-884.

Winkelmann, R. (2000). *Econometric Analysis of Count Data* (Third Edition). New York: Springer, 133-135.

Xu, T., Zhu, G., and Han, S. (2017). Study of depression influencing factors with zero-inflated regression models in a large-scale population survey. *BMJ open*, 7(11), e016471.

Yang, J., Xie, M. and Goh, T.N. (2009). Outlier identification and robust parameter estimation in a zero-inflated Poisson model. *Journal of Applied Statistics*, 38(2), 421-430.

Yau, K. K., Wang, K., and Lee, A. H. (2003). Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. *Biometrical Journal: journal of mathematical methods in biosciences*, 45(4), 437-452.

Yau, K. K., Lee, A. H., and Carrivick, P. J. (2004). Modeling zero-inflated count series with application to occupational health. *Computer methods and programs in biomedicine*, 74(1), 47-52.

Yesilova, A., Kaydan, M. B., and Yilmaz, K. (2010). Modeling insect-egg data with excess zeros using zero-inflated regression models. *Hacettepe Journal of Mathematics and Statistics*, 39(2), 273-282.

Zeileis, A., Kleiber, C., and Jackman, S. (2008). Regression models for count data in R. *Journal of statistical software,* 27(8), 1-25.

Zhuo, L., Stacey, K., Lawrence, J., C., Lisa, K., H., and Richard H., L., M., O. (2008). Modeling motor vehicle crashes for street racers using zero-inflated models. *Accident Analysis and Prevention*, 40, 835–839.



Gazili olmak ayrıcalıktır…