

**UKUROVA NİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

YÜKSEK LİSANS TEZİ

Gizem YILDIRIM

POİSSON VE NEGATİF BİNOM REGRESYON MODELLERİ

İSTATİSTİK ANABİLİM DALI

ADANA-2019

**ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

POISSON VE NEGATİF BİNOM REGRESYON MODELLERİ

Gizem YILDIRIM

YÜKSEK LİSANS TEZİ

İSTATİSTİK ANABİLİM DALI

Bu Tez 26/08/2019 Tarihinde Aşağıdaki Jüri Üyeleri Tarafından
Oybirliği/Oyçokluğu ile Kabul Edilmiştir.

.....
Prof. Dr. Selahattin KAÇIRANLAR
DANIŞMAN

.....
Dr. Öğr. Üyesi Selma TOKER KUTAY
ÜYE

.....
Dr. Öğr. Üyesi Yakup Murat BULUT
ÜYE

Bu Tez Enstitümüz İstatistik Anabilim Dalında hazırlanmıştır.
Kod No:

Prof. Dr. Mustafa GÖK
Enstitü Müdürü

Not: Bu tezde kullanılan özgün ve başka kaynaktan yapılan bildirişlerin, çizelge ve fotoğrafların kaynak gösterilmeden kullanımı, 5846 sayılı Fikir ve Sanat Eserleri Kanunundaki hükümlere tabidir.

ÖZ

YÜKSEK LİSANS TEZİ

POISSON VE NEGATİF BİNOM REGRESYON MODELLERİ

Gizem YILDIRIM

ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
İSTATİSTİK ANABİLİM DALI

Danışman : Prof. Dr. Selahattin KAÇIRANLAR
Yıl: 2019, Sayfa: 91
Jüri : Prof. Dr. Selahattin KAÇIRANLAR
: Dr.Öğr. Üyesi Selma TOKER KUTAY
: Dr.Öğr. Üyesi Yakup Murat BULUT

Son yıllarda, bilgi çağının bir sonucu olarak artan veri boyutuyla ihtiyaç duyulan gelişmiş tahmin modellerini içeren temel alanlardan biri olan regresyon analizi birçok disiplinde daha yaygın olarak kullanılmaktadır. Çoğunlukla sürekli bir yanıt değişkeninin tahmin edilmesi amacıyla kullanılsa da söz konusu değişkenin kesikli ve sayılabilir değerler aldığı durumda kullanılan poisson ve negatif binom regresyon modellerine ilgi gitgide artmaktadır. Bu tez çalışmasında sayım verilerinde yaygın olarak kullanılan, poisson ve negatif binom regresyon modelleri ve onların bol sıfırlı veriler için karşılık gelen modellerinin yanı sıra ridge ve Liu tahmin edici gibi yanlı tahmin edicilere dayanan modeller incelenmiştir. Bahsedilen modellerin teorik özelliklerinin yanı sıra gerçek bir veri seti üzerindeki performans karşılaştırmaları da yapılmıştır. Tez kapsamında toplam on farklı model ele alınmıştır. Bu modellerin performans karşılaştırmaları Akaike bilgi kriteri, log olabilirlik değeri, rezidü kareler ortalaması ve rezidülerin ortalama mutlak hata değerleri gibi kriterlerle yapılmıştır. Elde edilen sonuçlara göre bol sıfırlı negatif binom regresyon modeli en uygun model olarak bulunmuş ve bu modelin katsayı tahminleri elde edilmiştir. Katsayı tahminlerinin yanı sıra yorumları da detaylı olarak verilmiştir.

Anahtar kelimeler: Poisson regresyon, negatif binom regresyon, yanlı tahmin ediciler, frekans verisi

ABSTRACT

MSc. THESIS

POISSON AND NEGATIVE BINOMIAL REGRESSION MODELS

Gizem YILDIRIM

ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES
DEPARTMENT OF STATISTICS

Supervisor : Prof. Dr. Selahattin KAÇIRANLAR

Year: 2019, Pages: 91

Jury : Prof. Dr. Selahattin KAÇIRANLAR

: Asst. Prof. Dr. Selma TOKER KUTAY

: Asst. Prof. Dr. Yakup Murat BULUT

In recent years, regression analysis, which is one of the fundamental fields that includes advanced estimation models needed as with the increasing data sizes as a result of the information era, is used more extensively in many disciplines. Although regression analysis is mostly used to estimate a continuous response variable, poisson and negative binomial regression models are increasingly used when the response variable takes discrete and countable values. In this thesis study, in addition to poisson and negative binomial regression models with their corresponding models for zero inflated data, which are commonly used in count data, models based on biased estimators such as ridge and Liu estimator, have been examined. In addition to the theoretical properties of the mentioned models, performance comparisons on a real data set have been also performed. A total of ten different models have been studied within the scope of the thesis. The performance comparisons of these models have been made with criteria such as Akaike information criterion, log likelihood value, mean squared residuals, and mean absolute residuals. According to the results, zero inflated negative binomial regression model has been found as the most suitable model and coefficient estimation of this model has been obtained. Apart from the coefficient estimations, comments have been given in detail.

Key Words: Poisson regression, negative binomial regression, biased estimators, count data

GENİŞLETİLMİŞ ÖZET

Son yıllarda gelişen teknoloji toplumsal etkisinin yanı sıra bireylerin de alışkanlıklarını önemli derecede değiştirmiştir. Yaşamın her alanında teknoloji, hayatlarını kolaylaştırmak ve daha doğru kararlar alabilmelerini sağlamak amacıyla bireyleri bir takım bilgileri paylaşmaya ikna edebilmeyi başarmıştır. Gerek sosyal medya gerekse de gerçek yaşam içinde hemen hemen her alanda çeşitli teknolojik araçlar ile bireylerden çok farklı özellikte veriler toplanmakta ve saklanmaktadır. Hiç şüphesiz ki toplanan bu bilginin boyutu gün geçtikçe katlanarak artmaktadır. Sahip olunan bu veri yığınının anlamlandırılması gerek kurumlar gerekse de bireyler için bazen hayati derecede önem arz etmektedir.

Verilerin bilimsel olarak doğru ve güvenilir bir şekilde analiz edilmesinde istatistik bilimi kritik bir konuma sahiptir. Yazılım teknolojileri ile birleştirildiğinde farklı kullanım alanlarına yönelik olarak son yıllarda istatistiğe ve modellere duyulan ilgi giderek artmaktadır. Bu modeller çok çeşitli ve fazla sayıda olmasına karşın en temel alt alanlarından biri hiç şüphesiz regresyon analizidir. Birçok modeli içerisinde barındıran ve hemen hemen her disiplinden araştırmacı tarafından çeşitli amaçlarla kullanılan regresyon analizi, en genel manada bir yanıt (bağımlı) değişken üzerinde etkili olduğu düşünülen faktörlerin (bağımsız değişkenlerin) incelenmesi, yeni verilerde etkili ve tutarlı tahminler yapabilen farklı yapıda matematiksel modeller geliştirme süreci olarak tanımlanabilir.

Regresyon analizinde sürekli bir yanıt değişkenin tahmin edilmesi daha yaygın bir amaçtır. Bununla birlikte yanıt değişkenin ikili veya çok düzeyli kategorik, sayılabilir veya kesikli yapıda olduğu özel durumlar için de farklı modeller geliştirilmiştir. Bu modeller kendi içerisinde doğrusal ve genelleştirilmiş doğrusal modeller olarak iki alt gruba ayrılabilir. Doğrusal olarak ifade edilen modeller yanıt değişkenin dağılımının normal dağılımdan geldiğini varsayarken, genelleştirilmiş modeller bu değişkenin normal dağılımın yanı sıra farklı dağılımlardan geldiği varsayımıyla geliştirilmiştir.

Bu tez çalışmasında genelleştirilmiş doğrusal modeller sınıfına ait poisson ve negatif binom regresyon modelleri ele alınmıştır. Bu kapsamda Poisson (PR) ve negatif binom regresyon (NBR) modellerinin yanı sıra bu modellere dayanan bol sıfırlı verilerde kullanılan bol sıfırlı poisson (ZIPR), bol sıfırlı negatif binom (ZINBR), poisson hurdle (PH) ve negatif binom hurdle (NBHR) regresyon modelleri de incelenmiştir. Ayrıca her iki modelin ridge ve Liu gibi yanlı tahmin edicilere dayanan versiyonları da gerek teorik gerekse de uygulamada değerlendirmeye alınmıştır. Sonuç olarak tezin kapsamı toplam 10 modelle sınırlandırılmıştır.

Göz önüne alınan her bir modelin genel yapısı, teorik özellikleri ve model uyum kriterleri detaylı olarak verilmiştir. Modellerin birbiriyle ilişkisi ve farkları açıklanmaya çalışılmıştır. Değerlenen modellerin uygulama performanslarını karşılaştırmak amacıyla gerçek bir veri seti kullanılmıştır. Bu veri setinde yer alan değişkenler

- Hastanede kalınan gün sayısı,
- Bireyin kendi hakkındaki sağlık algısı,
- Kronik rahatsızlık sayısı,
- Günlük yaşamdaki kısıtlılık durumu,
- Yaş,
- Cinsiyet,
- Medeni hal,
- Eğitim durumu,
- Ailenin geliri,
- İş durumu,
- Özel sağlık sigortasının varlığı

olarak listelenebilir. Yanıt (bağımlı) değişken olarak doktorların ziyaret (vizite) sayısı değişkeni ele alınmıştır. Veri setinde toplam birey sayısı 4406'dır. Herhangi bir gözlem kayıp değer içermemektedir. Uygulama çalışmasına geçmeden önce

verinin aykırı değer içerip içermediği kontrol edilmiş ve aykırı olduğu belirlenen 224 birey işlem dışı bırakılmıştır. Sonuç olarak kalan 4182 birey üzerinde analizler yapılmıştır.

Tez içerisinde değinilen regresyon modellerini performanslarını karşılaştırmak için veri sırasıyla %70 ve %30 oranlarında eğitim ve test verisi olarak parçalanmıştır. Modellerin eğitim verisiyle eğitime işlemi yapıp, test verileriyle genelleştirme performansları ölçülmüştür. Eğitim sırasında yanlış tahmin edicilere dayanan poisson ve negatif binom regresyon modellerin dışında kalan modellerde eğitim performansı Akaike bilgi kriteri ve log olabilirlik değerleriyle değerlendirilmiştir. Aynı veri için yanlış tahmin edicilerin katsayı değerleri hesaplanmıştır. Yanlış tahmin ediciler dışında kalan modellerden en iyi bulunan modeller negatif binom regresyon, bol sıfırlı negatif binom ve negatif binom hurdle regresyon modelleri olmuştur. En zayıf model olarak ise poisson regresyon modeli bulunmuştur. Ayrıca negatif binom regresyon modellerinde aşırı yayılım parametresinin gerekli olup olmadığı belirlemek amacıyla iç içe modeller kendi aralarında log olabilirlik oran testleri ile anlamlılık testine tabi tutulmuştur. Elde edilen sonuçlara göre negatif binom regresyon modellerinde aşırı yayılım parametresinin gerekli olduğu ve bu modellerin poisson karşılıklarından anlamlı derece ayrıldığı görülmüştür. İç içe modellerin iç içe olmayan modellerde olası ikililer kendi aralarında Vuong testiyle karşılaştırılmıştır. Elde edilen Vuong test sonuçlarına göre PHR modeli PR modeliyle, ZIPR modeli PR modeliyle, ZINBR modeli NBR modeliyle, NBHR modeli NBR modeliyle karşılaştırıldığında daha tercih edilebilir modeller olarak bulunmuş ve bu modeller arasındaki farkın istatistiksel olarak anlamlı olduğu görülmüştür. Bunların dışında kalan ikili modellerden ZINBR ve NBHR ile ZIPR ve PHR model karşılaştırmalarında modeller arasında farklılık anlamlı bulunmamıştır. Model anlamlılık testlerinin yanı sıra eğitim verisinde elde edilen tahmin değerleri rootogram grafikleriyle görsel olarak değerlendirilmiştir. Rootogram grafiklerinde görsel olarak en iyi

uyumu negatif binom temelli modeller (negatif binom, bol sıfırlı negatif binom, negatif binom hurdle) göstermiştir.

Modellerin test edilmesi sürecinde daha önce eğitimde kullanılmayan ve test verisi olarak ayrılan %30'luk veri seti için oluşturulan modellerin tahmin değerleri elde edilmiştir. Elde edilen tahmin değerleri ile gerçek yanıtlar arasındaki fark olarak tanımlanan rezidü değerleri kullanılarak rezidü kareler ortalaması, rezidü kareler ortalamasının karekökü ve rezidülerin ortalama mutlak hata değeri karşılaştırma kriteri olarak kullanılmıştır. Elde edilen sonuçlara göre test verisinde en iyi performansı gösteren modeller bol sıfırlı poisson regresyon, bol sıfırlı negatif binom regresyon ve poisson hurdle regresyon bulunmuştur.

Eğitim performansı, rootogram grafikleri ve test performansının bir arada değerlendirildiğinde bol sıfırlı negatif binom regresyon modelinin kullanılan veriye en uygun model olduğu görülmüş olup bu model için tüm veri eğitilmiş ve katsayı değerleri elde edilmiştir. Elde edilen katsayıların anlamlılıkları değerlendirildiğinde sıfırdan farklı frekansları tahmin için oluşturulan modelde hastanede kalınan gün sayısı, sağlık algısı, kronik rahatsızlık sayısı, günlük yaşamda kısıtlılık durumu, yaş, medeni hal, eğitim durumu ve özel sigortası durumu anlamlı değişkenler olarak bulunmuşken, sıfır değerli frekansları tahmin amacıyla oluşturulan modelde ise kronik rahatsızlık sayısı, yaş, cinsiyet, medeni hal, eğitim durumu ve özel sağlık sigortası durumu anlamlı bulunmuştur. Uygulama bölümünün sonunda anlamlı bulunan bu değişkenlerin yorumlarına yer verilmiştir.

Sonuç olarak tez kapsamında değinilen, poisson ve negatif binom regresyon modellerine dayanan modellerin sayım verilerinde, özellikle bol sıfırlı sayım verilerinde etkin birer alternatif olabileceği görülmüştür. Özellikle tıp ve sigortacılık alanları başta olmak üzere birçok alanda kullanım alanı bulan bu yaklaşımların teorik olarak geliştirilmesi ve alternatif tahmin edicilerle desteklenmesi geleceğe ilişkin hedeflenen araştırma konuları olarak görülmektedir.

TEŞEKKÜR

Tez çalışmamın planlanmasında, araştırılmasında, yürütülmesinde ve oluşumunda ilgi ve desteğini esirgemeyen, engin bilgi ve tecrübelerinden yararlandığım, her aşamada yönlendirme ve bilgilendirmeleri ile çalışmamı şekillendiren tez danışmanım saygıdeğer hocam Prof. Dr. Selahattin KAÇIRANLAR'a, bilgi ve birikimleriyle her zaman kendilerinden çok şey öğreneceğime inandığım jüri üyeleri, Dr. Öğr. Üyesi Selma TOKER KUTAY'a ve Dr. Öğr. Üyesi Yakup Murat BULUT'a teşekkürlerimi sunarım.

Tezimin başından sonuna kadar desteğini esirgemeyen ve tezimi okuyup gereken düzeltmeleri yapma sürecinde bana büyük destek olan, bu hayattaki en büyük şansım olan değerli eşim Hasan YILDIRIM'a ve çalışmalarımda bana en büyük motivasyonu sağlayan biricik oğlum Ali Selim'e teşekkür ederim.

Son olarak, tüm hayatım boyunca bana en büyük sevgiyi ve desteği vererek her zaman yanımda olan sevgili aileme sonsuz teşekkür ederim.

İÇİNDEKİLER

SAYFA

ÖZ	I
ABSTRACT.....	II
GENİŞLETİLMİŞ ÖZET	III
TEŞEKKÜR.....	VII
İÇİNDEKİLER	VIII
ÇİZELGELER DİZİNİ	XII
ŞEKİLLER DİZİNİ	XIV
SİMGELER VE KISALTMALAR DİZİNİ	XVI
1. GİRİŞ	1
2. ÖNCEKİ ÇALIŞMALAR	7
3. GENELLEŞTİRİLMİŞ DOĞRUSAL MODELLER	11
3.1. Genelleştirilmiş Doğrusal Modellerin Genel Yapısı.....	11
3.2. Model Yeterliğinin Sınanması	18
3.2.1. Parametre Tahminlerinin Sınanması	18
3.2.1.1. β için Hipotez Testleri ve Güven Aralığı	18
3.2.1.2. Modelin Uyum İyiiliğinin Sınanması ve Sapma İstatistiği.....	19
3.2.1.2.(1). Pearson İstatistiği	19
3.2.1.2.(2). Sapma İstatistiği.....	20
3.2.1.2.(3). Akaike Bilgi Ölçütü	20
3.2.1.2.(4). Bayes Bilgi Ölçütü.....	20
4. REGRESYON ANALİZİNDE BAZI YANLI TAHMİN EDİCİLER	23
4.1. Ridge Regresyon Tahmin Edici	23
4.1.1. Ridge Regresyon Tahmin Edicinin Özellikleri.....	25
4.1.2. Ridge Regresyon Tahmin Edicinin Varyans ve Yanlılığı.....	27
4.1.3. Ridge Regresyon Tahmin Edicide Parametre Seçimi	28
4.1.4. Ridge Regresyon Tahmin Edicinin Performansının Değerlendirmesi.....	29

4.2. Liu Regresyon Tahmin Edici	30
5. POISSON REGRESYON ANALİZİ	31
5.1. Poisson Regresyon Modeli	31
5.2. Bol Sıfırlı Poisson Regresyon Modeli	32
5.3. Bol Sıfırlı Genelleştirilmiş Poisson Regresyon Modeli	33
5.4. En Çok Olabilirlik Tahmin Edici (MLE)	33
5.5. Poisson Regresyon Katsayılarının Yorumlanması	36
5.6. Poisson Regresyonda Bazı Yanlı Tahmin Ediciler	36
5.6.1. Poisson Ridge Tahmin Edici	36
5.6.1.1. Önerilen Ridge Parametresi Tahmin Yaklaşımları	38
5.6.2. Bol Sıfırlı Modellerde Poisson Ridge Tahmin Edici	40
5.6.2.1. Önerilen Ridge Parametresi Tahmin Yaklaşımları	42
5.6.3. Modified Jackknifed Poisson Ridge Tahmin Edici	44
5.6.3.1. Önerilen Ridge Parametresi Tahmin Yaklaşımları	46
5.6.4. Poisson Liu Tahmin Edici	47
5.6.5. Liu Tahmin Edicide Yanlılık Parametresinin Seçimi	47
5.7. Log Normal Poisson Regresyon Modeli	48
5.8. Poisson Hurdle Regresyon Modeli	49
5.9. Poisson Regresyonda Uyum İyiliği Testleri	51
5.9.1. Sapma İstatistiği (G^2)	51
5.9.2. Pearson Ki-Kare İstatistiği (χ^2)	51
5.9.3. Yapay R^2 İstatistiği	54
5.9.4. Akaike Bilgi Ölçütü	54
5.9.5. Pearson İstatistiği	54
5.10. Tahmin Edicilerin Performanslarının Değerlendirilmesi	54
6. NEGATİF BİNOM REGRESYON ANALİZİ	57
6.1. Negatif Binom Regresyon Modeli	57
6.1.1. Doğrusal Varyans Fonksiyonu ile Negatif Binom Dağılımı (NB1)	58

6.1.2. Karesel Varyans Fonksiyonu ile Negatif Binom Dağılımı (NB2).....	59
6.2. En Çok Olabilirlik Tahmin Edici (MLE).....	59
6.2.1. NB1 Model ve En Çok Olabilirlik Tahmini	59
6.2.2. NB2 Model ve En Çok Olabilirlik Tahmini	60
6.3. Negatif Binom Regresyonda Bazı Yanlı Tahmin Ediciler.....	61
6.3.1. Negatif Binom Ridge Tahmin Edici	61
6.3.1.1. Negatif Binom Ridge Tahmin Edicide k Parametresinin Seçimi	62
6.3.2. Negatif Binom Liu Tahmin Edici	63
6.3.2.1. Negatif Binom Liu Tahmin Edicide Parametre Seçimi.....	64
6.3.3. İki Parametrelili Negatif Binom Tahmin Edici.....	65
6.3.3.1. İki Parametrelili Negatif Binom Tahmin Edicide Parametrelerin Seçimi.....	66
6.4. Lognormal ve Gamma Mixed Negatif Binom Regresyon Modeli.....	68
6.5. Negatif Binom Hurdle Regresyon Modeli	69
7. UYGULAMA ÇALIŞMASI.....	71
7.1. Verinin Yapısı ve Kökeni	71
7.2. Eğitim Verisinin Modellenmesi	74
7.3. Test Verisi Kullanılarak Performans Karşılaştırması	77
7.4. En İyi Modelin Belirlenmesi ve Yorumlanması	78
8. SONUÇ VE ÖNERİLER.....	81
KAYNAKLAR	83
ÖZGEÇMİŞ	91



ÇİZELGELER DİZİNİ

SAYFA

Çizelge 7.1. Eğitim modellerinin AIC ve LL değerleri.....	75
Çizelge 7.2. Eğitim modellerinde yürütülen ikili karşılaştırmaların LRT ve Vuong test sonuçları.....	76
Çizelge 7.3. Test verisinde modellerin performans sonuçları	78
Çizelge 7.4. Bol sıfırlı negatif binom regresyon (ZINBR) uyumu sonuçları	79





ŞEKİLLER DİZİNİ

SAYFA

Şekil 1.1.	Yıllara göre poisson regresyon konusunda yapılan çalışma sayıları	4
Şekil 1.2.	Yıllara göre negatif binom regresyon konusunda yapılan çalışma sayıları	5
Şekil 7.1.	Bütün, eğitim ve test verilerinde vizite sayılarının frekans grafikler	72
Şekil 7.2.	Bireylerin sağlık algılarına ve hastanede kalınan gün sayılarına göre vizite sayılarının frekans grafikleri.....	73
Şekil 7.3.	Bireylerin özel sigorta durumları ve cinsiyetlerine göre vizite sayılarının frekans grafikleri.....	74
Şekil 7.4.	Eğitim verisinde kullanılan modellerin rootogram grafikleri.....	77



SİMGELER VE KISALTMALAR

X	: Bağımsız değişkenler kümesi
Y	: Negatif değerli olmayan tamsayı değerli bağımlı (yanıt) değişken
ε	: Hata terimi
β	: Bilinmeyen parametre vektörü
η	: Doğrusal ön kestiricidir
θ	: Yayılım parametresi
μ	: Bağımlı değişken ortalama değeri
σ^2	: Varyans
n	: Örneklem büyüklüğü
N	: Kitle büyüklüğü
W	: Ağırlık matrisi
z	: β 'nin bir fonksiyonu
L	: Olabilirlik değeri
$S_{\hat{\beta}_j}$	: $\hat{\beta}_j$ 'nin standart hatası
$\hat{\beta}_j$	: β_j 'nin tahmin edici
G^2	: Sapma istatistiği
$\hat{\mu}$	: μ 'nün tahmin edici
p	: Modelin parametre sayısı
q	: Modeldeki açıklayıcı değişken sayısı
β_0	: Doğrusal regresyon modelindeki sabit parametre
β_1	: Doğrusal regresyon modelindeki eğim parametresi
d_i	: i.gözleme ait sapma artışı
R^2	: Belirleme katsayısı
AHKT	: Ağırlıklı Hata Kareler Toplamı
AIC	: Akaike Information Criterion (Akaike Bilgi Ölçütü)
BIC	: Bayesian Information Criterion (Bayes Bilgi Ölçütü)

EKK : En Küçük Kareler
IWLS : Iteratively Weighted Least Squares (Adımsal Ağırlıklı En Küçük Kareler)
JRR : Jackknifed Ridge Regresyon
LGNB : Lognormal Gamma Mixed Negatif Binom Regresyon Modeli
LIU : Lineer Birleştirilmiş Regresyon Tahmin Edici
MAE : Mean Absolute Error (Ortalama Mutlak Hata)
MJPR : Modified Jackknifed Poisson Regresyon
MJRR : Modified Jackknifed Ridge Regresyon
MLE : Maksimum Likelihood Estimator (En Çok Olabilirlik Tahmin Edici)
MSE : Mean Square Error (Hata Kareler Ortalaması)
NBHR : Negatif Binom Hurdle Regresyon
NBLR : Negatif Binom Liu Regresyon
NBR : Negatif Binom Regresyon
NBRR : Negatif Binom Ridge Regresyon
PHR : Poisson Hurdle Regresyon
PLR : Poisson Liu Regresyon
PR : Poisson Regresyon
PRR : Poisson Ridge Regresyon
RR : Ridge Regresyon
VIF : Variance Inflation Factor (Varyans Şişirme Faktörü)
WMSE : Weighted Mean Square Error (Ağırlıklandırılmış Hata Kareler Ortalaması)
ZIGPR : Zero-Inflated Generalized Poisson Regression (Bol Sıfırlı Genelleştirilmiş Poisson Regresyon)
ZINBR : Zero-Inflated Negatif Binom Regression (Bol Sıfırlı Negatif Binom Regresyon)
ZIPR : Zero-Inflated Poisson Regression (Bol Sıfırlı Poisson Regresyon)
ZIPRR : Zero-Inflated Poisson Ridge Regression (Bol Sıfırlı Poisson Ridge Regresyon)

1. GİRİŞ

Modelleme ilgilenilen bir rassal değişkenin değişimini tanımlayan matematiksel ifadeler geliştirme süreci olarak tanımlanır. Bu değişken örneğin bir ürünün fiyat tahmini olabileceği gibi bazen de bir büyüme oranı olabilir. Araştırma alanına göre ilgi odağında olan değişken değişebileceği gibi bu değişkene bağlı modelin de yapısı doğal olarak değişim gösterecektir. Regresyon analizi farklı yapıda modeller kurmaya imkân tanıyan ve istatistik biliminin en köklü çalışma alanlarından biridir. Uzun yıllardır çok farklı disiplinlerden araştırmacılar bu alanda yoğun olarak çalışma yürütmektedirler. Değişkenler arası ilişkileri görmeye, bu ilişkilere bağlı farklı yapıda modeller oluşturmaya ve tahminler yapmaya imkan tanıyan regresyon analizi günümüzde hala popülerliğini korumaktadır.

Gelişen teknoloji, artan veri miktarı ve beraberinde ortaya çıkan veriyi anlamlandırma isteği birçok disiplinde olduğu gibi istatistik biliminde de olumlu etkiler doğurmuştur. Çok sayıda ve farklı yapıda ortaya çıkan değişkenler arasında ilişki kurmanın önemi daha da artmıştır. Bununla birlikte klasik modelleme yöntemlerine alternatif olarak yeni modeller kullanmak zorunlu bir ihtiyaç halini almıştır.

Bir araştırmada araştırmanın odağı olan, değişiminin kaynağı araştırılmak istenen değişken olarak tanımlanan bağımlı değişken ile bu değişken üzerine etkisi olduğu düşünülen değişken olarak tanımlanan bir veya daha çok bağımsız değişken arasındaki bağıntıyı incelemek amacıyla kullanılan yöntemle regresyon analizi denir. Bir diğer deyişle değişkenler arasında fonksiyonel ilişkileri araştırmak için kullanılan kavramsal olarak basit bir metottur (Chatterjee ve ark., 2006). Regresyon analizi bu ilişkileri incelerken değişkenler arasında sebep-sonuç bağlamında bir yaklaşım kullanır. Daha önce bahsedilen ihtiyaçlardan ötürü bu alanda sürekli, kesikli veya kategorik yapıda olabilen değişkenlerle çalışmayı mümkün kılan çok sayıda regresyon modelleri ortaya koyulmuş ve başarılı bir şekilde uygulanmıştır. Bu uygulamalar çok farklı disiplinlerde görülebilmektedir.

Regresyon analizinin başlıca kullanım alanları aşağıda listelenmiştir (Chatterjee ve ark., 2006; Wikipedia):

- Biyoloji
- Tıp
- Ekonomi
- Fizik
- Kimya
- Zirai bilimler
- İşletme
- Pazarlama
- İstatistik
- Tarih
- Politika
- Çevre bilimleri
- Sosyal bilimler (sosyoloji, eğitim, psikoloji, antropoloji,)
- Eczacılık
- Bilgisayar bilimleri
- Astronomi ve yer bilimleri
- Genetik

Artan veri miktarı ve karşılaşılan farklı yapıda değişkenler farklı regresyon modelleriyle fonksiyonel olarak ilişkilendirilir. Modelleme yapılırken bağımlı değişkenin yapısı dikkate alınmaktadır. Bu değişken sürekli olduğunda klasik en küçük kareler (EKK) yöntemi ile modellenenebilir. Bu gibi durumlarda bağımsız değişken yapısına bağlı olarak tek değişken olması durumunda basit, iki veya daha fazla değişken olması durumunda ise çoklu doğrusal regresyon analizi başarılı bir şekilde uygulanabilir. Benzer şekilde bağımlı değişken yapısı kategorik ise düzey sayısına bağlı olarak iki düzeyli durumda ikili lojistik regresyon analizi, ikiden

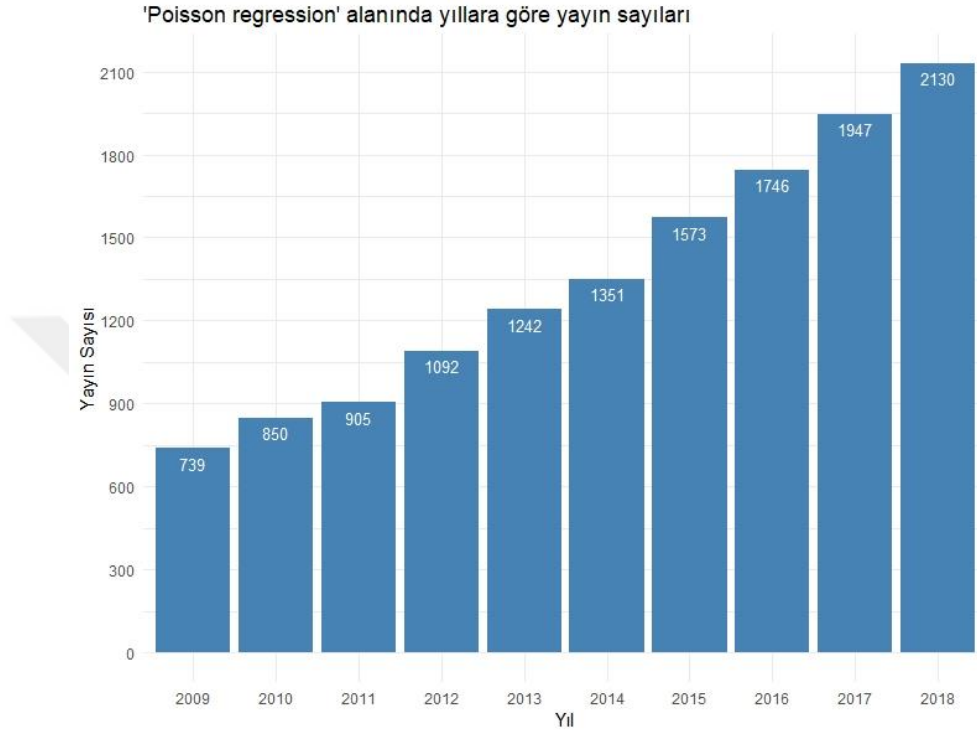
fazla düzey olması durumunda ise çok düzeyli lojistik regresyon analizi ile değişkenler arasındaki fonksiyonel ilişki araştırılabilir.

Yapılan çalışmada bağımlı değişkenin bir sayım veya frekans verisi olması durumunda ise bahsedilen modeller uygulanamaz. Bağımlı değişkenin, olasılık dağılımının negatif olmayan tamsayılardan oluşan sayım verisi tanımlayan kesikli bir dağılım olması durumunda poisson regresyon analizi uygulanır. Bu analiz örneğin bir grup insanın hastanede kaldıkları gün sayısını tahmin etmek veya bir afet sırasında birtakım değişkenlere bağlı ölüm sayısını modellemek amacıyla kullanılabilir (Chatterjee ve ark., 2006). Bu modelin kullanılabilmesi için bağımlı değişkenin dağılımının kesikli bir dağılım olan Poisson dağılımına sahip olması gerekir. Olayların gerçekleşme sayıları çoğu zaman verinin en yaygın görülen formudur. Bu nedenle genellikle belirli bir zaman diliminde gerçekleşen olay sayısı bu model aracılığı ile tahmin edilebilir. Poisson regresyon modelinin kullanıldığı alanlar şu şekilde listelenebilir (Dobson ve Barnett, 2008; Chatterjee ve ark., 2006; Wikipedia):

- Tıp – Biyoistatistik
- Astronomi
- Biyoloji
- İşletme ve Yönetim
- Finans ve Sigortacılık
- Demografi
- Yer bilimleri
- İletişim

Thomson Reuters tarafından hazırlanan SCI, SCI-Expanded, SSCI ve AHCI indekslerince taranan çalışmalara ulaşılabilen ISI Web of Knowledge SM (apps.isiknowledge.com) arama sayfasında arama kutucuğuna “poisson regression” terimi yazılıp yıllara göre arama yapılarak 2009 ile 2018 yılları arasında içerisinde

“poisson regression” kavramı geçen çalışma sayıları Şekil 1.1’de verilmiştir. Grafikten konuya ilişkin çalışma sayısının her yıl artış gösterdiği görülmektedir.



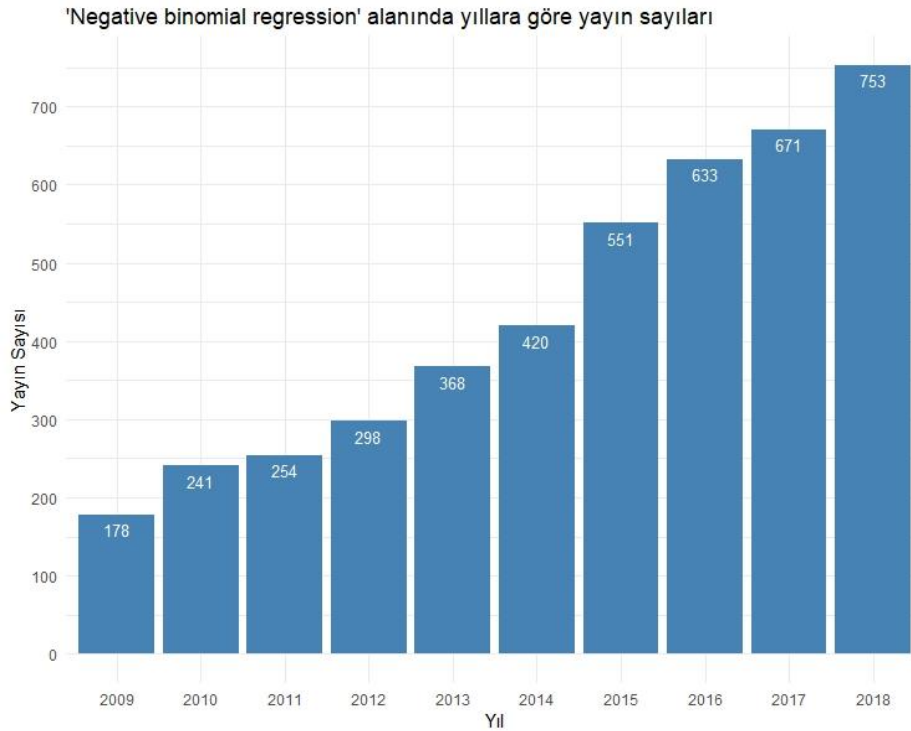
Şekil 1.1. Yıllara göre poisson regresyon konusunda yapılan çalışma sayıları

Poisson regresyonda bağımlı değişkenin dağılımı poisson dağılımına sahiptir. Bu dağılımın karakteristik özelliği ortalamasının varyansına eşit olmasıdır. Ancak gerçek yaşam verilerinde bu durum sıklıkla ihlal edilir. Genellikle ölçüm değerlerinden varyans ortalamadan daha büyük olarak gözlenir. Bu durum aşırı yayılım olarak adlandırılır ve klasik poisson regresyon modeli üzerinde bozucu bir etkiye sahip olur. Bu nedenle poisson regresyonu doğrudan kullanmak hatalı olduğundan aşırı yayılım sorununun çözüm yollarından biri olarak negatif binom dağılımına dayanan negatif binom regresyon modeli kullanılmaktadır.

Gerçek yaşam verilerinde yaygın olarak gözlenen bir diğer problem ise aşırı sayıda sıfır değerli ölçüm olmasıdır. Bu durumda da negatif binom regresyon

modeli kullanılabilir. Yaygın olarak bir olayın ya hiç gözlenmiyor ya da en az bir defa gözleniyor olduğu durumlarda kullanılmaktadır. Örneğin toplumda bireylerin günlük hayatta sigara tüketimi üzerine yapılan bir araştırmada hiç tüketmeyen bireylerin çoğunlukta olabileceği ve bağımlı değişkenin çok kez sıfır değerini alabileceği ihtimalinden ötürü negatif binom regresyon modeli kullanılabilir.

Negatif binom regresyon analizi üzerine yapılan çalışma sayılarının değişimini incelemek amacıyla Thomson Reuters tarafından hazırlanan SCI, SCI-Expanded, SSCI ve AHCI indekslerince taranan çalışmalara ulaşılabilen ISI Web of Knowledge SM (apps.isiknowledge.com) arama sayfasında arama kutucuğuna “negative binomial regression” terimi yazılıp yıllara göre arama yapılarak 2009 ile 2018 yılları arasında içerisinde “negative binomial regression” kavramı geçen çalışma sayıları Şekil 1.2’de verilmiştir. Grafikten bu konuya ilişkin çalışma sayısının her yıl artış gösterdiği görülmektedir.



Şekil 1.2. Yıllara göre negatif binom regresyon konusunda yapılan çalışma sayıları

Poisson ve negatif binom regresyon modelleri sayım verilerini modellemek amacıyla yoğun olarak kullanılmakta olan modellerdir. Farklı özellikte veri yapılarında bir model diğerinden daha iyi performans ortaya koyabilmektedir. Bu tez çalışmasında da bu modeller gerçek veri üzerinde test edilerek farklı performans ölçütlerine göre karşılaştırılmıştır.

Tez sekiz bölümden oluşmaktadır. Tez konusuyla ilgili kavramsal ve genel bilgilerin verildiği bu giriş bölümünün ardından ikinci bölümde literatürde yer alan çalışmalar verilmiştir.

Üçüncü bölümde ise genelleştirilmiş doğrusal modellerin genel yapısı, parametre tahminleri ve uyum iyiliği ölçülerine yer verilmiştir. Regresyon analizinde yaygın olarak kullanılan ridge ve Liu regresyon tahmin edicileri en genel haliyle dördüncü bölümde özetlenmiştir. Beşinci bölümde ise poisson regresyon analizi ile ilgili kavramlar ve tanımlara yer verilip model yapısı ve parametre tahmin yöntemlerine değinilip poisson regresyonu üzerinde uygulanan ridge ve Liu yanlı tahmin edicilere yer verilmiştir. Benzer yaklaşımla negatif binom regresyon analizi ile ilgili model yapısı, parametre tahminleri ve yanlı tahmin edicilerin özellikleri altıncı bölümde incelenmiştir.

Tezin yedinci bölümünde ise gerçek veri kullanılarak anlatılan tahmin edicilere bağlı modeller uydurulup, performansları karşılaştırmalı olarak ele alınmıştır. Elde edilen sonuçlara bağlı yorumlara ve gelecek çalışmalarla ilgili önerilere tezin sekizinci bölümünde yer verilmiştir.

2. ÖNCEKİ ÇALIŞMALAR

Tahmin yapmak birçok bilimsel disiplinde modellemenin öncelikli amaçlarından biridir. Etkin olarak modelleme ve dolayısıyla tahmin yapmaya imkan tanıyan araştırma alanlarından biri de regresyon analizidir. Temel amacı rastlantısal olarak değişen bir odak değişkenini, bu değişken üzerinde etkili olduğu düşünülen bir takım bağımsız değişkenlerle matematiksel bir model içerisinde incelemek ve tahmin etmektir. Bu model aracılığı ile değişkenler arası ilişkilerin yönü, şiddeti ve bağımlı değişken üzerinde ki bireysel etkileri incelenebilir. Regresyon analizinin sayım verilerinde uygulamalarını içeren alt dalı poisson regresyon analizidir. Poisson regresyon analizinin yanı sıra veride gözlenen varyansın ortalamadan büyük olduğu yani aşırı yayılım durumunda uygulanan regresyon analizi türü ise negatif binom regresyon analizidir. Söz konusu iki regresyon türü ile ilgili literatürde yer alan çalışmalardan bazıları aşağıda verilmiştir.

Poisson dağılımı ilk defa Siméon-Denis Poisson (1781–1840) tarafından diğer olasılık hakkındaki yazıları ile birlikte 1838'de yayınlanan *Recherches sur la probabilité des jugements en matières criminelles et matière civile* ("Ceza hukuku ve medeni hukuk alanlarındaki hükümlerin olasılığı üzerinde araştırmalar") adındaki eserinde ortaya atılmıştır (Wikipedia).

Wang ve Famoye (1997), çalışmalarında hane halkı doğurganlık kararının modellenmesinde standart poisson regresyon ve genelleştirilmiş poisson regresyon modellerini kullanmışlardır. En iyi modelin belirlenmesi için poisson regresyon modellerini karşılaştırmışlardır.

Selim ve Üçdoğruk (2003), "Sayma veri modelleri ile çocuk sayısı belirleyicileri: Türkiye'deki seçilmiş iller için sosyoekonomik analizler" başlıklı çalışmasında poisson quasi en çok olabilirlik yöntemini kullanmışlardır. Bu yöntemle çocuk sayısı belirleyicilerini modellemeye çalışmışlardır.

Heinzl ve Mittlböck (2003) aşırı ya da az yayılım gösteren poisson regresyon modeli için yapay R^2 ölçütlerini önermişlerdir.

Sezgin ve Deniz (2004) Türkiye grev sayılarına etki eden faktörleri bulmak amacıyla aşırı yayılım modeli ve negatif binom regresyon modeli ile bir takım analizler yapmışlardır.

Yeşilyurt (2005), “Poisson regresyon modeli ve Türkiye’deki boşanma istatistiklerine uygulanması” başlıklı yüksek lisans tezinde ise yanıt değişkeni olarak 1989 ile 1998 yılları arasında Türkiye’deki boşanma sayısı verilerini kullanmış olup, yıl ve yaş değişkenleri ise açıklayıcı değişkenler olarak ele almıştır. Uygulamada aşırı yayılım durumu ile karşılaşmış olduğundan ağırlıklandırma parametresi kullanılarak bazı değerlerde düzeltmeler yoluna gidilmiştir.

Köleoğlu (2006), doktora tezinde olay zamanı analizinde kullanılan poisson regresyon modeli, poisson regresyon modelinde aşırı yayılımla karşılaşıldığında kullanılan Poisson-gamma regresyon modeli ve olay zamanı verilerinde gözlemlenemeyen heterojenlikle karşılaşıldığında kullanılan tesadüfi etkiler Poisson-gamma regresyon modelini incelemiştir.

Famoye ve Singh (2006) aile içi şiddet verilerine eğitim, iş, gelir, ilaç kullanımı gibi bağımsız değişkenleri kullanarak bol sıfırlı genelleştirilmiş poisson regresyon modelini uygulamışlardır.

Yeşilova ve Atlıhan (2007), “Farklı sıcaklıkların *Scymnus Subvillosus*’un bıraktığı yumurta sayıları üzerine etkilerinin karışımı Poisson regresyon ile analiz edilmesi” başlıklı çalışmalarında *Scymnus Subvillosus* dişilerinin oviposizyon sürecinde bıraktıkları yumurta sayılarını Poisson regresyon modeli ile analiz etmişlerdir.

Fallah ve ark. (2009), yapay sinir ağı kullanarak doğrusal olmayan poisson regresyon analizi üzerine çalışmalar yürütmüşlerdir.

Yeşilova ve ark. (2011) sıfır değer ağırlıklı genelleştirilmiş poisson regresyonu yardımıyla Van Gölü’nde suda yüzen yeşil böcekler, 1909’in

popülasyon değişimi üzerinde fiziko-kimyasal çevresel koşullarının etkilerinin araştırılması konulu bir çalışma yapmıştır.

Bandyopadhyay ve ark. (2011) madde bağımlılığı çalışmalarında bol sıfırlı veriler için geliştirilen bol sıfırlı regresyon modellerini karşılaştırmışlar ve bol sıfırlı hurdle modellerini en iyi tahmini veren modeller olarak önermişlerdir.

Buu ve ark. (2011) madde bağımlılığı araştırmalarında önemli olabilecek değişkenlerin belirlenmesi amacıyla bol sıfırlı poisson modelinde değişken seçim yaklaşımları önermişlerdir.

Mouatassim ve Ezzahid (2012) özel sağlık sigortası verilerinde yaş, medeni hal, cinsiyet gibi bağımsız değişkenler kullanarak poisson ve bol sıfırlı poisson regresyon modellerini uygulamışlardır.

Ghitany ve ark. (2012) çalışmalarında çok değişkenli karışımli poisson regresyon modellerini sınıflandırmak için en çok olabilirlik tahminini kolaylaştırmada genel bir EM algoritması önermişlerdir.

Xie ve ark. (2013) sigara alışkanlığı ile ilgili çalışmalarında poisson ve negatif binom regresyon modellerinin yanı sıra bol sıfırlı versiyonlarını ve hurdle modelleri kullanarak istatistiksel bir analiz çalışması yapmışlardır. Elde edilen sonuçlara göre negatif binom regresyon modelinin de bol sıfırlı modeller karşısında başarılı olabileceğini göstermişlerdir.

Avşar (2014) Türkiye’de yerel sanayi üretim ve ithalat verilerini kullanarak antidamping soruşturma sayısını etkileyen faktörleri negatif binom regresyon modeli ile araştırmıştır.

Liyanage ve ark. (2015) dünya genelince böbrek nakli gereken birey sayısını tahmin etmek ve hastalığın yaygınlığını araştırmak amacıyla poisson regresyon modelini kullanmışlardır. Elde ettikleri bulguları kullanarak dünya genelinde 2030 yılına kadar oluşabilecek ihtiyaca ilişkin bir projeksiyon sunmuşlardır.

Martinez ve ark. (2015) özel bir ilaç türünün şiddetli koah hastaları üzerindeki etkilerini araştırmak amacıyla poisson ve negatif binom regresyon

modellerini kullanmışlardır. Elde edilen sonuçları çeşitli performans kriterlerine göre karşılaştırmalı olarak sunmuşlardır.

Escario ve Wilkinson (2015) poisson, negatif binom, bol sıfırlı poisson ve bol sıfırlı negatif binom regresyon modellerini kullanarak sigara içme alışkanlığının nesiller arasındaki geçişgenliğini incelemiş ve bulguları yorumlamışlardır.

Oliveira ve ark. (2016) radyasyon maruziyeti sonrası kromozom bozulma sayılarını tahmin amaçlı olarak bol sıfırlı poisson ve bol sıfırlı negatif binom regresyon modellerini kullanmışlardır.

Tang ve ark. (2017) depresyonu etkileyen faktörleri araştırmak amacıyla geniş ölçekli bir kitle üzerinde bol sıfırlı regresyon modellerini araştırmış ve bol sıfırlı negatif binom regresyon modelinin en iyi tahminleri üreten model olduğunu bulmuşlardır.

Faroughi ve İsmail (2017) iki değişkenli ve bol sıfırlı verilere özgü olarak önerilen bol sıfırlı negatif binom modellerini incelemiş, iç içe bol sıfırlı negatif binom regresyon modelini önermiş ve uygulamalarına değinmişlerdir.

Avcı (2018) şizofren bireylerin hastane yatış sayılarını etkileyen faktörleri aşırı yayılım durumunda önerilen negatif binom ve Conway-Maxwell-Poisson (CMO-Poisson) regresyon modellerini karşılaştırmalı olarak incelemiştir.

Chai ve ark. (2018) gemi batma kazalarını modellemek amacıyla bol sıfırlı negatif binom regresyon modelini kullanmış ve elde ettikleri sonuçları detaylı olarak yorumlamışlardır.

3. GENELLEŞTİRİLMİŞ DOĞRUSAL MODELLER

3.1. Genelleştirilmiş Doğrusal Modellerin Genel Yapısı

Doğrusal bir model $E(\varepsilon_i) = 0$ olmak üzere

$$y_i = x_i' \beta + \varepsilon_i, \quad i = 1, 2, \dots, n \quad (3.1)$$

olarak yazılabilir. Burada $x_i' = [1, x_{i1}, x_{i2}, \dots, x_{ip}]$, $(p + 1) \times n$ boyutlu i . kümeye ilişkin satır vektörü, $\beta' = [\beta_0, \beta_1, \dots, \beta_p]$, $1 \times (p + 1)$ boyutlu bilinmeyen parametreler ve y_i ise üstel dağılım ailesinden rasgele değişkendir. Genelleştirilmiş doğrusal modeller 3 temel bileşene sahiptir. Bunlar üstel aile formunda yazılabilir olması, sistematik bileşen içermesi ve bağlantı fonksiyonuna sahip olmasıdır.

1) Üstel dağılım ailesinin olasılık yoğunluk fonksiyonunun genel formu $\mu_i = E(y_i)$ ve $g(\mu_i) = x_i' \beta$ olmak üzere $\eta_i = x_i' \beta$ link fonksiyonu olarak adlandırılmak üzere,

$$f(y_i, \Phi_i, \theta) = \exp \left(\frac{y_i \Phi_i - b(\Phi_i)}{a(\theta)} + k(y_i, \theta) \right) \quad (3.2)$$

şeklinde verilmektedir (Montgomery ve ark, 2012; Agresti, 2015). Burada

- $a(\cdot)$, $b(\cdot)$ ve $k(\cdot)$ bir üstel aileden diğerine değişen fonksiyonlardır.
- θ , yayılım parametresidir.
- Φ_i , doğrusal ön kestiricidir.

Bu genel form kullanılarak üstel ailenin üyesi olan dağılımlardan poisson ve bernoulli dağılımları için olasılık yoğunluk fonksiyonlarının özel durumları yazılabilir. Poisson dağılımının olasılık yoğunluk fonksiyonu,

$$f(y_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} = \exp(y_i \ln \lambda_i - \lambda_i - \ln(y_i!)) \quad (3.3)$$

şeklinde üstel formda yazılabilir. (3.2)'de yer alan olasılık yoğunluk fonksiyonunun genel formunu oluşturan terimler, (3.3)'de,

$$\Phi_i = \ln \lambda_i \text{ ve } \lambda_i = e^{\Phi_i}$$

$$b(\Phi_i) = \lambda_i$$

$$a(\theta) = 1$$

$$k(y_i, \theta) = -\ln(\lambda_i!)$$

olur. Bernoulli dağılımı için ise olasılık yoğunluk fonksiyonu,

$$\begin{aligned} f(y_i) &= \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \\ &= \exp(y_i \ln \pi_i + (1 - y_i) \ln(1 - \pi_i)) \\ &= \exp(y_i \ln \pi_i - \ln(1 - \pi_i) - y_i \ln(1 - \pi_i)) \\ &= \exp\left(y_i \ln\left(\frac{\pi_i}{1 - \pi_i}\right) + \ln(1 - \pi_i)\right) \end{aligned} \quad (3.4)$$

şeklinde üstel formda yazılabilir (Montgomery ve ark,2012; Dobson ve Barnett, 2008).(3.2)'de yer alan olasılık yoğunluk fonksiyonunun genel formunu oluşturan terimler, (3.4)'de

$$\begin{aligned} \Phi_i &= \ln\left(\frac{\pi_i}{1 - \pi_i}\right) \\ \pi_i &= \frac{e^{\Phi_i}}{1 + e^{\Phi_i}} \\ b(\Phi_i) &= -\ln(1 - \pi_i) \\ a(\theta) &= 1 \\ k(y_i, \theta) &= 0 \end{aligned}$$

olarak tanımlanır.

Genelleştirilmiş doğrusal modellerde olabilirlik fonksiyonu,

$$L(\beta) = \prod_{i=1}^n \left(\frac{y_i \Phi_i - b(\Phi_i)}{a(\theta)} + k(y_i, \theta) \right) \quad (3.5)$$

şeklinde gösterilir. n birimlik örnekler için $l = l(y_i; \Phi_i, \theta) = \ln L(y_i; \Phi_i, \theta)$ olmak üzere log olabilirlik fonksiyonu ise

$$l(\beta) = \sum_{i=1}^n l_i(\beta) = \sum_{i=1}^n \left(\frac{y_i \Phi_i - b(\Phi_i)}{a(\theta)} + k(y_i, \theta) \right) \quad i = 1, 2, \dots, n \quad (3.6)$$

şeklinde gösterilir.

y 'nin dağılımı üstel ailenin bir dağılımına üyesi olduğundan l 'nin Φ_i 'ye göre birinci türevinin alınmasıyla ortalama

$$\frac{\partial l}{\partial \Phi_i} = \frac{y_i - b'(\Phi_i)}{a(\theta)} \quad (3.7)$$

ve 2. türevin alınmasıyla da varyans

$$\frac{\partial^2 l}{\partial \Phi_i^2} = \frac{-b''(\Phi_i)}{a(\theta)} \quad (3.8)$$

elde edilir.

Eşitlik (3.7)'yi sıfıra eşitleyerek, olabilirlik fonksiyonunun eşitliğinin $a(\theta)$ 'dan bağımsız olduğu gözlemlenir. $\Phi = x'\beta$ 'deki Φ ve β 'nin tahminine uygunluğu ile ilgilenildiği için $a(\theta) = 1$ varsayılmaktadır.

Fisher skorlama ile,

$$E\left(\frac{\partial l}{\partial \Phi_i}\right) = 0 \quad (3.9)$$

$$-E\left(\frac{\partial^2 l}{\partial \Phi_i^2}\right) = E\left(\frac{\partial l}{\partial \Phi_i}\right)^2 \quad (3.10)$$

eşitlikleri elde edilir ve Eşitlik (3.7) ile Eşitlik (3.9) birleştirilirse

$$E\left(\frac{y_i - b'(\Phi_i)}{a(\theta)}\right) = 0 \quad (3.11)$$

$$E(y_i) - b'(\Phi_i) = 0 \quad (3.12)$$

elde edilir. Eşitlik (3.12)'den

$$E(y_i) = b'(\Phi_i) = \mu_i \quad (3.13)$$

ile genelleştirilmiş doğrusal modeller için beklenen değerin genel formu bulunmuş olur. Benzer olarak eşitlik (3.8) ve (3.10) birleştirilirse

$$E\left(\frac{-b''(\Phi_i)}{a(\theta)}\right) = E\left(\frac{y_i - b'(\Phi_i)}{a(\theta)}\right)^2 \quad (3.14)$$

$$\frac{-b''(\Phi_i)}{a(\theta)} = \frac{V(y_i)}{a(\theta)^2} \quad (3.15)$$

$$V(y_i) = a(\theta)b''(\Phi_i) = \frac{\partial E(y_i)}{\partial \eta_i} = \frac{\partial \mu_i}{\partial \Phi_i} \quad (3.16)$$

genelleştirilmiş doğrusal modeller için varyansın genel formu elde edilmiş olur (Agresti, 2015). Eşitlik (3.16)'da varyansın, beklenen değerin bir fonksiyonu olduğu; ancak doğrusal regresyon varsayımlarından olan sabit varyanslılık koşulunu sağlamadığı görülmektedir.

2) Sistemik bileşen: Modelin ortalama parametresinin hesaplanmasında kullanılan doğrusal ön kestirim, bağımsız değişkenlerin doğrusal kombinasyonu,

$$\eta_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_p X_{pi} = \sum_{i=1}^n x_i \beta \quad (3.17)$$

biçiminde ifade edilmektedir (Montgomery ve ark,2012; Dobson ve Barnett, 2008).

3) Rasgele ve sistemik bileşen arasındaki bağ fonksiyonu ($g(.)$), bağımlı değişkenin beklentisini ifade eder. $\mu_i = E(y_i)$ için doğrusal ön -kestirici,

$$g(\mu_i)=\eta_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_p X_{pi} \quad (3.18)$$

biçiminde tanımlanır. Böylece bağ fonksiyonu y 'nin ortalaması ile η_i 'yi ilişkilendirir. Genel olarak η_i, μ_i 'nin bir fonksiyonudur.

$$g(\mu_i)=\eta_i \quad (3.19)$$

Bağ fonksiyonu tersinir bir fonksiyon olduğu için,

$$\mu_i = g^{-1}(\eta_i) = g^{-1}(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_p X_{pi}) \quad (3.20)$$

şeklinde yazılabilir (Montgomery ve ark,2012).

Böylece genelleştirilmiş doğrusal modeller, bağımlı değişken için bir doğrusal olmayan regresyon modeli olarak veya beklenen bir bağımlı değişkenin dönüşümü için ise, bir doğrusal model olarak düşünülmektedir. Bu ters bağa $g^{-1}(.)$ ortalama fonksiyonu da denir.

Genelleştirilmiş doğrusal modellerde tahmin en çok olabilirlik yöntemine dayanmaktadır. $y_1, \dots, y_n$ gözlenmiş değerleri için log-olabilirlik fonksiyonu Eşitlik (3.5)' de verildiği gibidir.

Log-olabilirlik fonksiyonuna zincir kuralı uygulanarak β 'ya göre türevi alınıp sıfıra eşitlenirse,

$$\frac{\partial l_i(\beta)}{\partial \beta} = \frac{\partial l_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta} = 0 \quad (3.21)$$

elde edilir. Eşitlik (3.21)'deki türevler

$$\begin{aligned} \frac{\partial l_i}{\partial \eta_i} &= \frac{\partial}{\partial \eta_i} \left(\sum_{i=1}^n \left(\frac{y_i \eta_i - b(\eta_i)}{a(\theta)} + k(y_i, \theta) \right) \right) \\ &= \sum_{i=1}^n \left(\left(y_i - \frac{\partial b(\eta_i)}{\partial \eta_i} \right) \frac{1}{a(\theta)} + \frac{\partial k(y_i, \theta)}{\partial \eta_i} \right) \\ &= \frac{1}{a(\theta)} \sum_{i=1}^n \left(y_i - \frac{\partial b(\eta_i)}{\partial \eta_i} \right) \end{aligned} \quad (3.22)$$

ve

$$\frac{\partial \eta_i}{\partial \beta} = \frac{\partial}{\partial \beta} (x_i' \beta) = x_i \quad (3.23)$$

şeklinde bulunur. Eşitlik (3.22) ve Eşitlik (3.23)'ü Eşitlik (3.21)'de yerine yazarsak,

$$\frac{\partial l}{\partial \beta} = \frac{\partial l}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta} = \frac{1}{a(\theta)} \sum_{i=1}^n \left(y_i - \frac{\partial b(\eta_i)}{\partial \eta_i} \right) x_i = 0 \quad (3.24)$$

şeklinde bulunur. Eşitlik (3.24)'ün β için çözümlenmesi gerekmektedir. Fakat $l(\beta)$, β 'nın doğrusal bir fonksiyonu olmadığı için doğrudan çözülmez ve Newton-Raphson metodu gibi iteratif yöntemler yardımı ile çözülebilmektedir. Maksimum $l(\beta)$ fonksiyonunu bulabilmek için $\frac{\partial l(\beta)}{\partial \beta}$ ifadesi Taylor serisi ile β^0 etrafında açılırsa,

$$\frac{\partial l(\beta)}{\partial \beta} = l'(\beta) = l'(\beta^0) + \frac{\partial^2 l(\beta)}{\partial \beta \partial \beta'} (\beta - \beta^0) \quad (3.25)$$

$$l'(\beta^0) + \frac{\partial^2 l(\beta)}{\partial \beta \partial \beta'} (\beta - \beta^0) = 0 \quad (3.26)$$

$$\beta = \beta^0 - \left[\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta'} \right]^{-1} l'(\beta^0) \quad (3.27)$$

olmak üzere kök için çözüm bulunur. Kökün tahminini geliştirmek için Eşitlik (3.27) yinelenmekte ve böylece gerçek köke yakınsamada,

$$\beta^{(k+1)} = \beta^{(k)} - \left[\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta'} \right]^{-1} l'(\beta^{(k)}) \quad (3.28)$$

eşitliği kullanılabilir. Genelleştirilmiş doğrusal modellerdeki durum için Newton-Raphson Metoduna göre çözüm $\frac{\partial l}{\partial \beta}$ ifadesi β etrafında,

$$\beta_{E\check{C}O} = \beta - \left[\frac{\partial^2 l}{\partial \beta \partial \beta'} \right]^{-1} \frac{\partial l}{\partial \beta} \quad (3.29)$$

biçiminde elde edilir. Eşitlik (3.21)'de verilen zincir kuralı genişletilirse,

$$\begin{aligned} \frac{\partial l(\beta)}{\partial \beta} &= \sum_{i=1}^n \frac{\partial l_i}{\partial \Phi_i} \frac{\partial \Phi_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta} \\ &= \sum_{i=1}^n x_i \frac{\partial \mu_i}{\partial \eta_i} \frac{(y_i - \mu_i)}{V(y_i)} \end{aligned} \quad (3.30)$$

$$\frac{\partial l(\beta)}{\partial \beta'} = \sum_{i=1}^n \frac{\partial \mu_i}{\partial \eta_i} \frac{(y_i - \mu_i)}{V(y_i)} x_i' \quad (3.31)$$

bulunur. β bileşenlerini bulmak için olabilirlik denklemi ise,

$$\sum_{i=1}^n x_i \frac{\partial \mu_i}{\partial \eta_i} \frac{(y_i - \mu_i)}{V(y_i)} = 0 \quad (3.32)$$

şeklindedir. Eşitlik (3.29)'da ki $\frac{\partial^2 l}{\partial \beta \partial \beta'}$ ifadesi, eşitlik (3.10)'da verilen Fisher Skorlama ile,

$$\begin{aligned} E \left[\frac{\partial^2 l}{\partial \beta \partial \beta'} \right] &= -E \left[\frac{\partial l}{\partial \beta} \frac{\partial l}{\partial \beta'} \right] \\ &= -E \left[\sum_{i=1}^n x_i \frac{\partial \mu_i}{\partial \eta_i} \frac{(y_i - \mu_i)}{V(y_i)} \frac{(y_i - \mu_i)}{V(y_i)} \frac{\partial \mu_i}{\partial \eta_i} x_i' \right] \end{aligned} \quad (3.33)$$

$$\begin{aligned}
&= - \sum_{i=1}^n x_i x_i' \left[\frac{\partial \mu_i}{\partial \eta_i} \right]^2 \frac{1}{V(y_i)} \\
&= -\theta
\end{aligned}$$

olarak hesaplanır. Burada, $W = \text{diag} \left[\frac{\left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2}{V(y_i)} \right]$ olmak üzere, $\theta = X'WX$ bilgi matrisi

olarak adlandırılır. Eşitlik (3.30) ve eşitlik (3.33), eşitlik (3.29)'da yerine yazılıp her iki tarafı $X'WX$ ile çarpılırsa,

$$\beta^{(k+1)}(X'WX) = \beta^{(k)}(X'WX) - \frac{\partial l}{\partial \beta} \Big|_{\beta=\beta^{(k)}} \quad (3.34)$$

şeklinde olur. Eşitlik (3.34)'ün sağ tarafı eşitlik (3.30) ve eşitlik (3.33)'ün birleşimi olarak yazılırsa

$$\left[\sum_{i=1}^n x_i x_i' \frac{\partial \mu_i}{\partial \eta_i} \frac{1}{V(y_i)} \beta^{(k)} \right] + \sum_{i=1}^n x_i \frac{\partial \mu_i}{\partial \eta_i} \frac{(y_i - \mu_i)}{V(y_i)} \quad (3.35)$$

elde edilir ve (3.35) eşitliği

$$X'Wz \quad (3.36)$$

olarak yazılabilir. (3.36)'daki z değeri,

$$\begin{aligned}
z &= \sum_{i=1}^n x_i \beta^{(k)} + (y_i - \mu_i) \left(\frac{\partial \eta_i}{\partial \mu_i} \right) \\
&= \eta_i + (y_i - \mu_i) \left(\frac{\partial \eta_i}{\partial \mu_i} \right)
\end{aligned} \quad (3.37)$$

şeklinde bulunur. Eşitlik (3.34)'e (3.37)'deki değer yazılırsa genelleştirilmiş doğrusal model için en çok olabilirlik iteratif tahmin edici,

$$\hat{\beta}^{(k+1)} = (X'W^{(k)}X)^{-1} X'W^{(k)} \left[\eta_i + (y_i - \mu_i) \frac{\partial \eta_i}{\partial \mu_i} \right]^{(k)} \quad (3.38)$$

şeklini alır. Buradan,

$$\hat{\beta}^{(k+1)} = (X'W^{(k)}X)^{-1} X'W^{(k)} z^{(k)} \quad (3.39)$$

olarak yazılır (Dobson ve Barnett, 2008; Montgomery ve ark,2012; Deniz, 2005).

3.2. Model Yeterliğinin Sınanması

Regresyon analizi sonucunda elde edilen modelin doğru olduğu biliniyorsa daha ileri bir analize gerek kalmadan çalışma sonlandırılabilir. Model yeterliğinin sınanması, parametre tahmin edicilerinin sınanması ve modelin uyum iyiliğinin test edilmesi sürecinden meydana gelmektedir. Bu testlerin sonucunda parametre tahmin edicilerinin modele uygun tahmin ediciler olup olmadığına ve kullanılan verilerin modele uyum gösterip göstermediğine karar verilebilir.

3.2.1 Parametre Tahminlerinin Sınanması

3.2.1.1 Bıçın Hipotez Testleri ve Güven Aralığı

Olabilirlik oranı ve Wald istatistiği modellerde kullanılan temel çıkarım araçlarındandır. Olabilirlik oranı testi, H_0 'ın H_1 'e karşı testi için kullanılan genel bir yöntemdir.

Regresyon modeli için hipotezler;

$$H_0 = \beta_1 = \dots = \beta_p = 0$$

$$H_1 = \text{En az biri } \beta_j \neq 0 \quad j = 1, 2, \dots, p$$

şeklinde gösterilir ve χ^2_{sd} istatistik değeri ile karşılaştırılır. Burada serbestlik derecesi parametre sayıları arasındaki farka eşittir. Herhangi bir gözlemin regresyon katsayısı β_j

için önem testleri;

$$H_0 = \beta_j = 0 \quad (\beta_j \text{ parametresi anlamsızdır.})$$

$$H_1 = \beta_j \neq 0 \quad (\beta_j \text{ parametresi anlamlıdır.})$$

şeklinde gösterilebilir.

Olabilirlik oranı testi t-testine karşılık gelen Wald testi ile eşdeğerdir ve $\hat{\beta} \sim N(\beta, (X'WX)^{-1})$ dir. Test istatistiği,

$$Z_i = \frac{\hat{\beta}_j}{S_{\hat{\beta}_j}} \quad (3.40)$$

biçimindedir. Güven aralığı ise, $\hat{\beta}_j \pm Z_{\alpha/2} S_{\hat{\beta}_j}$ ile bulunur. Wald ve olabilirlik oranı testlerinin hipotezleri asimptotik olarak eşdeğerdir (Dobson ve Barnett, 2008; Deniz, 2005).

3.2.1.2 Modelin Uyum İyiliğinin Sınanması ve Sapma İstatistiği

Parametreler tahmin edildikten sonra gözlemlerin modelin şekli etrafındaki dağılımlarının ölçülmesi gerekmektedir. Çünkü gözlemler tahmin edilen modelin çizilen şekline ne kadar yakınsa modelin uyum iyiliği de o kadar yüksek olacaktır. Başka bir şekilde ifade etmek gerekirse bağımlı değişkendeki değişimin açıklayıcı değişkendeki değişimlerle açıklanması o kadar iyi olacaktır.

Poisson regresyon modelinin uyum iyiliğinin sınanmasında; Pearson χ^2 , sapmalar istatistiği, yapay R^2 ölçümü, Akaike Bilgi Ölçütü (AIC) ve Bayes Bilgi Ölçütü (BIC) yaygın olarak kullanılan ölçütlerdir.

3.2.1.2.(1). Pearson İstatistiği

μ_i ortalamalı ve ω_i varyanslı yanıt değişkeni y_i 'ye ait herhangi bir model için standart uyum iyiliği ölçüm yöntemi Pearson istatistiğidir ve

$$\chi^2 = \sum_{i=1}^n \frac{(y_i - \hat{\mu})^2}{\hat{\omega}_i} \quad (3.41)$$

olarak ifade edilir. Bu değer serinin yayılımının aşırı olup olmadığını belirlemede kullanılır. Burada $\hat{\mu}_i$ ve $\hat{\omega}_i$ değerleri μ_i ve ω_i parametrelerinin tahmin değerleridir. Hesaplanan χ^2 değeri $\hat{\mu}$ için belirlenmiş serbestlik derecesi (n-k) ile karşılaştırılır. Hesaplanan χ^2 değerinin serbestlik derecesine oranı 1'den büyük değer alırsa verilerin modele uygun olmadığı ve aşırı yayılım durumunun olduğu söylenebilir (Dobson ve Barnett, 2008; Deniz, 2005).

3.2.1.2.(2).Sapma İstatistiği

Uyum iyiliğinin test edilmesi sürecinde kullanılan yöntemlerden bir diğeri de sapma istatistiğidir. Bu istatistik değerine " G^2 " istatistiği de denilmektedir ve

$$G^2 = 2 \sum_{i=1}^n y_i \ln \frac{y_i}{\mu_i} \quad (3.42)$$

şeklinde ifade edilir. Bu istatistik değeri 0'a yakınsıyorsa model uyumu artıyor denilebilir. Eğer bu istatistik değeri sıfıra eşitse model uyumunun mükemmel olduğu söylenebilir (Dobson ve Barnett, 2008; Deniz, 2005).

3.2.1.2.(3). Akaike Bilgi Ölçütü

Akaike Bilgi Ölçütü, verilen sonucu farklı modeller ile karşılaştırmak için bir yol olarak 1973 yılında Akaike tarafından geliştirilmiştir. Uygun değerlerin belirli beklenen değerler açısından gerçek değerler olma eğilimine ne kadar yakın bir model olduğunu belirler. Model karmaşıklığı ile uyumu arasında bir denge kurmaya yarar. Modeli açıklayabilecek kadar karmaşık ancak olası modeller arasından mümkün olduğu kadar en basit olan model aranır. Akaike bilgi ölçütü

$$AIC = -2\log L + 2q \quad (3.43)$$

şeklinde ifade edilmektedir. Bu eşitlikte L , log-olabilirlik fonksiyonunun maksimum değerini, q ; açıklayıcı değişken sayısını göstermektedir. AIC değeri ne kadar küçükse modelin o kadar iyi olduğu söylenebilir (Dobson ve Barnett, 2008; Deniz, 2005).

3.2.1.2.(4). Bayes Bilgi Ölçütü

Bayes bilgi ölçütü Akaike bilgi ölçütü gibi veriler ve model arasında uygunluğu ölçen yöntemlerden biridir. Bayes bilgi ölçütü

$$BIC = -2\ln(L) + q\ln(N) \quad (3.44)$$

őeklinde ifade edilecektir. Bu eőitlikte L , log-olabilirlik fonksiyonunun maksimum deėerini, q , aıklayıcı deėiőken sayısını, N rnek byklėn ifade etmektedir (Dobson ve Barnett, 2008; Deniz, 2005).





4. REGRESYON ANALİZİNDE BAZI YANLI TAHMİN EDİCİLER

Gizem YILDIRIM

4. REGRESYON ANALİZİNDE BAZI YANLI TAHMİN EDİCİLER

4.1. Ridge Regresyon Tahmin Edici

Çoklu doğrusal bağlantı probleminin en önemli sonuçlarından birisi regresyon katsayılarının varyans ve kovaryans değerlerinin artmasıdır. Yani çoklu doğrusal bağlantı olması halinde önemli değişkenlere ait regresyon katsayılarının standart hataları büyüyecek ve bu değişkenlere ait regresyon katsayılarının kısmi t testleri anlamsız çıkacaktır. Çoklu doğrusal bağlantı olması halinde herhangi bir bağımsız değişken veya birime ait veriler modelden çıkartıldığında veya modele eklendiğinde kısmi regresyon katsayılarında çok önemli değişiklikler olacaktır. Diğer bir deyişle çoklu doğrusal bağlantılı verilerle hesaplanan standartlaştırılmış regresyon katsayıları durağanlığını ve kararlılığını koruyamamaktadır.

Çoklu bağlantının yukarıda belirttiğimiz tüm bu etkilerini en aza indirebilmek için kullanılan yanlı tahmin yöntemlerinden biri de ridge regresyon tahmin yöntemidir. Ridge regresyon tahmin yöntemi Hoerl ve Kennard (1970a) tarafından bağımsız değişkenler arasında çoklu doğrusal bağlantı olması durumunda EKK tahmin yönteminin yetersiz kalması nedeniyle geliştirilen bir yanlı tahmin yöntemidir. Hoerl ve Kennard, ridge regresyon tahmin edici aşağıdaki durumlar için önermişlerdir:

- Kuvvetli çoklu doğrusal bağlantının varlığı halinde katsayılarda meydana gelen kararsızlıkların grafikte gösterilmesinde,
- Çoklu doğrusal regresyon modelinde bağımsız değişkenlerin birbirleri ile ilişkili oldukları durumlarda EKK tahmininden daha küçük varyanslı tahminlerin elde edilmesinde,

Ridge regresyon tahmin yöntemi yanlı regresyon tahmin yöntemi olmasına karşın EKK tahmin yöntemine göre iki önemli etkisi vardır. Bunlar:

1. Bağımsız değişkenlerdeki çoklu doğrusal bağlantıyı gidermek
2. Regresyonda hata kareler ortalamasını (HKO) azaltmak

4. REGRESYON ANALİZİNDE BAZI YANLI TAHMİN EDİCİLER

Gizem YILDIRIM

Bu bakımdan ridge regresyon tahmin edicileri yanlı tahmin ediciler olmasına karşın EKK tahminlerinden daha kararlı olabilmektedirler. Ridge regresyon tahmin edicileri varyansta sağladığı azalıştan dolayı çok çeşitli alanlarda kullanılmışlardır. Varyans ve kovaryans matrisinin köşegen değerlerine yanlılık sabiti “ k ” ilave etmenin haricinde ridge regresyon ve EKK tahmin yöntemlerinin işleyişleri aynıdır. Ridge regresyon ile bir taraftan tahminlerin varyansları azaltılırken diğer taraftan k katsayısı kadar yanlı tahminler elde edilmektedir. Yansız tahminlerde yüksek varyans, yanlı tahminlerde düşük varyans söz konusudur.

Değişkenler arasındaki ilişkiler, EKK katsayı tahmin edicilerinin varyanslarının artmasına neden olur. Modeldeki her bir terim için VIF (varyans şişirme faktörü) değerleri, katsayıların varyansları üzerinde basit korelasyonların toplam etkisini gösterir. Çoklu bağlantı durumunda, korelasyon matrisinin ters matrisindeki köşegen elemanlar olan VIF değerleri, her bir kestirimin diğer kestirimlerle olan çoklu korelasyonunda sonsuz hale gelir.

Bu durumda $y = X\beta + \varepsilon$ şeklinde tanımlanan klasik regresyon modeline dayalı olarak elde edilen $\hat{\beta}$ kestirimleri en küçük varyanslı kestirimler olmaz, çünkü çoklu doğrusal bağlantı $\hat{\beta}$ ve β arasında sapmaya neden olur. L_1 , $\hat{\beta}$ 'nin β 'den sapması olmak üzere yanlılığın karesi

$$L_1^2 = (\hat{\beta} - \beta)'(\hat{\beta} - \beta) \quad (4.1)$$

olarak gösterilebilir.

$$E(L_1^2) = \text{tr}(\text{Var}(\hat{\beta})) = \sigma^2 \text{tr}((X'X)^{-1}) \quad (4.2)$$

Burada $\text{tr}(\text{iz})$, bir kare matrisin esas köşegen üzerindeki elemanlarının toplamıdır. $(X'X)$ matrisinin öz değerleri $\lambda_{\max} = \lambda_1 > \lambda_2 > \dots > \lambda_k = \lambda_{\min} > 0$ ile gösterilir. Bu durumda

$$E(L_1^2) = \sigma^2 \sum_{j=1}^k \lambda_j^{-1} \quad (4.3)$$

olarak bulunur.

4. REGRESYON ANALİZİNDE BAZI YANLI TAHMİN EDİCİLER

Gizem YILDIRIM

Çoklu bağlantı olması durumunda, bir veya daha fazla öz değerin küçük değer alması $\hat{\beta}$ 'nın β 'dan sapmalarının beklenen değerinin büyümesine neden olacaktır. Çoklu doğrusal bağlantının hata kareler toplamı üzerindeki olumsuz etkisini giderebilmek için Hoerl ve Kennard (1970a, 1970b), ridge tahmin ediciyi aşağıdaki gibi tanımlamışlardır.

$$\hat{\beta}_{RR}(k) = (X'X + kI)^{-1}X'y \text{ veya } \hat{\beta}_{RR}(k) = (R + kI)^{-1}X'y, k > 0 \quad (4.4)$$

Yukarıdaki eşitlikte değişkenler standartlaştırıldığında R bağımsız değişkenler arasındaki korelasyon matrisini gösterdiğinden $X'X = R$ dir. Ridge tahmin edici, k yanlılık parametresine bağlıdır ve k 'nın seçimi ridge tahmin edicinin performansını etkiler. Burada k , pozitif değerler almaktadır. $k = 0$ için ridge çözümünün EKK çözümüne eşdeğer olduğu aşıkardır. Regresyon problemi için genelde k yanlılık parametresinin bir en uygun çözümü vardır. Bu optimal k yanlılık parametresi değeri, EKK yönteminden daha küçük hata kareler ortalaması elde edilmesini sağlar.

4.1.1. Ridge Regresyon Tahmin Edicinin Özellikleri

1- EKK tahmin edicinin doğrusal bir dönüşümü ridge tahmin edicidir:

$$\hat{\beta}_{RR}(k) = (X'X + kI)^{-1}X'y = (X'X + kI)^{-1}X'X\hat{\beta} \quad (4.5)$$

olarak yazılabileceğinden

$$\kappa = (X'X + kI)^{-1}X'X \quad (4.6)$$

olarak alınırsa

$$\hat{\beta}_{RR} = \kappa\hat{\beta} \quad (4.7)$$

olur.

2.Ridge tahmin edici yanlıdır:

T , $X'X$ matrisinin öz vektörlerinin ortogonal matrisi ve Λ , $X'X$ matrisinin öz değerlerinin diagonal matrisi olmak üzere

$$(X'X) = T\Lambda T' \quad (4.8)$$

4. REGRESYON ANALİZİNDE BAZI YANLI TAHMİN EDİCİLER

Gizem YILDIRIM

dönüşümü yapılırsa $i=1,2,\dots,n$ için

$$\kappa = (X'X + kI)^{-1}X'X = T' \text{diag}(\delta_i)T \quad (4.9)$$

olur. Burada

$$\delta_i = \frac{\lambda_i}{(\lambda_i + k)^{-1}} \quad (4.10)$$

dir. Buradan

$$\hat{\beta}_{RR}(k) = T' \text{diag}(\delta_i)T \hat{\beta} = \kappa \hat{\beta} \quad (4.11)$$

olur. $k > 0$ olduğundan

$$E(\hat{\beta}_{RR}) \neq \hat{\beta} \quad (4.12)$$

olur ve $\hat{\beta}_{RR}$ yanlı bir tahmin edicidir. Burada $k = 0$ olursa ridge tahmin edici, EKK tahmin ediciye eşit olacaktır.

3.Ridge Tahmin Edicinin Hata Kareler Ortalaması ve Özellikleri

(4.1) ve (4.2) nolu ifadeler ridge tahmin ediciye uygulanırsa

$$L_1^2(k) = (\hat{\beta}_{RR} - \beta)'(\hat{\beta}_{RR} - \beta) \quad (4.13)$$

$$\begin{aligned} E(L_1^2(k)) &= E[(\hat{\beta}_{RR} - \beta)'(\hat{\beta}_{RR} - \beta)] \\ &= E[(\hat{\beta} - \beta)'Z'Z(\hat{\beta} - \beta)] + (Z\beta - \beta)'(Z\beta - \beta) \\ &= \sigma^2 \text{tr}(X'X)^{-1}Z'Z + \beta'(Z - I)'(Z - I)\beta \\ &= \sigma^2 [\text{tr}(X'X + kI)^{-1} - k \text{tr}(X'X + kI)^{-2}] \\ &\quad + k^2 \beta'(X'X + kI)^{-2} \beta \\ &= \sigma^2 \sum_{i=1}^n \frac{\lambda_i}{(\lambda_i + k)^2} + k^2 \beta'(X'X + kI)^{-2} \beta \\ &= \gamma_{1(k)} + \gamma_{2(k)} \end{aligned} \quad (4.14)$$

olduğu Hoerl ve Kennard tarafından (1970a)'da gösterilmiştir. Burada $\gamma_{1(k)}\hat{\beta}_{RR}$ parametrelerinin varyans toplamını, $\gamma_{2(k)}$ ise yanlılığın karesini göstermektedir. Başka bir deyişle, $\gamma_{1(k)}$ parametrelerin toplam varyansıdır. $\gamma_{2(k)}$ ise $Z\beta$ 'dan β 'ya

4. REGRESYON ANALİZİNDE BAZI YANLI TAHMİN EDİCİLER

Gizem YILDIRIM

olan uzaklığın karesidir. $k = 0$ olduğunda $\gamma_{2(k)} = 0$ olacaktır, çünkü bu durumda $Z = I$ olacaktır.

Ridge tahmin edicinin hata kareler ortalaması

$$\text{MSE}(\hat{\beta}_{RR}) = E(\hat{\beta}_{RR} - \beta)^2 = V(\hat{\beta}_{RR}) + [E(\hat{\beta}_{RR}) - \beta]^2 \quad (4.15)$$

$$\text{MSE}(\hat{\beta}_{RR}) = \text{Var}(\hat{\beta}_{RR}) + (\text{bias}\hat{\beta}_{RR})^2 \quad (4.16)$$

olarak gösterilebilir.

4.Ridge Tahmin Edicinin Hata Kareler Ortalamasının Minimum Olması

Hoerl ve Kennard (1970a, 1970b), daha önceki başlıklarda gösterilen (4.2) ve (4.14) denklemleri arasındaki ilişkiyi Varlık Teoremi olarak adlandırıp $E(L_1^2(k)) < E(L_1^2)$ olacak şekilde bir $k > 0$ bulmak her zaman mümkündür.” şeklinde açıklamışlardır.

Bu eşitsizliğe göre çoklu bağlantı durumunda ridge regresyon yöntemiyle, EKK yöntemine göre daha küçük hata kareler ortalamasına sahip olan kestirimler yapmak mümkündür.

4.1.2. Ridge Regresyon Tahmin Edicinin Varyans ve Yanlılığı:

$\gamma_{1(k)}$ 'nin $\hat{\beta}_{RR}$ parametrelerinin varyans toplamı olduğunu Eşitlik (4.14) de açıklamıştık. Çoklu bağlantı olması durumunda regresyon modeli oluşturulurken ridge regresyon kullanılıyorsa, ridge tahmin edicinin varyansının nasıl bir değişime uğradığını bilmemiz gerekir. y rasgele değişkenimiz için

$$\hat{\beta}_{RR} = \kappa \hat{\beta} = \kappa(X'X)^{-1}X'y$$

yazılabilir.

$$\begin{aligned} \text{Var}(\hat{\beta}_{RR}) &= \text{Var}(\kappa(X'X)^{-1}X'y) \\ &= \kappa(X'X)^{-1}X'\text{Var}(y)X(X'X)^{-1}\kappa' \\ &= \sigma^2\kappa(X'X)^{-1}X'X(X'X)^{-1}\kappa' \\ &= \sigma^2\kappa(X'X)^{-1}\kappa' \end{aligned} \quad (4.17)$$

olur.

4. REGRESYON ANALİZİNDE BAZI YANLI TAHMİN EDİCİLER

Gizem YILDIRIM

$k > 0$ olduğunda ridge tahmin edicideki yanlılık k ile birlikte artmaktadır. k artarken varyans azalmaktadır. Bu durumla ilgili olarak Hoerl ve Kennard (1970a, 1970b) aşağıdaki teoremleri önermişlerdir.

Teorem 1: $\gamma_{1(k)}$ toplam varyans, k 'nın sürekli düzgün azalan bir fonksiyonudur.

Teorem 2: $\gamma_{2(k)}$ yanlılığın karesi, k 'nın sürekli artan düzgün bir fonksiyonudur.

Ridge regresyon yönteminde yanlılığın karesindeki artıştan daha büyük varyans azalmasını sağlayacak bir k değerinin belirlenmesi durumunda ridge tahmin edicideki hata kareler ortalaması EKK tahmin edicideki hata kareler ortalamasından daha küçük olacaktır. Bu yüzden k 'nın seçimi çok önemli bir yere sahiptir.

4.1.3. Ridge Regresyon Tahmin Edicide Parametre Seçimi

Çoklu doğrusal bağlantı durumunda ridge regresyon yöntemi ile yapılacak kestirimlerin kararlılığı, ridge yanlılık parametresi k 'nin en iyi değerin belirlenmesine bağlıdır. Uygun olan k değerinin belirlenmesi k 'nın bilinmeyen β parametre vektörüne ve varyansına bağlı olması sebebiyle kesin bir şekilde belirlenemez. Kesin olarak belirlenemeyen k ridge parametresini tahmin için birden çok öneri getirilmiştir. Bunlardan birkaçı aşağıdaki gibidir:

1. Hoerl ve Kennard k 'nın seçimi için ridge trace (ridge izi) kullanımını önermişlerdir.
2. Marquardt ve Snee (1975), VIF'in 1 ile 10 değerleri arasında olduğunda k 'nın seçilebileceğini belirtmişlerdir.
3. Andersson (2001) bağımsız değişkenler için 1'e yaklaşan VIF değerlerini sağlayan k sabitinin alınabileceği belirtmiştir.
4. Hoerl, Kennard ve Baldwin (1975)'de

$$k = \frac{p\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}} \quad (4.18)$$

eşitliğini önermişlerdir.

4. REGRESYON ANALİZİNDE BAZI YANLI TAHMİN EDİCİLER

Gizem YILDIRIM

5. McDonald ve Galarneau (1975)'de

$$\hat{\beta}'_{RR}\hat{\beta}_{RR} = \hat{\beta}'\hat{\beta} - \sigma^2 \sum_{j=1}^p \lambda_j^{-1} \quad (4.19)$$

eşitliğini sağlayan k değerinin uygun olacağını belirtmişlerdir. Burada λ_j , j.özdeğere karşılık gelmektedir.

Bunların dışında çoklu bağlantının çok güçlü olması durumunda EKK tahmin edicileri elde edilemeyeceğinden (4.19) da ki formül yerine

$$k = \frac{p\hat{\sigma}^2}{\sum_{i=1}^k \hat{\alpha}_i^2} \quad (4.20)$$

eşitliği kullanılır ve burada $\hat{\alpha}$ temel bileşenler tahmin edicidir.

4.1.4. Ridge Regresyon Tahmin Edicinin Performansının Değerlendirmesi

Ridge regresyon ve en çok olabilirlik tahmin yöntemlerinin performansını incelemek için ortalama hata kareler (MSE) değeri

$$MSE = \frac{1}{R} \sum_{i=1}^R SE_i = \frac{\sum_{i=1}^R (\hat{\beta} - \beta)'_i (\hat{\beta} - \beta)_i}{R} \quad (4.21)$$

şeklinde tanımlanıp $\hat{\beta}$, β 'nin en çok olabilirlik veya ridge regresyon tarafından elde edilen tahmin edicidir. R ise Monte Carlo simülasyonunda kullanılan tekrar sayısını ifade eder. MSE hesaplanırken sabit terim dışındaki parametreler kullanılır.

Ekonomik yayınlarda, çalışmanın odağı parametre tahminlerinden çok kısmi etkileridir. Her bir eğim katsayısının kısmi etkileri için yaygın uygulanan formül

$$\frac{1}{n} \sum_{i=1}^n \hat{\mu}_i \hat{\beta}_j \quad (4.22)$$

olup, $\hat{\mu}_i$ değerleri ML veya ridge tahmin edicileri ile elde edilen katsayılara karşılık gelir. Kısmi etkiler için MSE değeri ise eşitlik (4.22)'de ki her bir tahmin edilen eğim katsayısı yerine $\frac{1}{n} \sum_{i=1}^n \hat{\mu}_i \hat{\beta}_j$ yazarak elde edilebilir. Gerçek eğim katsayı değerleri için $\frac{1}{n} \sum_{i=1}^n \mu_i \beta_j$ kullanılır. Bu sayede açıklayıcı değişkenlerin kısmi

4. REGRESYON ANALİZİNDE BAZI YANLI TAHMİN EDİCİLER

Gizem YILDIRIM

etkileri tahmin edilen parametrelerin bir fonksiyonu olacak ve tahmin edilen parametreleri iyileştirmek ayrıca kısmi etkilerin tahminlerinin de iyileştirilmesi sağlayacaktır.

4.2 Liu Regresyon Tahmin Edici

Liu tahmin edici, çoklu iç ilişki problemi ile baş edebilen ridge regresyon tahmin ediciye alternatif olarak geliştirilmiş yanlı bir tahmin edicidir. Matris formunda gösterilen çoklu doğrusal regresyon modeli

$$y = X\beta + \varepsilon \quad (4.23)$$

şeklinde. Bu model için Liu tahmin edici, $0 < d < 1$ yanlılık çarpanı olmak üzere,

$$\hat{\beta}_{LR}(d) = (X'X + I)^{-1}(X'X + dI)\hat{\beta} \quad (4.24)$$

şeklinde ifade edilmektedir. Tahmin edicinin varyans-kovaryans matrisi,

$$Var(\hat{\beta}_{LR}) = \sigma^2(X'X + I)^{-1}(X'X + dI)(X'X)^{-1}(X'X + dI)(X'X + I)^{-1} \quad (4.25)$$

biçiminde ifade edilir. Liu için MSE, λ_i , $X'X$ matrisinin öz değerleri olmak üzere,

$$MSE(\hat{\beta}_{LR}) = \sum_{i=1}^P \frac{(\lambda_i + d)^2}{\lambda_i(\lambda_i + 1)^2} + (d - 1)^2 \sum_{i=1}^P \frac{(\beta_i)^2}{(\lambda_i + 1)^2} \quad (4.26)$$

dir. Liu tahmin edicinin ridge tahmin ediciye göre avantajı d yanlılık parametresinin bir lineer fonksiyonu olmasıdır ve böylece d 'nin seçimi k 'dan daha kolay olmaktadır (Liu, 1993).

5. POISSON REGRESYON ANALİZİ

Poisson regresyon modeli, bağımlı değişkenin sayılabilir olduğu durumlarda kullanılan bir regresyon modelidir. İlgilenilen bir olayın meydana gelme oranı veya bir olayın meydana gelme sayısı poisson regresyon ile modellenilebilir. Poisson regresyon, incelenmek istenen zaman aralığında meydana gelmiş olan olaylar ile belirlenen bağımsız değişkenler arasında bağlantı kurmaya çalışılan problemlerde uygulanmaktadır. Model, süresiz ve negatif olmayan sayılabilir verileri içermesinden dolayı, beklenen sayıların logaritmasının bağımsız değişkenlerin bir fonksiyonu olduğu varsayılmaktadır.

5.1. Poisson Regresyon Modeli

Poisson regresyon modeli, bağımsız değişken $X' = [1, X_{i1}, X_{i2}, \dots, X_{ip}]_{n \times (p+1)}$ ve μ parametresi arasındaki ilişkiyi ifade eden Poisson dağılımından üretilir. Poisson regresyon modeli, poisson dağılımının ortalamasına göre belirlenir ve model

$$f(y_i|x_i) = \begin{cases} \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} & y_i = 0, 1, 2, \dots \end{cases} \quad (5.1)$$

şeklinde verilmektedir (Cameron ve Trivedi, 2013). Burada μ_i ,

$$E(y_i|x_i) = \mu(x_i) = c_i f(x_i, \beta) = \mu_i \quad i = 1, 2, \dots, n \quad (5.2)$$

şeklindedir.

Poisson regresyon modelinde, parametrelerin tahmini genel olarak en çok olabilirlik yöntemi ile gerçekleştirilmektedir. En çok olabilirlik yöntemi bağımlı değişken y_i 'nin dağılımına göre $\hat{\beta}$ tahmin edicileri hesaplamada en çok kullanılan yöntemdir. Bu yöntemde β tahminleri, log olabilirlik fonksiyonunu en büyük yapacak şekilde seçilmelidir.

5.2. Bol Sıfırlı Poisson Regresyon Modeli (ZIPR)

Poisson dağılımında ortalama ile varyansın eşit olduğundan önceki bölümlerde bahsedilmiştir. Bununla birlikte bu eşitliğin sağlanamaması durumunda, genellikle aşırı yayılım nadir de olsa az yayılımın görüldüğüne değinilmiştir. Bu tür verilerin modellenmesi için bol sıfırlı modeller kullanılmaktadır. Bol sıfırlı modeller, veride mevcut olan fazla sıfırlara ilişkin yoğunluğun temel sayı dağılımından ayrı tutularak hesaplanmasından dolayı iki durumlu süreçler (dual state process) de denilmektedir.

Bol sıfırlı poisson regresyon modelinde veride sadece aşırı yayılımın sıfır değerlerindeki yoğunluktan kaynaklandığı varsayımı söz konusu iken, bol sıfırlı negatif binom regresyon modeli ve bol sıfırlı genelleştirilmiş poisson regresyon modellerinde ise aşırı yayılım haricinde birçok başka nedenden de kaynaklanabileceği varsayımı mevcuttur. Bu yüzden bol sıfırlı regresyon modellerinin hepsinin ortalaması birbirine eşitken varyansları farklılık göstermektedir. Bol sıfırlı poisson regresyon modelleri tamamen bol sıfırlı sayma verilerinin incelenmesi için geliştirilmiştir.

Bol sıfırlı poisson regresyon modeli

$$P(Y_i = 0) = \pi_i + (1 - \pi_i)e^{-\mu_i}y_i = 0 \quad (5.3)$$

$$P(Y_i = y_i) = (1 - \pi_i) \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!} y_i > 0 \quad (5.4)$$

şeklinde (Cameron ve Trivedi, 2013; Agresti, 2015).

Bol sıfırlı poisson regresyon modelinin beklenen değeri ve varyansı aşağıda gösterildiği gibi olacaktır:

$$E(Y_i) = (1 - \pi_i)\mu_i \quad (5.5)$$

$$V(Y_i) = \mu_i(1 - \pi_i)(1 + \mu_i\pi_i) \quad (5.6)$$

Bol sıfırlı poisson regresyonda $\pi_i = 0$ olduğunda poisson regresyona ve $\pi_i > 0$ olduğunda ise regresyon aşırı yayılım durumuna gider.

5.3. Bol Sıfırlı Genelleştirilmiş Poisson Regresyon Modeli (ZIGPR)

Bu model bol sıfırlı negatif binom regresyona bir alternatif olması amacıyla sayım verilerinde sıfırların çokluğunu incelemek ve aşırı yayılım sorununu önlemek için Consul ve Famoye (1992) tarafından önerilmiştir. Önerilen model genelleştirilmiş poisson dağılımına dayanır.

y_i , $i = 1, 2, \dots, n$ sayım verilerinde ölçülen bağımlı değişkenlerin ölçümünü gösterebilir. Y_i 'nin olasılık yoğunluk fonksiyonu aşağıdaki gibi gösterilir:

$$f(Y_i = y_i) = \left(\frac{\mu_i}{1 + \theta \mu_i} \right)^{y_i} \left(\frac{1 + \theta \mu_i}{y_i!} \right)^{y_i-1} \exp \left\{ \frac{-\mu_i(1 + \theta \mu_i)}{1 + \theta \mu_i} \right\}, y_i \geq 0 \quad (5.7)$$

Genelleştirilmiş poisson dağılımının ortalama ve varyansı sırasıyla;

$$E(Y_i) = \mu_i \quad (5.8)$$

$$V(Y_i) = \mu_i(1 + \theta \mu_i)^2 \quad (5.9)$$

şeklinde gösterilir. Bol sıfırlı poisson regresyon modelinden farkı bu modelde yer alan θ parametresinin aldığı değere göre aşırı veya zayıf yayılım durumunun oluşmasıdır. $\theta > 0$ iken bu model aşırı yayılım (over-dispersion) durumuna, $\theta < 0$ iken zayıf yayılım (under-dispersion) durumuna karşılık gelir (Famoye ve Singh, 2006). Beklenenden daha fazla sayıda sıfır gözlemi olduğunda genelleştirilmiş poisson regresyon modeli iyi uyum sağlamaz (Cameron ve Trivedi, 2013; Agresti, 2015).

5.4. En Çok Olabilirlik Tahmin Edici (MLE)

En çok olabilirlik yönteminde $\hat{\beta}$ kestirimleri, log-olabilirlik fonksiyonunu en büyük yapacak şekilde seçilmektedir. Poisson regresyon modelinde ortalama parametresi

$$E(y_i|x_i) = \mu_i = \exp(x_i'\beta) \quad (5.10)$$

şeklindedir.

Bu ifade olasılık fonksiyonunda yerine yazılırsa

$$f(y_i|x_i) = \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} \quad y_i = 0, 1, 2, \dots \quad (5.11)$$

Bağımsız gözlemler için, en çok olabilirlik fonksiyonu

$$L(y, \beta) = \prod_{i=1}^n f(y_i) = \prod_{i=1}^n \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} = \frac{(\prod_{i=1}^n \mu_i^{y_i}) e^{-\sum_{i=1}^n \mu_i}}{\prod_{i=1}^n (y_i!)} \quad (5.12)$$

olarak yazılabilir.

Log-olabilirlik fonksiyonu

$$l(\beta) = \sum_{i=1}^n y_i \ln(\mu_i) - \sum_{i=1}^n (\mu_i) - \sum_{i=1}^n \ln(y_i!) \quad (5.13)$$

şeklindedir. $\mu_i = \exp(x_i' \beta)$ eşitlikte yerine yazılırsa

$$l(\beta) = \sum_{i=1}^n y_i x_i' \beta - \sum_{i=1}^n e^{x_i' \beta} - \sum_{i=1}^n \ln(y_i!) \quad (5.14)$$

elde edilir.

Poisson regresyon modelinin parametrelerinin en çok olabilirlik tahmini bulunmak istendiğinden; yukarıdaki log-olabilirlik fonksiyonunun β 'ya göre birinci dereceden kısmi türevi alınıp sıfıra eşitlenerek

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^n y_i x_i' - \sum_{i=1}^n x_i e^{-x_i' \beta} = \sum_{i=1}^n x_i (y_i - \mu_i) = 0 \quad (5.15)$$

$$\sum_{i=1}^n x_i' (y_i - \exp(x_i' \beta)) = 0 \quad (5.16)$$

$$\sum_{i=1}^n x_i' (y_i - \mu_i) = 0 \quad (5.17)$$

şeklinde yazılır (Cameron ve Trivedi, 2013).

Bu eşitlikten eğer açıklayıcı değişkenler bir sabit terim içerirse ($y_i - \mu_i$) artıklar toplamının sıfır olacağı görülebilir. Ortalama parametresinin üstel olması

neticesinde log-olabilirlik fonksiyonları doğrusal olmadığı için, en çok olabilirlik yöntemi ile tek adımda çözüme ulaşılmaz.

Burada $\hat{\beta}_p$ değerini hesaplamak için analitik çözüm bulunmamaktadır. Çözüm için adimsal ağırlıklı en küçük kareler (IWLS) yöntemi kullanılabilir. İteratif ağırlıklandırılmış en küçük kareler yöntemi algoritması

$$\hat{\beta}_{ML} = (x' \hat{W} \hat{z})^{-1} x' \hat{W} \hat{z} \quad (5.18)$$

şeklindedir. Burada $\hat{W} = \text{diag}[\hat{\mu}_i]$ ve $\hat{z}$,

$$\hat{z}_i = \log(\hat{\mu}_i) + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} \quad (5.19)$$

olarak tanımlanır. En çok olabilirlik tahmin edicide, ikinci türevler matrisinin tersine karşılık gelen asimptotik olarak normal dağılımlı kovaryans matrisi

$$\text{Var}(\hat{\beta}_{ML}) = \left[-E \left(\frac{\partial^2 l}{\partial \beta_j \partial \beta'_k} \right) \right]^{-1} = (x' \hat{W} x)^{-1} \quad (5.20)$$

şeklindedir. Ayrıca MSE eşitliği

$$E(L_{ML}^2) = E(\hat{\beta}_{ML} - \beta)' (\hat{\beta}_{ML} - \beta) = \text{tr} \left[(x' \hat{W} x)^{-1} \right] = \sum_{j=1}^J \frac{1}{\lambda_j} \quad (5.21)$$

olarak tanımlanır. Burada λ_j , $x' \hat{W} x$ matrisinin j.inci öz değeridir. Açıklayıcı değişkenler yüksek korelasyonlu olduğunda ağırlıklandırılmış çapraz çarpımlar matrisi $x' \hat{W} x$ kötü koşullu olur ki bu en çok olabilirlik tahmininin yüksek varyanslı ve değişken olmasına neden olur. Bu durumda, tahmin edilmiş katsayıların ortalama vektörü çok uzun olduğundan, tahmin edilmiş parametreleri yorumlamak oldukça zordur. Verilen bilgiler ile uygulanan modeller doğrultusunda varyans değeri için

$$\hat{\beta}_{PR} \sim N \left[\beta, V_{ML}[\hat{\beta}_{PR}] \right]$$

$$V_{ML}[\hat{\beta}_{PR}] = \left(\sum_{i=1}^n \mu_i x_i x_i' \right)^{-1} \quad (5.22)$$

sonuçlarına ulaşılır (Cameron ve Trivedi, 2013; Agresti, 2015).

5.5. Poisson Regresyon Katsayılarının Yorumlanması

Poisson regresyon sürecinde elde edilen katsayılar ile bağımlı değişken y_i değerleri için tahminler yapılabilmektedir. Doğrusal modeller için $E(Y|x_i) = x_i' \beta$ ile β katsayıları i. bağımsız değişkenler içinde bir birimlik değişkenin etkisi olarak yorumlanır. Poisson regresyon modeli üstel bir yapıda olduğundan bu yorumun değişmesi gerekir. Üstel koşullu ortalama i. bağımsız değişken değeri için,

$$\frac{\partial E(y|x_i)}{\partial x_i} = \beta_i \exp(x_i' \beta) \quad (5.23)$$

şeklinde gösterilir. Burada β_i için x_i 'deki bir birimlik değişime bağlı olan $E(y|x_i)$ 'deki nispi değişimi ölçtüğü yorumu yapılabilir.

5.6. Poisson Regresyon Analizinde Bazı Yanlı Tahmin Ediciler

5.6.1. Poisson Ridge Tahmin Edici

Çoklu iç ilişki kaynaklı probleme bir çözüm olarak Mansson ve Shukur (2011) tarafından frekans verilerine uygulanan poisson ridge regresyon yöntemi önerilmiştir. Bu metodun çözümü elde edilirken en çok olabilirlik yaklaşımının ağırlıklı hata kareler toplamını (AHKT) yaklaşık olarak minimize edeceği bilgisi kullanılır. Bu nedenle, $\hat{\beta}_{MLE}$ ağırlıklı hata kareler toplamı açısından optimal bir tahmin edici olarak görülebilir. Bu yaklaşımla elde edilen tahmin edici

$$\hat{\beta}_{PRR} = (X' \widehat{W} X + kI)^{-1} X' \widehat{W} X \hat{\beta}_{ML} = Z \hat{\beta}_{ML} \quad (5.24)$$

$$\begin{aligned} E(L_{PRR}^2) &= E(\hat{\beta}_{PRR} - \beta)' (\hat{\beta}_{PRR} - \beta) \\ &= E[(\hat{\beta}_{ML} - \beta)' Z' Z (\hat{\beta}_{ML} - \beta)] + (Z\beta - \beta)' (Z\beta - \beta) \end{aligned} \quad (5.25)$$

$$= \sum_{j=1}^J \frac{\lambda_j}{(\lambda_j + k)^2} + \beta' k^2 (X' \widehat{W} X + kI)^{-2} \beta = \gamma_{1(k)} + \gamma_{2(k)} \quad (5.26)$$

olarak tanımlanır. Burada $\gamma_{1(k)}$ varyansa ve $\gamma_{2(k)}$ yanlılık teriminin karesine karşılık gelir. PRR yöntemi iki açıdan dikkat çekicidir. İlk olarak yazılımsal açıdan herhangi bir değişiklik gerektirmediği için çok basit bir yöntemdir. İkincisi varyans terimindeki azalmayı yanlılık teriminin karesindeki artıştan daha fazla yapacak şekilde uygun bir k değerinin belirlenmesi durumunda en çok olabilirlik tahmininden daha küçük MSE değeri sağlar.

Hoerl ve Kennard (1970a, 1970b) ridge regresyon için MSE'nin daima EKK'den daha küçük olduğu, sıfırdan büyük olan bir k değerinin olduğunu göstermişlerdir. Burada, aynı zamanda poisson ridge tahmin edicinin MSE'sinin, en çok olabilirlik tahmin edicinin MSE'sinden daha küçük olduğu bir k değerinin olduğu gösterilecektir. Bunun için ilk olarak $\gamma_{1(k)}$ 'in k 'nın monoton azalan bir fonksiyonu olduğu ve $\gamma_{2(k)}$ 'in k 'nın monoton artan bir fonksiyonu olduğu gösterilmelidir.

$$\gamma_{2(k)} = k^2 \beta' k^2 (x' \widehat{W} x + kI)^{-2} \beta = k^2 \sum_{i=1}^p \frac{\alpha_i^2}{(\lambda_i + k)^2} \quad (5.27)$$

Burada $\alpha_i^2 = \gamma \hat{\beta}_{ML}$ dir ve $\gamma, x' \widehat{W} x$ matrisinin özvektörüdür.

Tüm i değerleri için $k \geq 0$ ve $\lambda_i > 0$ olduğundan $\gamma_{2(k)}$ 'in k 'nın monoton artan bir fonksiyonu olduğu sonucuna varabiliriz. $E(L_{PRR}^2) < E(L_{ML}^2)$ olduğunu göstermek için $E(L_{PRR}^2)$ 'nin k 'ya göre birinci türevini alırsak;

$$\frac{\delta E(L_{PRR}^2)}{\delta k} = \frac{\delta \gamma_{1(k)}}{\delta k} + \frac{\delta \gamma_{2(k)}}{\delta k} = -2 \sum_{i=1}^p \frac{\lambda_i}{(\lambda_i + k)^3} + 2k \sum_{i=1}^p \frac{\lambda_i \alpha_i^2}{(\lambda_i + k)^3} \quad (5.28)$$

Bu sayede $\gamma_{1(k)}$ 'in k 'nın monoton azalan bir fonksiyonu olduğunu ve $\gamma_{2(k)}$ 'in k 'nın monoton artan bir fonksiyonu olduğu gösterilmiştir. Aynı zamanda sırasıyla $\gamma_{1(k)}$ ve $\gamma_{2(k)}$ 'nin birinci türevlerinin daima pozitif olmayan ve negatif

olmayan olduğu gösterilmiştir. Bu nedenle $E(L_{PRR}^2) < E(L_{ML}^2)$ için $\frac{\delta E(L_{PRR}^2)}{\delta k} < 0$ eşitliğindeki gibi daima sıfırdan büyük bir k değerinin olduğunu göstermek gereklidir.

$$k < \frac{1}{\alpha_{maksimum}^2} \quad (5.29)$$

Burada $\alpha_{maksimum}^2$, α_i^2 'nin en büyük elemanı olarak tanımlanmaktadır. Eşitlik (5.29) sıfıra eşitlenerek k için çözüldüğünde k 'nın en iyi değeri bulunmuş olacaktır.

$$k = \frac{1}{\alpha_i^2} \quad (5.30)$$

5.6.1.1. Önerilen Ridge Parametresi Tahmin Yaklaşımları

Bu bölümde ridge parametresi k 'yı tahmin edebilmek için farklı tahmin edicilere değinilmiştir. Bunlardan en çok bilinen ridge parametresi Hoerl ve Kennard (1970a, 1970b) tarafından önerilen

$$K1 = \hat{k}_{HK1} = \frac{\hat{\sigma}^2}{\hat{\alpha}_{maksimum}^2} \quad (5.31)$$

tahmin edicidir. Burada

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{\mu}_i)^2}{n - p - 1} \quad (5.32)$$

Bu tahmin edicinin değiştirilmiş bir versiyonu

$$K2 = \hat{k}_{HKM} = \frac{1}{\hat{\alpha}_{maksimum}^2} \quad (5.33)$$

olarak önerilmiştir. PRR eşitliğindeki $\frac{1}{\hat{\alpha}_{maksimum}^2}$ için eşitlik (5.30)'da ifade edilen en iyi k değerine karşılık gelmektedir.

Her ne kadar yukarıda gösterdiğimiz bu tahmin ediciler birçok çalışmada (Schaefer, 1986; Kibria, 2003; Alkhamisi ve Shukur, 2008) k 'nın en iyi değerini

tahmin etmek için gösterilmiş olsa da, biz Kibria(2003)'den iki tahmin ediciye daha yer vereceğiz:

$$K3 = \hat{k}_{GM} = \frac{\sigma^2}{(\prod_{i=1}^p \hat{\alpha}_i^2)^{\frac{1}{p}}} \quad (5.34)$$

ve

$$K4 = \hat{k}_{MED} = \text{Medyan}\{m_i^2\} \quad (5.35)$$

Bu eşitlikte

$$m_i = \sqrt{\frac{\hat{\sigma}^2}{\hat{\alpha}_i^2}} \quad (5.36)$$

olacaktır.

Diğer bir k tahmin edici ilk defa Alkhamisi ve ark. (2006) tarafından önerilen

$$K5 = \hat{k}_{maksimum}^{KS} = \text{maksimum}(s_i) \quad (5.37)$$

tahmin edicidir.

Burada

$$s_i = \frac{t_i \hat{\sigma}^2}{(n-p)\hat{\sigma}^2 + t_i \hat{\alpha}_i^2} \quad (5.38)$$

dir ve t_i , $X'X$ matrisinin özdeğeridir.

Hoerl ve Kennard (1970a, 1970b) ve Kibria (2003) tarafından önerilen ridge parametre tahmin edicileri aynı özelliği taşımaktadırlar, bu da tahmin edilen katsayı vektörü ortalama olarak daha uzun normlu olduğunda, korelasyon (ρ) değerleri arttıkça tahmin edilen değerlerin azalacağıdır. Bu özellik bazı durumlarda yararlı olmayabilir çünkü artan korelasyon (ρ) değerlerine karşılık $x'\hat{W}x$ matrisinin yakın tekillik problemini çözmek için k'nın daha geniş değer aralığına ihtiyaç duyulur. Daha önceki tahmin edicilerden ve Alkhamisi ve Shukur (2008)'den karekök dönüşümleri düşüncesinden yola çıkılarak Muniz ve Kibria (2009) aşağıdaki tahmin edicileri önermişlerdir:

$$K6 = \hat{k}_{KM2} = \text{maksimum}\left(\frac{1}{m_i}\right) \quad (5.39)$$

$$K7 = \hat{k}_{KM4} = \left(\prod_{i=1}^p \frac{1}{m_i} \right)^{1/p} \quad (5.40)$$

$$K8 = \hat{k}_{KM6} = \text{medyan} \left(\frac{1}{m_i} \right) \quad (5.41)$$

Bu tahmin ediciler m_i 'nin tersine bağlıdır bu yüzden korelasyon (ρ) değeri büyüdükçe tahmin ediciler de artar. Sonuç olarak bu tahmin edicilerin çoklu iç ilişki sorununa karşılık en dayanıklı tahmin ediciler olduğu varsayılmaktadır.

5.6.2. Bol Sıfırlı Modellerde Poisson Ridge Regresyon (ZIPRR)

Bol sıfırlı poisson regresyon modeli, bağımlı değişkenlerin aşırı yayılımı ve sıfırların fazlalığını açıklamasından dolayı, negatif olmayan ekonomik verilerin analizinde yaygın olarak kullanılan bir modeldir. Bu model, firmaların patent sayısı, ipotek ve kredi kartının finansal yükümlülüklerini ne sıklıkla yerine getirmediği, tekliflerin şirketlere devredilme miktarlarını ve farklı firmaların ihracat sayısı gibi farklı ekonomik faktörler üzerindeki etkiyi görmek için inovasyon analizlerinde kullanılabilir (Kibria ve ark., 2013).

Çoklu iç ilişki sorunu yüksek varyansa ve katsayı tahminlerini yorumlamada zorluklara yol açar. Bundan dolayı çoklu iç ilişki, geçerli istatistiksel çıkarımda bulunmayı ve farklı ekonomik faktörlerin bağımlı değişken üzerindeki etkisini araştırmayı zorlaştırır. Doğrusal regresyon modelleri için bu problemin bir çözümü olarak Hoerl ve Kennard (1970a, 1970b) sıradan EKK tahmin edicinin yerine yanlış bir tahmin edici olan ridge regresyonun kullanılabileceğini öne sürmüşlerdir. Hoerl ve Kennard (1970a, 1970b) küçük bir miktar yanlışlık ile birlikte tahmin edilen katsayıların varyansının önemli ölçüde azaltılabileceğini göstermişlerdir. Bu model daha sonra, Schaffer ve ark. (1984) tarafından logit ve poisson modelleri gibi ML ile tahmin edilen modellerde kullanılmak üzere genelleştirilmiştir (Kibria ve ark., 2013).

Kibria ve ark. (2013) çalışmalarında bol sıfırlı poisson regresyon (ZIPR) modeli için bir ridge regresyon (RR) tahmin edici önermek ve ridge parametresini tahmin etmek için ilk olarak Hoerl ve Kennard (1970a, 1970b), Kibria (2003), Muniz ve Kibria (2009) ve Kibria ve ark. (2011) tarafından önerilen bazı tahmin edicileri genelleştirmişlerdir. Böylelikle bu genelleştirilen tahmin ediciler ZIPRR için kullanılabilir olmuştur. Önerilen farklı tahmin edicilerin performanslarını karşılaştırmak için Monte Carlo Simülasyonunu kullanarak MSE ve MAE değerlerini hesaplanmıştır. Simulasyon çalışmasında örneklem büyüklüğü, açıklayıcı değişkenler arasındaki korelasyon derecesi ve logit modellerin kesişim değerleri gibi faktörleri değiştirilmiştir. Simulasyon çalışmasından elde edilen sonuçlar ZIPR modeli için yeni bir RR tahmin edicinin ML tahmin ediciye üstün geldiğini açıkça göstermektedir.

β 'yı tahmin etmek için en yaygın metot

$$l = \sum_{i=1}^N y_i \log(\pi_i) + \sum_{i=1}^N (1 - y_i) \log(1 - \pi_i) \quad (5.42)$$

şeklinde tanımlanan ortak log-olabilirlik fonksiyonunu maksimize eden ML yöntemidir.

Bu karmaşık olabilirlik fonksiyonu başlangıç değerlerinin ayrı ayrı poisson ve logit kestirimine bağlı iteratif ağırlıklı en küçük kareler (IWLS) algoritması ile elde edilir ve Nelder ve Mead (1965) tarafından önerilen simpleks yöntemi kullanılarak tahmin edilir. Bu yöneme alternatif olarak, Kibria ve ark. (2013) poisson modeli için aşağıdaki tahmin ediciyi önermişlerdir:

$$\hat{\beta}_{RR} = (X' \hat{W} X + kI)^{-1} X' \hat{W} X \hat{\beta}_{ML} \quad (5.43)$$

Burada $\hat{\beta}_{ML}$ tahminleri simpleks yöntem kullanılarak elde edilen tahminlerdir. $\hat{W}$, köşegen dışı elemanları sıfıra eşit olan ve i. köşegen elemanı $\hat{\mu}_i$ ya eşit olan bir matristir. Bu tahmin edici sırasıyla Hoerl ve Kennard (1970a,

1970b), Schaeffer ve ark. (1984) ve Mansson ve Shukur (2010a, 2010b) tarafından önerilmiş olan doğrusal, Logit ve poisson modellerle aynı mantığa sahip yanlış tahmin edicidir. Bu yanlışlık parametresi k sıfır ile sonsuz arasında değer alabilir. Sıfıra eşit olması durumunda $\hat{\beta}_{RR} = \hat{\beta}_{ML}$ olacaktır. Sıfırdan büyük olduğunda $\|\hat{\beta}_{RR}\| \leq \|\hat{\beta}_{ML}\|$ olacaktır. Çoklu iç ilişki olması durumunda en çok olabilirlik tahmin edici, $\hat{\beta}_{ML}$, ortalama olarak çok uzun (normu büyük) olacağından, $\hat{\beta}_{RR}$ 'nin $\hat{\beta}_{ML}$ 'den daha iyi performans göstermesi beklenir (Kibria ve ark., 2013).

5.6.2.1. Önerilen Ridge Parametresi Tahmin Yaklaşımları

Ridge parametresinin nasıl tahmin edileceği konusunda kesinleşmiş bir kural bulunmamaktadır. Bununla birlikte lineer RR modeli için birçok yöntem belirlenmiş olup, bu yöntemler ZIPR modeli için genelleştirilmiştir (Kibria ve ark., 2013). İlk tahmin edici,

$$K1 = \hat{k}_{HK1} = \frac{\hat{\sigma}^2}{\hat{\alpha}_{\text{maksimum}}^2} \quad (5.44)$$

Hoerl ve Kennard (1970a, 1970b) tarafından önerilmiştir. Burada $\hat{\alpha}_{\text{maksimum}}^2 \gamma \hat{\beta}_{ML}$ 'nin en büyük elemanıdır. γ , $X'WX$ matrisinin özvektörü ve $\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{\mu}_i)^2}{n-p-1}$ dir.

İkinci bir k tahmin edici Schaeffer ve ark. tarafından (1984) de önerilen tahmin edicidir. Lojistik RR için

$$K2 = \hat{k}_{HKM} = \frac{1}{\hat{\sigma}_{\text{maksimum}}^2} \quad (\text{Schaefer ve ark.}, 1984) \quad (5.45)$$

Sıradaki iki tahmin edici lineer regresyon için Kibria (2003) tarafından önerilmiş tahmin edicilerdir.

$$K3 = \hat{k}_{GM} = \frac{\hat{\sigma}^2}{(\prod_{i=1}^p \hat{\alpha}_i^2)^{1/p}} \quad (5.46)$$

$$K4 = \hat{k}_{MED} = \text{medyan} \{m_j^2\} \quad (5.47)$$

Burada;

$$m_i = \sqrt{\hat{\sigma}^2 / \hat{\alpha}_i^2} \quad (5.48)$$

olarak tanımlanır. k parametresinin seçimi için Eşitlik (5.49) – (5.60) ile verilen tahmin ediciler de kullanılabilir.

$$K5 = \hat{k}_{m5} = \text{Maksimum} \left(\frac{1}{m_j} \right), \quad (5.49)$$

$$K6 = \hat{k}_{m6} = \text{Maksimum}(m_j), \quad (5.50)$$

$$K7 = \hat{k}_{m7} = \prod_{j=1}^p \left(\frac{1}{m_j} \right)^{\frac{1}{p}}, \quad (5.51)$$

$$K8 = \hat{k}_{m8} = \prod_{j=1}^p (m_j)^{\frac{1}{p}}, \quad (5.52)$$

$$K9 = \hat{k}_{m9} = \text{medyan} \left(\frac{1}{m_j} \right), \quad (5.53)$$

$$K10 = \hat{k}_{m10} = \text{medyan}(m_j), \quad (5.54)$$

$$K11 = \hat{k}_{m11} = \text{Maksimum} \left(\frac{1}{q_j} \right), \quad (5.55)$$

$$K12 = \hat{k}_{m12} = \text{Maksimum}(q_j), \quad (5.56)$$

$$K13 = \hat{k}_{m13} = \prod_{j=1}^p \left(\frac{1}{q_j} \right)^{\frac{1}{p}}, \quad (5.57)$$

$$K14 = \hat{k}_{m14} = \prod_{j=1}^p (q_j)^{\frac{1}{p}}, \quad (5.58)$$

$$K15 = \hat{k}_{m15} = \text{medyan} \left(\frac{1}{q_j} \right), \quad (5.59)$$

$$K16 = \hat{k}_{m16} = \text{medyan}(q_j). \quad (5.60)$$

Burada $q_j = \frac{\lambda_{maksimum}}{(n-p)\hat{\sigma}^2 + \lambda_{maksimum}\hat{\alpha}_j^2}$ olarak tanımlanır (Muniz ve Kibria, 2009;

Kibria ve ark., 2012). $\lambda_{maksimum}$ değeri $X'WX$ matrisinin en büyük özdeğeridir (Kibria ve ark., 2013).

5.6.3. Modified Jackknifed Poisson Ridge Tahmin Edici (MJPR)

Poisson regresyon sayılabilir veri analizinde, uygulamalı araştırmalarda oldukça yaygındır. Bununla birlikte, bağımsız değişkenler yüksek olarak birbiri ile bağlantılı olduğunda poisson regresyon modelinde çoklu doğrusal bağlantı sorunu ortaya çıkmaktadır. Çoklu doğrusal bağlantı olmasından dolayı ortaya çıkan genel problemlere uygulanan yaygın çözüm yanlı tahmin edicilerin kullanılmasıdır. Son dönemlerde, poisson regresyon modelindeki ridge parametresi k 'yı tahmin edebilmek için çeşitli metodlar önerilmiştir. Bu metotlara dayalı bazı tahmin edicilerin, yaygın olarak kullanılan en çok olabilirlik tahmin ediciden daha iyi olduğu görülmüştür.

Türkan ve Özel (2016), çoklu doğrusal bağlantı sorununa çözüm getirmek için yapmış oldukları çalışmada modified jackknifed poisson regresyon (MJPR) tahmin ediciyi önermişlerdir. Bir simülasyon çalışması ve bir gerçek veri örneğini kullanarak tahmin edicilerin performansları değerlendirilmiştir. Hem hata kareler ortalaması hem de göreceli hata yüzdesi performansı tahmin edicileri karşılaştırmada performans kriteri olarak kabul edilmiştir. Simülasyon çalışması ve gerçek veri örneği sonuçları, çalışmada değerlendirilen tüm farklı durumlar için

modified jackknifed poisson regresyon (MJPR) metodunun poisson ridge regresyon, jackknifed poisson ridge regresyon ve MLE metotlarından çok daha iyi sonuçlar verdiğini göstermiştir. Jackknifed ridge regresyon (JRR) tahmin edici Singh ve ark. (1986) tarafından

$$\hat{\gamma}_{JRR} = (I + kA^{-1})^{-1}\hat{\gamma}_{RR} = (I - k^2A^{-2})\hat{\gamma}_{OLS} \quad (5.61)$$

şeklinde önerilmiştir. Bu tahmin ediciyi referans alarak Batah ve ark. (2008) tarafından önerilen modified jackknifed ridge regresyon (MJRR) tahmin edici ise

$$\begin{aligned} \hat{\gamma}_{MJRR} &= (I - k^2A^{-2})\hat{\gamma}_{RR} \\ &= (I - k^2A^{-2})(I - kA^{-1})\hat{\gamma}_{OLS} \end{aligned} \quad (5.62)$$

şeklinde tanımlanmıştır. Burada $A = (\Lambda + kI)$ olup $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$ özdeğerler matrisidir.

$G = (g_1, g_2, \dots, g_p)$ $p \times p$ boyutlu, $X'\hat{W}X$ matrisinin normalleştirilmiş öz vektörlerinden oluşan kolon matrisi, $\Lambda_{PR} = \text{diag}(\lambda_{1PR}, \lambda_{2PR}, \dots, \lambda_{pPR})$ öyle ki $G'X'\hat{W}XG = Z'\hat{W}Z = \Lambda_{PR}$ ve $Z = XG$ olmak üzere, $\hat{\beta}^{ML}$ 'nin poisson tahmin edici aşağıdaki gibi yazılabilir:

$$\hat{\gamma}^{ML} = \Lambda_{PR}^{-1}Z'\hat{W}\hat{s} \quad (5.63)$$

$$\hat{\beta}_{ML} = G\hat{\gamma}^{ML} \quad (5.64)$$

Tekrardan poisson ridge regresyon tahmin edici,

$$\hat{\gamma}^{PRR} = (\Lambda_{PR} + kI)^{-1}Z'\hat{W}\hat{s} = (I - kB^{-1})\hat{\gamma}^{ML} \quad (5.65)$$

burada $B = (\Lambda_{PR} + kI)$ ve $k > 0$ 'dır.

$$\hat{\gamma}^{PRR} = (Z'\hat{W}Z + kI)^{-1}Z'\hat{W}\hat{s} \quad (5.66)$$

Hinkley (1977) tarafından uyarlanmış JPR tahmin edici

$$\hat{\gamma}^{JPR} = \hat{\gamma}^{PRR} + (Z'\hat{W}Z + kI)^{-1} \sum_{i=1}^n z_i' \hat{\mu}_i(\hat{s}_i - z_i' \hat{\gamma}^{PRR}) \quad (5.67)$$

olarak elde edilir ve bu eşitlik sadeleştirilse,

$$\hat{\gamma}^{JPR} = (I + B^{-1}k)\hat{\gamma}^{PRR} = (I - (B^{-1}k)^2)\hat{\gamma}^{ML} \quad (5.68)$$

eşitliği elde edilmiş olur. Burada,

$$B=(Z'\widehat{W}Z + kI) \quad (5.69)$$

dir ve yukarıdaki eşitlikten modified jackknifed poisson regresyon(MJPR) tahmin edicisi

$$\hat{\gamma}^{MJPR} = (I - B^{-2}k^2)(I - B^{-1}k)\hat{\gamma}^{ML} \quad (5.70)$$

olarak elde edilir.

$\hat{\gamma}^{PRR}$, $\hat{\gamma}^{JPR}$ ve $\hat{\gamma}^{MJPR}$ 'nin MSE değerleri sırasıyla aşağıdaki gibi elde edilir:

$$\begin{aligned} MSE(\hat{\gamma}^{PRR}) &= Var(\hat{\gamma}^{PRR}) + [bias(\hat{\gamma}^{PRR})][bias(\hat{\gamma}^{PRR})]' \\ &= (I - kB^{-1})\Lambda^{-1}_{PR}(I - kB^{-1}) + k^2B^{-1}\gamma\gamma'B^{-1} \end{aligned} \quad (5.71)$$

$$\begin{aligned} MSE(\hat{\gamma}^{JPR}) &= Var(\hat{\gamma}^{JPR}) + [bias(\hat{\gamma}^{JPR})][bias(\hat{\gamma}^{JPR})]' \\ &= (I - k^2B^{-2})\Lambda^{-1}_{PR}(I - k^2B^{-2}) + k^4B^{-2}\gamma\gamma'B^{-2} \end{aligned} \quad (5.72)$$

$$\begin{aligned} MSE(\hat{\gamma}^{MJPR}) &= Var(\hat{\gamma}^{MJPR}) + [bias(\hat{\gamma}^{MJPR})][bias(\hat{\gamma}^{MJPR})]' \\ &= (I - k^2B^{-2})(I - kB^{-1})\Lambda^{-1}_{PR}(I - kB^{-1})(I - k^2B^{-2}) \\ &\quad + k^2w_pB^{-1}\gamma\gamma'w_p \end{aligned} \quad (5.73)$$

Burada

$$w_p = (I + kB^{-1} - k^2B^{-2}) \quad (5.74)$$

dir.

5.6.3.1. Önerilen Ridge Parametresi Tahmin Yaklaşımları

Ridge parametresi k yı tahmin etmek için belirli bir kural yoktur. Bunun yanı sıra ridge regresyon modeli için birçok yöntem önerilmiştir ve bu yöntemler modified jackknifed poisson regresyon (MJPR) için de kullanılabilir. Ridge parametre tahmininde en çok kullanılan beş tahmin edici aşağıda gösterilmektedir. Bunlardan ilki, Hoerl ve Kennard (1970a, 1970b) tarafından önerilen aşağıdaki tahmin edicidir:

$$MJPR1 = \hat{k}_1 = \frac{\hat{\sigma}^2}{\hat{\alpha}_{maksimum}^2} \quad (5.75)$$

Burada $\hat{\alpha}_{maksimum}^2$, $\delta^T \hat{\beta}^{ML}$ nin en büyük elemanıdır. k için ikinci tahmin edici Schaeffer ve ark. (1984) tarafından geliştirilmiştir.

$$MJPR2 = \hat{k}_2 = \frac{1}{\hat{\alpha}^2_{maksimum}} \quad (5.76)$$

Aşağıdaki diğer tahmin ediciler ise Kibria ve ark. (2011) tarafından önerilmişlerdir.

$$MJPR3 = \hat{k}_3 = medyan(m_i^2) \quad (5.77)$$

$$MJPR4 = \hat{k}_4 = maksimum(m_i) \quad (5.78)$$

$$MJPR5 = \hat{k}_5 = maksimum\left(\frac{1}{m_i}\right) \quad (5.79)$$

Burada;

$$m_i = \sqrt{\hat{\sigma}^2 / \hat{\alpha}_i^2} \text{ ve } \hat{\alpha}_i^2 = \delta^T \hat{\beta}^{ML} \text{ nin i. elemanıdır.}$$

5.6.4. Poisson Liu Tahmin Edici

Poisson Liu tahmin edici, iteratif ağırlıklandırılmış en küçük kareler yaklaşımı kullanılarak elde edilen en çok olabilirlik çözümlerine bağlı elde edilen hata kareler ortalamasının çoklu iç ilişki durumunda bozulması problemine çözüm olarak sayım verilerinde daha başarılı sonuçlar ortaya koyabilmesi amacıyla Mannson ve ark. (2012) tarafından önerilmiştir. $\hat{\beta}_{ML}$, klasik Poisson regresyon tahminleri olmak üzere poisson Liu regresyon (PLR) tahmin edici

$$\hat{\beta}_{PLR} = (X' \hat{W} X + I)^{-1} (X' \hat{W} X + dI) \hat{\beta}_{ML} \quad (5.80)$$

olarak tanımlanır.

5.6.5. Liu Tahmin Edicide Yanlılık Parametresinin Seçimi

Mansson ve ark. (2012) tarafından bazı tahmin ediciler önerilmiştir. Bu tahmin ediciler, Liu parametresinin MSE'sinin d parametresine göre diferansiyelinin alınması suretiyle türetilbilirler ve her bir parametre için sonuç fonksiyonunun sıfıra eşitlenmesi ile çözülebilir.

Mansson ve ark. (2012) tarafından önerilen bazı tahmin ediciler aşağıdaki gibidir:

$$d2 = \text{maksimum} \left(0, \frac{1}{p} \sum_{j=1}^p \frac{\hat{\alpha}_j^2 - 1}{1/\lambda_j + \hat{\alpha}_j^2} \right) \quad (5.81)$$

$$d3 = \text{maksimum} \left(0, \text{medyan} \left(\frac{\hat{\alpha}_j^2 - 1}{1/\lambda_j + \hat{\alpha}_j^2} \right) \right) \quad (5.82)$$

$$d4 = \text{maksimum} \left(0, \text{maksimum} \left(\frac{\hat{\alpha}_j^2 - 1}{1/\lambda_j + \hat{\alpha}_j^2} \right) \right) \quad (5.83)$$

$$d5 = \text{maksimum} \left(0, \text{minimum} \left(\frac{\hat{\alpha}_j^2 - 1}{1/\lambda_j + \hat{\alpha}_j^2} \right) \right) \quad (5.84)$$

Burada,

$\hat{\alpha}_j$, $\hat{\alpha} = \phi \hat{\beta}_{MLE}$ matrisinin j. bileşenidir.

ϕ ise $X'WX$ matrisinin öz vektörlerinin sütunudur.

Simülasyon çalışmasının yapılarak açıklayıcı değişkenlerin birbiri ile ilişkili olması durumunda klasik MLE, ridge ve Liu tahmin edicilerinin MSE performanslarını karşılaştırılarak en uygun parametreler belirlenebilir.

5.7. Log Normal Poisson Regresyon Modeli

Bir log-normal poisson regresyon modeli (Breslow, 1984) oluşturabilmek için öncelikle log-normal dağılımı incelememiz gerekir. $E(\epsilon_i) = e^{\sigma^2/2}$ ve $\text{Var}(\epsilon_i) = e^{\sigma^2}(e^{\sigma^2} - 1)$ olmak üzere ϵ_i log-normal dağılıma sahip bir rasgele değişken ise

$$\epsilon_i \sim \ln N(0, \sigma^2)$$

olarak gösterilir. Bu değişkenin beklenen değeri ve varyansı ise

$$E(y_i|x_i) = \exp(x_i^T \beta + \frac{\sigma^2}{2})$$

$$\text{Var}(y_i|x_i) = E(y_i|x_i) + (e^{\sigma^2} - 1)E^2(y_i|x_i)$$

şeklinde tanımlanır (Breslow, 1984; Winkelmann, 2008; Agresti, 2015).

NBR modeline kıyasla, e_i dışlanmışsa (farklı bir formdaysa), y_i dağılımı için analitik bir form yoktur ve MLE doğrudan hesaplanamayabilir. Bununla birlikte, Winkelmann (2008), teorik olarak avantajlı olduğu ve verilere daha iyi uyabildiği için log-normal poisson modelini yeniden değerlendirmeyi önermiştir (Zhou ve ark., 2012).

5.8. Poisson Hurdle Regresyon Modeli

Regresyon modelinde yanıt verisinin sayım (frekans) ölçümlerinden oluştuğu durumda gözlenen yanıt değerleri çok sayıda sıfır değeri içeriyorsa, bu değerlerin model üzerinde olumsuz etkiye sahip olur ve modelleme süreci için farklı bir yaklaşım gerekir. Bol sıfırlı modellerle benzer amaçlarla bu soruna çözüm amacıyla hurdle regresyon modelleri kullanılabilir. Hurdle regresyon yaklaşımı ilk olarak Cragg (1971) tarafından sansürlü veriler için önerilmekle birlikte Mullahy (1986) tarafından sayım verilerinde uygulamaları ile birlikte genel yapısı paylaşılmıştır.

Hurdle regresyon modeli iki aşamalı bir süreç içerir. Bu amaçla yanıt değişkeninin aldığı değerin sıfır veya sıfırdan farklı olma durumuna göre farklı hesaplamalar yapılır. Bu amaçla yanıt değişkenine sıfır ve pozitif değerlere karşılık olarak ikili yanıt muamelesi yapılır. Yanıt değişkeninin sıfır değerini aldığı durumlara karşılık bir olasılık tanımlar ve iki aşamalı sürecin ilk aşamasıdır. Sıfırdan farklı değerler için ise kesilmiş (truncated) dağılım kullanır ve bu da süreçinde ikinci aşamasıdır (Cameron ve Trivedi, 2013; Agresti, 2015). Hurdle modelleri aşırı yayılım (over-dispersion) ve eksik yayılım(under-dispersion) durumunda da esnek ve kullanılabilir (Saffari ve ark., 2012).

Bol sıfırlı modeller ile Hurdle modeller benzer kullanım amaçlarına sahip olmalarına rağmen aralarında kavramsal ve yapısal olarak farklılıklar bulunur. Her iki sınıftaki modeller iki aşamalı süreçten oluşur ancak bol sıfırlı modellerde sıfırdan farklı değerler tahmin için model oluşturulurken yanıtın hala sıfır değeri alma durumu söz konusudur. Hurdle modelde ise sıfırdan farklı yanıt değerlerinin

modellemesi yapılırken yanıtın pozitif olma koşulu vardır. Yapısal olarak ise her iki model sıfır değerlerinin tahmin modelini oluştururken yanıtta ikili değişken muamelesi yaparken, sıfırdan farklı değerlerde bol sıfırlı modeller klasik kesikli dağılımları kullanırken, Hurdle modeller kesilmiş (truncated) kesikli dağılımları sürece dahil eder. Hurdle regresyon modellerinin log olabilirlik fonksiyonun genel yapısı,

$$\begin{aligned} L(\delta_1, \delta_2) &= L_1(\delta_1) + L_2(\delta_2) \\ &= \sum_{i=1}^n [(1 - c_i) \ln f_1(0|x_i, \delta_1) + c_i \ln(1 - f_1(0|x_i, \delta_1))] \\ &\quad + \sum_{i=1}^n c_i [\ln f_2(y_i|x_i, \delta_2) - \ln(1 - f_2(0|x_i, \delta_2))] \end{aligned}$$

şeklindedir. Burada c_i ,

$$c_i = \begin{cases} 1, & y_i > 0 \\ 0, & y_i = 0 \end{cases}$$

olarak tanımlanan bir gösterge değişkendir. $f_1(0|x_i, \delta_1)$, sıfırların olasılıklarını ve $f_2(y_i|x_i, \delta_2)$ ise sıfırdan farklı yanıtların olasılıklarına karşılık gelir. Dolayısıyla $L_1(\delta_1)$ sıfır ve sıfırdan farklı değerli birimleri ayıran ikili bir sürece, $L_2(\delta_2)$ ise sıfırdan farklı (pozitif) yanıt değerleri için kesilmiş (truncated) dağılımdan elde edilen log olabilirlik fonksiyonuna karşılık gelir (Cameron ve Trivedi, 2013).

Hurdle regresyon modelleri oluşturulurken pratikte f_1 ve f_2 fonksiyonlarının (dağılımlarının) belirlenmesi gerekir. Genellikle ilk aşamayı oluşturan ikili süreç için f_1 fonksiyonunda logit veya probit modeller kullanılırken, sıfırdan farklı yanıt değerlerini tahmin için oluşturulan ikinci aşama için poisson veya negatif binom dağılımları tercih edilir (Cameron ve Trivedi, 2013). Burada ikinci aşamada kullanılan dağılımlar klasik dağılımlar olmayıp, kesilmiş (truncated) versiyonlarıdır.

5.9. Poisson Regresyonda Uyum İyiliği Testleri

Bir modelden elde edilen sonuçların kullanılabilir olması modelin geçerlilik derecesine bağlıdır. Sahip olduğumuz veri kümesi için birden fazla regresyon modeli olmasına rağmen bu modellerden veri kümesini en iyi açıklayacak modeli seçmek oldukça önemlidir. Poisson regresyonda normallik varsayımı sağlanmadığı için modelin uyum iyiliği testlerinden G^2 , Pearson χ^2 , Yapay R^2 ve AIC gibi uyum iyiliği testlerinden yararlanılabilir.

5.9.1 Sapma İstatistiği (D)

Uyum iyiliği ölçümünde kullanılan bir istatistiktir ve

$$D = 2 \sum_{i=1}^n \left[y_i \ln \left(\frac{y_i}{\mu_i} \right) - (y_i - \mu_i) \right] \quad (5.85)$$

biçiminde tanımlanır.

Eğer modelde

$$\sum_{i=1}^n y_i = \sum_{i=1}^n \hat{\mu}_i \quad (5.86)$$

ise olabilirlik oranı uyum iyiliği istatistiği

$$D = 2 \sum_{i=1}^n y_i \ln \left(\frac{y_i}{\mu_i} \right) \quad (5.87)$$

şeklinde olacaktır.

Elde edilen sapma istatistiğinin değeri 0'a yakınsıyor ise model uyumu artıyor denilebilir. Eğer bulunan sapma istatistiği değeri 0 ise model uyumunun mükemmel olduğu söylenebilir (Cameron ve Trivedi, 201; Deniz, 2005).

5.9.2 Pearson Ki-Kare İstatistiği (χ^2)

Pearson χ^2 istatistiği incelenilen model için beklenen frekanslar ile gözlenen frekanslar dikkate alınarak hesaplanmaktadır. Uyum iyiliğini test etmek için kullanılacak olan hipotezler

H_0 = Veriler $f(x)$ dağılımına uygunluk göstermektedir.

H_1 = Veriler $f(x)$ dağılımına uyum göstermemektedir.

biçiminde olacaktır. Pearson ki-kare istatistiği,

$$\chi^2 = \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{\hat{\omega}_i} \quad (5.88)$$

biçimindedir. Bu değer seride aşırı yayılım olup olmadığını belirlemek amacı ile kullanılır. Burada $\hat{\mu}_i$ ve $\hat{\omega}_i$ değerleri, μ_i ve ω_i ' nin tahmin değerleridir. Hesaplanan χ^2 değeri $\hat{\mu}$ için belirlenen serbestlik derecesi (n-k) ile karşılaştırılır.

Bu formül poisson regresyon modeli için uygulandığında $\omega_i = \mu_i$ olacaktır ve

$$\chi^2 = \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} \quad (5.89)$$

olarak elde edilir.

Hesaplanan χ^2 değerinin serbestlik derecesine oranının 1'den büyük değer alması seride aşırı yayılımın olduğunu ve verilerin modele uygun olmadığını, 1'den küçük değer alması seride az yayılımın olduğunu ifade etmektedir (Deniz, 2005).

5.9.3. Yapay R^2 İstatistiği

Poisson regresyon modellerinin değerlendirilmesinde bir ölçüm olarak yapay R^2 istatistiği kullanılmaktadır. Doğrusal regresyon modellerinde, R^2 'nin hesaplanabilmesi için öncelikle genel kareler toplamı ayrıştırılmalıdır. Bu toplam

$$\begin{aligned} \sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n (y_i - \hat{\mu}_i)^2 + \sum_{i=1}^n (\hat{\mu}_i - \bar{y})^2 \\ &+ 2 \sum_{i=1}^n (y_i - \hat{\mu}_i)(\hat{\mu}_i - \bar{y}) \end{aligned} \quad (5.90)$$

şeklinde gösterilir. Bu ifade poissonu da içeren ve doğrusal olmayan EKK ile üstel koşullu ortalamaya sahip tüm kestiriciler ve modeller için sıfırdan farklı değerli

olacaktır. Bu durumda R^2 'nin doğrusal regresyondan farklı bir yöntemle hesaplanması gerekmektedir.

Normallik varsayımı gerektirmeyen poisson regresyon modelinde R^2 ölçüsü olabilirlik oranı yaklaşımına dayanmaktadır. Doğrusal regresyon modeline ilişkin EKK tahmini AKT'nin en çok olabilirlik tahmini ve sapma değeri ile benzer özellikte olması nedeniyle R_p^2 ölçüsü

$$R_p^2 = 1 - \frac{l(y) - l(\hat{\mu})}{l(y) - l(\bar{y})} \quad (5.91)$$

şeklinde tanımlanmaktadır. Burada $l(y)$ doymuş (saturated) modelin log-olabilirliğini, $l(\hat{\mu})$ ilgili modelin log-olabilirliğini ve $l(\bar{y})$ sadece sabit terimin bulunduğu minimal modelin log-olabilirliğini göstermektedir.

$y_i \geq 0$ gözlenen değerler, $\hat{\mu}_i = \exp(x_i' \hat{\beta})$ tahmin edilen değerleri ve $\bar{y}_i = \exp(\hat{\beta}_0)$ ortalama değerleri olmak üzere log-olabilirlik fonksiyonları sırasıyla

$$l(y) = \sum_{i=1}^n (y_i \ln(y_i) - y_i - \ln(y_i!)) \quad (5.92)$$

$$l(\hat{\mu}) = \sum_{i=1}^n (y_i \ln(\hat{\mu}_i) - \hat{\mu}_i - \ln(y_i!)) \quad (5.93)$$

$$l(\bar{y}) = \sum_{i=1}^n (y_i \ln(\bar{y}_i) - \bar{y}_i - \ln(y_i!)) \quad (5.94)$$

biçiminde elde edilmektedir. Yukarıdaki log-olabilirlik fonksiyonları düzenlendiği takdirde yapay R^2 ölçüsüne ulaşılabilmektedir.

Poisson regresyon analizinde, modeldeki değişkenlerin modele katkısının testi serbestlik derecesine ilişkin $n - p \geq 0$ kısıtı sağlandıktan sonra sapma farkına bakılarak yapılmaktadır. Bu fark değeri, ilgili ki-kare tablo değeri ile karşılaştırılarak test edilen değişkenin modele olan katkısının önemli düzeyde olup olmadığı yorumu yapılabilir.

5.9.4 Akaike Bilgi Ölçütü

Akaike Bilgi Ölçütü 1973 yılında Akaike tarafından geliştirilmiştir. Bu kriter verilen sonucu farklı modeller ile karşılaştırmak amacı ile geliştirilmiştir. Basit model gerçek modelden daha karmaşık olmasına rağmen gerçek modelin bazı özelliklerinin tahminlerini daha iyi belirlediği için tercih edilme sebebi olabilir.

Akaike bilgi ölçütü

$$AIC = 2(q - l) \quad (5.95)$$

şeklinde verilir. Burada q tahmin edilen parametre sayısını göstermektedir (Cameron ve Trivedi, 2013; Deniz, 2005).

5.9.5 Pearson İstatistiği

μ_i ortalamalı ve ω_i varyanslı bağımlı değişken y_i 'ye ait herhangi bir model için standart uyum iyiliği ölçüm yöntemi Pearson istatistiğidir ve

$$P = \sum_{i=1}^n \frac{(y_i - \hat{\mu})^2}{\hat{\omega}_i} \quad (5.96)$$

olarak ifade edilir. Bu değer serinin yayılımının aşırı olup olmadığını belirlemede kullanılmaktadır. Burada $\hat{\mu}_i$ ve $\hat{\omega}_i$ değerleri μ_i ve ω_i 'nin kestirim değerleridir. Hesaplanan P değeri $\hat{\mu}$ değeri için belirlenmiş olan serbestlik derecesi ($n-q-1$) ile karşılaştırılır. Burada;

$$P_p > n - q - 1 \text{ ise seride aşırı yayılım,}$$

$$P_p < n - q - 1 \text{ ise seride eksik yayılım olduğu söylenebilir.}$$

Aşırı yayılım durumunda negatif binom regresyon modeli önerilmiştir (Cameron ve Trivedi, 2013; Deniz, 2005).

5.10. Tahmin Edicilerin Performansının Değerlendirilmesi

Tahmin metotlarının performanslarını incelemek için MSE değeri

$$MSE = \frac{\sum_{i=1}^R (\hat{\beta} - \beta)' (\hat{\beta} - \beta)}{R} \quad (5.97)$$

eşitliği ile hesaplanırken MAE değeri

$$MAE = \frac{\sum_{i=1}^R |\hat{\beta} - \beta|}{R} \quad (5.98)$$

ile bulunur. Burada $\hat{\beta}$, β 'nın tahmin edicisini ve R değeri Monte Carlo simülasyonunda kullanılan tekrar sayısını ifade eder (Kibria ve ark., 2011).



6. NEGATİF BİNOM REGRESYON ANALİZİ

Poisson regresyon analizi yapılırken birçok kez varyansın ortalamadan büyük olduğu yani aşırı yayılım olduğu durumla karşılaşılır. Bu gibi durumlarda farklı poisson regresyon türleri kullanılmalıdır. Bunlardan bir tanesi de karışık poisson regresyondur. Karışık poisson regresyon modellerinden en çok kullanılan model ise negatif binom regresyon modelidir.

Poisson regresyon modeli sayım verileri için genellikle kısıtlıdır. Bu durum, negatif binom regresyon modeli gibi alternatif modellerin ortaya çıkmasına neden olmaktadır. Bu kısıtlamanın kendini gösterdiği ilk durum poisson dağılımında sıfır sayılarının olasılığının tahminlerinin, verilerdeki gerçek gözlemlerden oldukça düşük olmasıdır. Verideki poisson tahminlerinden daha fazla sıfır içerdiği için, bu aşırı sıfır problemi olarak adlandırılır. İkinci ve en açık durum ise, poissonda varyansın genellikle ortalamadan yüksek olması sayım verileri için eksiklik olan özellik aşırı yayılım adını alır. Eğer dağılımda aşırı yayılım anlamlı ise, poisson regresyon modelinden elde edilen tahminler tutarlıdır fakat yeterli değildir. Poisson regresyon modelindeki tahmin edicilerin standart hataları yanlı-azalan olacaktır. Bu durum katsayıların anlamlılığı hakkında araştırmacının yanlış istatistiksel yorumlar yapmasına neden olabilir. Negatif binom regresyon modeli poisson regresyon modeli için bir seçenek sağlar. Negatif binom regresyon modeli sadece aşırı yayılım durumuna seçenek olarak kullanılır. Negatif binom regresyonuna poisson-gamma regresyon modeli de denilmektedir.

6.1 Negatif Binom Regresyon Modeli (NBR)

$Y \sim \text{poisson}(\mu)$ ve $\mu \sim \text{gamma}(\alpha, P)$ dağılımlarına sahip rastgele değişkenler olsun. $\text{Gamma}(\alpha, P)$, αP ortalamalı ve αP^2 varyanslı gamma dağılımıdır. Bu dağılıma ait olasılık-yoğunluk fonksiyonu;

$$p(\mu) = \frac{1}{P^\alpha \Gamma(\alpha)} \mu^{\alpha-1} \exp\left(-\frac{\mu}{P}\right), \quad \mu > 0 \quad (6.1)$$

$$= 0, \quad d.d.$$

$$f(y) = \frac{\Gamma(\alpha + y)}{\Gamma(\alpha)\Gamma(y + 1)} \left(\frac{P}{1 + P}\right)^y \left(\frac{1}{1 + P}\right)^\alpha \quad y = 0, 1, 2, \dots, P \quad \alpha > 0 \quad (6.2)$$

ile gösterilir. Yukarıda gösterilen dağılımın ortalaması ve varyansı

$$E(Y) = E(E(Y|\mu)) = E(\mu) = \alpha P \quad (6.3)$$

$$\begin{aligned} V(Y) &= E(V(Y|\mu)) + V(E(Y|\mu)) \\ &= V(\mu) + E(\mu) = \alpha P + \alpha P^2 \end{aligned} \quad (6.4)$$

şeklindedir. Negatif binom dağılımı, poisson ve gamma dağılımının birleşimlerinden elde edilmektedir (Cameron ve Trivedi, 2013; Agresti, 2015).

6.1.1. Doğrusal Varyans Fonksiyonu ile Negatif Binom Dağılımı(NB1)

Poisson regresyon modelinde y_i , $\mu_i = \exp(x_i'\beta)$ ortalamalı ve μ_i varyanslıdır. Çoğu zaman gerçek yaşam verilerinde ortalamaya ilişkin varsayım korunurken varyansın ortalamaya eşit olduğu varsayımı genellikle doğrulanmaz ve ortalamanın $\exp(x_i'\beta)$ varsayımı kabul edilir. Bu nedenle y_i 'nin koşullu varyansı için genel gösterim,

$$\omega_i = V[y_i/x_i] \quad (6.5)$$

biçimindedir. Ortalamanın bir fonksiyonu olarak varyans modeli;

$$\omega_i = \omega(\mu_i, \alpha) \quad (6.6)$$

şeklinde tanımlanabilir. Bu eşitlikte $\omega(.)$ özel tanımlı bir fonksiyon ve α sabit skaler bir parametredir. Çoğu model bu fonksiyonu aşağıda verilen eşitlik ile tanımlanan genel varyans fonksiyonuyla genelleştirir:

$$\omega_i = \mu_i + \alpha \mu_i^p \quad (6.7)$$

Yukarıdaki fonksiyonda p parametresi sabit olup özel olarak tanımlanır. Analiz genel olarak $\alpha = 0$ olan poisson durumuna ek olarak iki özel durumla sınırlandırılır.

İlk olarak, $p=1$ olarak alındığında elde edilen NB1 varyans fonksiyonunda varyans

$$\omega_i = (1 + \alpha)\mu_i \quad (6.8)$$

olarak ortalamanın katı olarak tanımlanır. Bu ifade genel doğrusal modeller sürecinde genellikle

$$\omega_i = \phi\mu_i \quad (6.9)$$

olarak elde edilir. Burada $\phi = 1 + \alpha$ 'dır (Cameron ve Trivedi, 2013; Agresti, 2015).

6.1.2. Karesel Varyans Fonksiyonu ile Negatif Binom Dağılımı(NB2)

İkinci olarak, $p=2$ olarak alındığında elde edilen NB2 varyans fonksiyonunda varyans ortalamanın karesel bir formu olup

$$\omega_i = \mu_i + \alpha\mu_i^2 \quad (6.10)$$

şeklinde tanımlanır.

Her iki durumda da yayılım parametresi α tahmin edilmesi gereken bir parametredir. Bu varyans fonksiyonları iki yerde kullanılabilir. Bunlardan birincisi poisson quasi en çok olabilirlik (QMLE)'den daha etkin tahmin ediciler sağlamak için, ikincisi ise poisson QMLE için alternatif standart hatalar elde etmek içindir (Cameron ve Trivedi, 2013; Agresti, 2015).

6.2. En Çok Olabilirlik Tahmin Edici (MLE)

6.2.1. NB1 Model ve En Çok Olabilirlik Tahmini

NB1 model, klasik doğrusal regresyon modeli yaklaşımında kullanılan varyans fonksiyonu ile benzerlik gösterir. NB1 log-olabilirlik fonksiyonu,

$$\ln L(\beta) = \sum_{i=1}^n \left\{ \sum_{j=0}^{y_i-1} \ln(j + \alpha^{-1} - \exp(x_i'\beta)) - \ln y_i! \right\} - (y_i \alpha^{-1} \exp(x_i'\beta)) \ln(1 + \alpha) + y_i \ln \alpha \quad (6.11)$$

ile gösterilir. NB1 en çok olabilirlik tahmininin ilk sıra koşulları çözümü aşağıdaki gibidir.

$$\sum_{i=1}^n \left\{ \left(\sum_{j=0}^{y_i-1} \frac{\alpha^{-1} \mu_i}{j + \alpha^{-1} \mu_i} \right) x_i + \alpha^{-1} \mu_i x_i \right\} = 0 \quad (6.12)$$

$$\frac{1}{\alpha^2} \left\{ - \left(\sum_{j=0}^{y_i-1} \frac{\mu_i}{j + \alpha} \right) - \alpha^{-1} \mu_i \ln(1 + \alpha) - \frac{1}{1 + \alpha} + y_i \alpha \right\} = 0 \quad (6.13)$$

NB1 olasılık yoğunluk fonksiyonunun ilk iki momentine dayalı tahminler NB1 en çok olabilirlik tahmini olarak adlandırılan poisson en çok olabilirlik tahminlerini verir (Cameron and Trivedi, 2013).

6.2.2. NB2 Model ve En Çok Olabilirlik Tahmini

Negatif binom dağılımının en genel uygulaması $\mu + \alpha\mu^2$ varyans fonksiyonu ile NB2 modeldir. Bu modelin yoğunluk fonksiyonu,

$$f(y|\mu, \alpha) = \frac{\Gamma(y + \alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(y + 1)} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu} \right)^{\alpha^{-1}} \left(\frac{\mu}{\alpha^{-1} + \mu} \right)^y, \quad (6.14)$$

$$\alpha \geq 0, y = 0, 1, ..$$

olarak bulunur. $\alpha = 0$ durumunda fonksiyon poisson dağılımına indirgenir. Üstel ortalama $\mu_i = \exp(x_i' \beta)$ için log olabilirlik fonksiyonu

$$\ln L(\alpha\beta) = \sum_{i=1}^n \left\{ \begin{array}{c} \sum_{j=0}^{y_i-1} \ln(j + \alpha^{-1}) - \ln y_i! \\ -(y_i + \alpha^{-1}) \ln(1 + \alpha \exp(x_i' \beta)) + y_i \ln \alpha + y_i x_i \beta \end{array} \right\} \quad (6.15)$$

şeklinde tanımlanır.

NB2 en çok olabilirlik tahmini

$$\sum_{i=1}^n \frac{y_i + \mu_i}{1 + \alpha\mu_i} x_i = 0 \quad (6.16)$$

$$\sum_{i=1}^n \left\{ \frac{1}{\alpha} \left(\ln(1 + \alpha\mu_i) - \sum_{j=0}^{y_i-1} \frac{1}{(j + \alpha^{-1})} \right) + \left\{ \frac{\mu_i - x_i}{\alpha(1 + \alpha\mu_i)} \right\} \right\} = 0 \quad (6.17)$$

sıra koşulunun çözümü olan (6.17) ile verilen denklemdir (Cameron ve Trivedi, 2013; Agresti, 2015).

6.3. Negatif Binom Regresyonda Bazı Yanlı Tahmin Ediciler

6.3.1. Negatif Binom Ridge Tahmin Edici (NBRR)

Negatif binom ridge regresyon modeli sayım verilerinin analizinde uygulamalı çalışmalarda oldukça yaygın kullanılmaktadır. Parametre tahmininde kullanılan en çok olabilirlik tahmin edici yüksek seviyede iç ilişkili değişkenlerin varlığı durumunda çok hassastır. Bu nedenle çoklu iç ilişkinin varlığı durumunda negatif binom regresyon modelinde parametrelerin tahmini için dayanıklı bir seçenek olarak Mansson (2012) tarafından negatif binom ridge regresyon tahmin edici (NBRR) önerilmiştir. Bu tahmin edici

$$\hat{\beta}_{NBR} = (X'W_{\hat{\beta}_{ML}}X + kI)^{-1}(X'W_{\hat{\beta}_{ML}}X)\hat{\beta}_{ML} \quad (6.18)$$

şeklinde tanımlanır. Bu tahmin edicinin MSE değeri

$$\begin{aligned} E(L_{NBR}^2) &= (\hat{\beta}_{NBR} - \beta)'(\hat{\beta}_{NBR} - \beta) \\ &= E \left[(\hat{\beta}_{ML} - \beta)'Z'Z(\hat{\beta}_{ML} - \beta) \right] + (Z\hat{\beta}_{ML} - \beta)'(Z\hat{\beta}_{ML} - \beta) \\ &= \sum_{j=1}^J \frac{\eta_j}{(\eta_j + k)^2} + k^2 \sum_{j=1}^J \frac{\alpha_j^2}{(\eta_j + k)^2} = \gamma_{1(k)} + \gamma_{2(k)} \end{aligned} \quad (6.19)$$

şeklinde elde edilebilir. Burada α_j^2 değeri $\gamma\hat{\beta}_{ML}$ 'nin j. elemanı ve γ değeri $\Lambda = \text{diag}(\eta_j)$ olan $X'W_{\hat{\beta}_{ML}}X = \gamma'\Lambda\gamma$ olacak şekilde tanımlanan özvektöre karşılık gelir (Mansson, 2012).

Negatif Binom Ridge Regresyon Tahmin Edicinin Özellikleri

1. Toplam varyans $\gamma_{1(k)}$, sürekli ve k'nın monoton azalan bir fonksiyonudur.
2. Karesel yanlılık $\gamma_{2(k)}$, sürekli ve k'nın monoton artan bir fonksiyonudur.
3. Her zaman $k > 0$ olan bir k değeri vardır:

$$E(L_{RR}^2) < E(L_{ML}^2) = \sum_{j=1}^J \frac{1}{\eta_j} \quad (6.20)$$

Ridge parametresi için optimal k değeri ise;

$$k = \frac{1}{\alpha_j^2} \quad (6.21)$$

dir (Hoerl and Kennard, 1970a).

6.3.1.1. Negatif Binom Ridge Tahmin Edicide k Parametresinin Seçimi

Ridge parametresini tahmin etmek için Hoerl ve Kennard (1970a, 1970b), Kibria (2003), Alkhamisi ve ark. (2006) ve Muniz ve Kibria (2009) tarafından önerilen birden çok metot bulunmaktadır.

$$K1 = \hat{k}_{HK1} = \frac{\hat{\sigma}^2}{\hat{\alpha}_{\text{maksimum}}^2} \quad (6.22)$$

Hoerl ve Kennard (1970a, 1970b) tarafından önerilmiştir.

Burada $\hat{\alpha}_{\text{maksimum}}^2$ $\gamma \hat{\beta}_{ML}$ 'nin en büyük elemanıdır. γ , $X'WX$ matrisinin özvektörü ve $\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{\mu}_i)^2}{n-p-1}$ 'dir.

Sıradaki iki tahmin edici lineer regresyon için Kibria (2003) tarafından önerilmiş tahmin edicilerdir.

$$K2 = \hat{k}_{GM} = \frac{\hat{\sigma}^2}{(\prod_{i=1}^p \hat{\alpha}_j^2)^{1/p}} \quad (6.23)$$

$$K3 = \hat{k}_{MED} = \text{medyan} \{m_j^2\} \quad (6.24)$$

Burada $m_j = \sqrt{\hat{\sigma}^2 / \hat{\alpha}_j^2}$ dir.

$$K4 = \hat{k}_{\text{maksimum}}^{KS} = \text{Maksimum}(s_j) \text{ (Alkamisi ve ark., 2006)}$$

Bu eşitlikte

$$s_j = \frac{t_j \hat{\sigma}^2}{(n-p)\hat{\sigma}^2 + t_j \hat{\alpha}_j^2} \quad (6.25)$$

ve t_j , $X'X$ matrisinin öz değeridir. Eşitlik (6.26) – (6.28) ile verilen tahmin ediciler ise Muniz ve Kibria (2009) tarafından önerilmiştir:

$$K5 = \hat{k}_{KM2} = \text{maksimum} \left(\frac{1}{m_j} \right), \quad (6.26)$$

$$K6 = \hat{k}_{KM4} = \prod_{j=1}^p \left(\frac{1}{m_j} \right)^{\frac{1}{p}}, \quad (6.27)$$

$$K7 = \hat{k}_{KM6} = \text{medyan} \left(\frac{1}{m_j} \right). \quad (6.28)$$

6.3.2. Negatif Binom Liu Tahmin Edici

Negatif binom regresyon modeli sağlık, sosyal, ekonomik ve fiziki bilimler alanlarında, bağımlı değişken negatif olmayan tamsayılar veya sayımlar şeklinde geldiğinde çok popülerdir. Bu model genellikle en çok olabilirlik tahmin edici (ML) ile tahmin edilir, doğrusal olmayan bir denklemin çözümü ise iteratif ağırlıklı en küçük kareler algoritması uygulanarak bulunur. Bu tahmin yönteminin bir problemi regresyon modelindeki açıklayıcı değişkenler birbiri ile yüksek derecede ilişkili ise ortaya çıkar. Bu problem MLE yöntemi ile tahmin edilen parametrelerin varyansının yüksek olmasına ve durağan olmamasına yol açar (Mannson, 2013).

Yanlı tahmin edicileri uygulamak günümüzde çoklu doğrusal bağlantının neden olduğu problemleri çözmek için etkili bir iyileştirici yaklaşım olarak uygulanmaktadır. Mannson (2013) tarafından doğrusal regresyon modeli için Liu tarafından önerilen tahmin ediciyi kullanarak negatif binom regresyon modeli için yeni bir yanlı tahmin edici önerilmiştir. Ayrıca yanlılık parametresi d'yi tahmin etmek için bazı metotlar da gösterilmiştir.

Negatif binom regresyon modeli genellikle en çok olabilirlik tahmin edici yöntemi ile tahmin edilir. Bu tahmin edici, ağırlıklandırılmış hata kareler ortalaması (WMSE) toplamını en aza indirir, ancak çoklu doğrusal bağlantı varlığında, ağırlıklandırılmış çapraz çarpım matrisi $X'\hat{W}X$ kötü koşullu olur ve bu

MLE yöntemi ile tahmin edilen parametrelerin varyansının yüksek olmasına ve durağan olmamasına yol açar. IWLS yaklaşık olarak ağırlıklı hata kareler ortalamasını en aza indirdiğinden doğrusal regresyon için Liu yönteminin bir genelleştirmesi olarak uygulanabilir ve aşağıdaki yanlı tahmin edici elde edilir (Mannson, 2013):

$$\hat{\beta}_d = (X'WX + I)^{-1} (X'WX + dI) \hat{\beta}_{ML} \quad (6.29)$$

Yanlılık parametresi d , 0 ile 1 arasında değer alır ve nasıl tahmin edileceğine ilişkin kesin bir kural yoktur.

6.3.2.1. Negatif Binom Liu Tahmin Edicide Parametre Seçimi

Yanlılık parametresi d 'nin en iyi değerini tahmin etmek için birçok yöntem önerilmiştir. Liu tahmincisinin tanımına göre, yanlılık parametresinin optimal değeri, tek bir d değeri ile tahmin edilecektir.

$$D_1 = \text{maksimum} \left(0, \frac{\hat{\alpha}_{\text{maksimum}}^2 - 1}{\frac{1}{\hat{\lambda}_{\text{maksimum}}} + \hat{\alpha}_{\text{maksimum}}^2} \right) \quad (6.30)$$

Burada $\hat{\alpha}_{\text{maksimum}}^2$ ve $\hat{\lambda}_{\text{maksimum}}$ değerleri sırasıyla $\hat{\alpha}_j^2$ ve $X'WX$ in özdeğerlerinin maksimum bileşenleridir (Mannson, 2013).

Bilinmeyen parametrelerin değerlerini Hoerl ve Kennard (1970a, 1970b)'dan esinlenilen fikre dayanarak yansız tahmin edicilerin maksimum değerleriyle değiştirilmiştir. Bununla birlikte Liu tahmin edici için yanlılık parametresinin negatif olmamasını sağlayan başka bir maksimum bileşeni kullanılmıştır (Mannson, 2013). Ayrıca

$$D_2 = \text{maksimum} \left(0, \text{medyan} \left(\frac{\hat{\alpha}_j^2 - 1}{\frac{1}{\hat{\lambda}_j} + \hat{\alpha}_j^2} \right) \right) \quad (6.31)$$

$$D_3 = \text{maksimum} \left(0, \frac{1}{p} \sum_j^J \frac{\alpha_j^2 - 1}{\frac{1}{\hat{\lambda}_j} + \hat{\alpha}_j^2} \right) \quad (6.32)$$

$$D_4 = \text{maksimum} \left(0, \text{maksimum} \left(\frac{\hat{\alpha}_j^2 - 1}{\frac{1}{\hat{\lambda}_j} + \hat{\alpha}_j^2} \right) \right) \quad (6.33)$$

$$D_5 = \text{maksimum} \left(0, \text{minimum} \left(\frac{\hat{\alpha}_j^2 - 1}{\frac{1}{\hat{\lambda}_j} + \hat{\alpha}_j^2} \right) \right) \quad (6.34)$$

tahmin edicileri kullanılmıştır (Mannson, 2013).

6.3.3. İki Parametrelili Negatif Binom Tahmin Edici

Huang ve Yang (2014) tarafından negatif binom regresyon modelinde çoklu doğrusal bağlantıya karşı çözüm olarak iki parametrelili tahmin edici önerilmiştir. Önerilen bu tahmin edici özel durumlarda en çok olabilirlik, ridge ve Liu tahmin edicilerine denk olan genel bir tahmin edici yapısındadır.

Sırasıyla Mannson (2012) ve Mannson (2013) tarafından önerilen negatif binom regresyon ridge tahmin edici (NBRR) ve negatif binom regresyon Liu tahmin edici (NBLR)

$$\hat{\beta}_{NBRR} = (X'WX + kI)^{-1}(X'WX)\hat{\beta}_{ML} \quad (6.35)$$

$$\hat{\beta}_{NBLR} = (X'WX + I)^{-1}(X'WX + dI)\hat{\beta}_{ML} \quad (6.36)$$

şeklinde önerilmiştir. Bu iki farklı tahmin edici kullanılarak Huang ve Yang (2014) tarafından negatif binom iki parametrelili tahmin edici (NBTP)

$$\hat{\beta}_{(k,d)} = (X'WX + I)^{-1}(X'WX + dI)(X'WX + kI)^{-1}(X'WX)\hat{\beta}_{ML} \quad (6.37)$$

olarak önerilmiştir. Burada $k \geq 0$ ve $0 \leq d < 1$ 'dir.

Negatif binom iki parametrelili tahmin edici bazı özel durumlarda en çok olabilirlik, ridge ve Liu tahmin edicilerine denktir.

$$\lim_{d \rightarrow 1} \hat{\beta}_{(0,d)} = \hat{\beta}_{ML} \quad (6.38)$$

$$\lim_{d \rightarrow 1} \hat{\beta}_{(k,d)} = (X'WX + kI)^{-1}(X'WX)\hat{\beta}_{ML} = \hat{\beta}_{Ridge} \quad (6.39)$$

$$\hat{\beta}_{(0,d)} = (X'WX + I)^{-1}(X'WX + dI)\hat{\beta}_{ML} = \hat{\beta}_{Liu} \quad (6.40)$$

$\hat{\beta} = Z\hat{\beta}_{ML}$ şeklinde tanımlanan bir tahmin edicinin asimptotik MSE değeri

$$\begin{aligned} MSE(\hat{\beta}) &= E(\hat{\beta} - \beta)'(\hat{\beta} - \beta) \\ &= E(\hat{\beta}_{ML} - \beta)'Z'Z(\hat{\beta}_{ML} - \beta) + (Z\beta - \beta)'(Z\beta - \beta) \end{aligned} \quad (6.41)$$

olarak tanımlanır.

İki parametrelili tahmin edicinin tanımından hareketle MSE değeri

$$\begin{aligned} MSE(\hat{\beta}_{(k,d)}) &= \sum_{j=1}^p \frac{(\lambda_j + d)^2 \lambda_j}{(\lambda_j + 1)^2 (\lambda_j + k)^2} + \sum_{j=1}^p \frac{((k + 1 - d)\lambda_j + k)^2 \alpha_j}{(\lambda_j + 1)^2 (\lambda_j + k)^2} \\ &= \gamma_1(k, d) + \gamma_2(k, d) \end{aligned} \quad (6.42)$$

şeklinde elde edilebilir. Bu eşitlikte ilk terim asimptotik varyans, ikinci terim yanlışlık teriminin karesine karşılık gelir. İki parametrelili tahmin edicinin amacı varyans terimindeki azalmanın yanlışlık teriminin karesindeki artmaktan daha fazla olacak şekilde k ve d parametrelerini seçmektir.

6.3.3.1. İki Parametrelili Negatif Binom Tahmin Edicide Parametrelerin Seçimi

Uygulamalarda k ve d parametrelerinin seçimi karmaşık bir problemdir. Hoerl ve Kennard (1970a, 1970b) ve Kibria (2003) tarafından doğrusal ridge regresyon modeli için yanlışlık parametresi seçimine yönelik önerilen birkaç metottan hareketle Huang ve Yang (2014) k ve d parametreleri için farklı tahmin ediciler önermişlerdir.

d^* parametresi sıfır ile bir arasında sabit iken, k için optimal tahmin edici

$$K = \min \left\{ \frac{(\hat{\lambda}_j + d^*) - \hat{\lambda}_j \hat{\alpha}_j^2 (1 - d^*)}{\hat{\alpha}_j^2 (\hat{\lambda}_j + 1)}, j = 1, 2, \dots, p \right\} \quad (6.43)$$

şeklinde önerilmiş olup, bu eşitlik $(\partial MSE(\hat{\beta}_{(k,d)})) / \partial k$ değerini sıfıra eşitleyerek ve λ_j ve α_j yerine yansız tahmin edicilerini yazarak elde edilir. $d^* = 1$ iken $K = 1/\hat{\alpha}_{max}^2$ olur ve bu değer Hoerl ve Kennard tarafından ridge parametresi k 'yı tahmin etmek için önerilmiştir.

Benzer şekilde $k^* > 0$ olacak şekilde sabit bir değer alındığında d parametresinin optimal değeri

$$D_1 = \frac{(k^* \hat{\lambda}_{maks.} + \hat{\lambda}_{maks.} + k^*) \hat{\alpha}_{maks.}^2 - \hat{\lambda}_{maks.}}{\hat{\alpha}_{maks.}^2 \hat{\lambda}_{maks.} + 1}, \quad (6.44)$$

eşitliği kullanarak elde edilebilir. Burada $\hat{\lambda}_{maks.}$, $X'WX$ matrisinin en büyük özdeğeri olarak tanımlanır.

Eşitlik (6.44)'de $k^* = 0$ alındığında

$$D_1 = \max \left(0, \frac{\hat{\alpha}_{maksimum}^2 - 1}{\frac{1}{\hat{\lambda}_{maksimum}} + \hat{\alpha}_{maksimum}^2} \right) \quad (6.45)$$

alınabilir.

$MSE(\hat{\beta}_{(k,d)})$ minimize etmek için $\frac{(\partial MSE(\hat{\beta}_{(k,d)}))}{\partial d} = 0$ alındığında d parametresi için alternatif bir tahmin edici

$$\hat{d}_2 = \frac{\sum_{j=1}^p \left(((k^* + 1) \hat{\lambda}_j + k^*) \hat{\lambda}_j \hat{\alpha}_j^2 - \hat{\lambda}_j^2 \right) / (\hat{\lambda}_j + 1)^2 (\hat{\lambda}_j + k^*)^2}{\sum_{j=1}^p (1 + \hat{\lambda}_j \hat{\alpha}_j^2) \hat{\lambda}_j / (\hat{\lambda}_j + 1)^2 (\hat{\lambda}_j + k^*)^2} \quad (6.46)$$

olarak bulunabilir. Pratik amaçla

$$D_2 = \text{maksimum}(0, \hat{d}_2) \quad (6.47)$$

eşitliği de kullanılabilir.

Sonuç olarak d parametresi üzerinde

$$\max \left(0, \frac{\hat{\lambda}_{maks.} (1 - \hat{\alpha}_{maks.}^2)}{\frac{1}{\hat{\alpha}_{maks.}^2} (1 + \hat{\lambda}_{maks.})} \right) < k < \frac{1}{\hat{\alpha}_{maks.}^2} \quad (6.48)$$

şeklinde bir kısıt elde edilir. Bu kısıt altında,

D_1 'in her zaman birden küçük ve sıfırdan büyük olduğu söylenebilir. Dahası D_2 'nin birden küçük ve negatif olmayacağı söylenebilir. k ve d parametrelerinin seçimi aşağıdaki adımlar kullanılarak bulunabilir:

Adım 1. Son eşitliğe uygun olarak k belirlenir.

Adım 2. Belirlenen k değeri kullanılarak D_1 ve D_2 hesaplanır.

Adım 3. D_1 ve D_2 değerleri kullanılarak K değeri elde edilir.

Adım 4. Elde edilen K negatif ise $K=k$ alınır.

6.4. Lognormal ve Gamma Mixed Negatif Binom Regresyon Modeli

Tahmindeki belirsizlikleri açık bir şekilde modellemek ve önceki bilgileri birleştirmek için, Bayesyen yaklaşımları avantajlı görünmektedir. Bununla birlikte, sayımların Bayes analizi, etkili çıkarımın olmayışı ile ciddi bir şekilde sınırlıdır, çünkü regresyon katsayısı β için Poisson ve NB olabilirlikleri altında önsel eşlenik dağılım ve NB yayılım parametresi Γ için önsel eşlenik dağılım da bilinmemektedir. Bu sorunları ele almak için, Zhou ve ark. (2012) tarafından LGNB modeli olarak adlandırılan, sayımlar için bir lognormal ve gamma mixed NB regresyon modeli önerilmiştir.

Önceden tanımlamış olduğumuz $\ln N(0, \sigma^2)$ log-normal dağılıma çarpımsal rasgele etki terimi ϵ_i ve daha önce bir gamma dağılımı parameteresi Γ eklenir.

$$p_i = \frac{e^{\Psi_i}}{1 + e^{\Psi_i}} = \frac{\exp(x_i' \beta) \epsilon_i}{1 + \exp(x_i' \beta) \epsilon_i} \quad (6.49)$$

$$\text{Logit}(p_i) = \ln \frac{p_i}{1 - p_i} \quad (6.50)$$

alındığı takdirde LGNB modeli

$$y_i \sim \text{NB}(r, p_i), \quad \Psi_i = \text{logit}(p_i) = x_i^T \beta + \ln \epsilon_i$$

$$\epsilon_i \sim \ln N(0, \varphi^{-1}), \quad \varphi \sim \text{gamma}(e_0, 1/f_0)$$

$$\beta \sim \prod_{p=0}^P N(0, \alpha_p^{-1}), \quad \alpha_p \sim \text{gamma}(c_0, 1/d_0)$$

$$r \sim \text{gamma}(a_0, 1/h), \quad h \sim \text{gamma}(b_0, 1/g_0)$$

olarak oluşturulabilir (Zhou ve ark., 2012). Burada $\varphi = \sigma^{-2}$ ve $a_0, b_0, c_0, d_0, e_0, f_0$ ve g_0 gamma hiper parametreleridir ve çalışmalarda 0.01 olarak alınırlar.

$$y_i \sim \text{NB}(r, p_i) \text{ iken } y_i \sim \text{Pois}(\lambda_i) \text{ ve } \lambda_i \sim \text{gamma}(r, \exp(x_i^T \beta) \epsilon_i)$$

alınarak gamma-poisson yapısında ifade edilebilir ve bu durumda LGNB modeli aynı zamanda lognormal-gamma-gamma-poisson regresyon modeli olarak adlandırılabilir. Bu model $X = [x_1^T, x_2^T, \dots, x_N^T]^T$ ve $\Psi = [\Psi_1, \dots, \Psi_N]^T$ alınır

$$\Psi \sim N(X\beta, \varphi^{-1}I)$$

şeklinde ifade edilebilir (Zhou ve ark., 2012).

Model Özellikleri ve Model Karşılaştırması

Toplam beklenen değer, toplam varyans değerlerini ve NB dağılımının momentlerini kullanarak;

$$E\{y_i|x_i\} = E_{\epsilon_i}[E\{y_i|x_i, \epsilon_i\}] = \exp(x_i^T \beta + \frac{\sigma^2}{2} + \ln r) \quad (6.51)$$

$$\begin{aligned} \text{Var}\{y_i|x_i\} &= E_{\epsilon_i}[\text{Var}\{y_i|x_i, \epsilon_i\}] + \text{Var}_{\epsilon_i}[E\{y_i|x_i, \epsilon_i\}] \\ &= E\{y_i|x_i, \epsilon_i\} + (e^{\sigma^2} (1 + r^{-1}) - 1)E^2\{y_i|x_i\} \end{aligned} \quad (6.52)$$

şeklinde yazabiliriz. Quasi yayılım parametresi $\kappa = (e^{\sigma^2} (1 + r^{-1}) - 1)$, varyans terimindeki ortalama karesel terimle ilişkili olan bir katsayı olarak tanımlanır. Negatif binom modelde $\kappa = \infty$ ve lognormal poisson modelde ise $\kappa = e^{\sigma^2} - 1$ 'dir. Yayılım parametreleri açısından farklı dağılım varsayımlarına sahip olmalarına rağmen, uyum iyiliği bakımından açıkça bir modelin diğerine oranla daha üstün olduğu henüz söylenememektedir. Önerilen LGNB modelinde ortalama ve $(e^{\sigma^2} (1 + r^{-1}) - 1)$ değerini ayarlamak amacıyla iki farklı parametre, r ve σ^2 , bulunmaktadır. Bu model $\sigma^2 = 0$ olduğunda negatif binom modeliyle, $\infty = r^{-1} = 0$ olduğunda ise lognormal poisson modeliyle denk olmaktadır. Böylece LGNB modeli her iki rastgele etkiyi dikkate alan ve fazladan bir serbestlik derecesine sahip olan bir modeldir (Zhou ve ark., 2012).

6.5. Negatif Binom Hurdle Regresyon Modeli

Negatif binom hurdle regresyon modeli, Bölüm (5.8)'de değinilen poisson hurdle modeliyle oldukça benzer yapıdadır. Bol sıfırlı modellerde olduğu gibi yine sıfır

değerli ve pozitif değerli yanıt değerlerini tahmin amacıyla oluşturulan iki aşamalı süreç içeriyor olup, aşırı yayılım (over-dispersion) ve eksik yayılım (under-dispersion) durumlarında kullanılabilen bir modeldir.

Poisson hurdle regresyonda olduğu gibi sıfır değerleri için yanıt ikili değişken muamelesi yapılır ve bu amaç için probit veya logit model kullanılır. Pozitif yanıt değerlerini tahmin ederken ise poisson hurdle regresyondan farklı olarak kesilmiş negatif binom dağılımı kullanılır. Uygulama alanı olarak her iki model de aynı amaçlarla kullanılabilir ve sıfır değerlerinden kaynaklanan sorunlara çözüm olabilir. Negatif binom hurdle modelinin log olabilirlik fonksiyonunun yapısı ve genel olarak Hurdle regresyon modellerinin işleyişi hakkında detaylı bilgi Bölüm (5.8)'de verilmiştir. Bu model hakkında daha fazla bilgi için McDowell (2003), Saffari ve ark. (2012), Cameron ve Trivedi (2013) ve Agresti (2015) incelenebilir.

7. UYGULAMA ÇALIŞMASI

Tez çalışması kapsamında değinilen modelleri kullanarak bir hastanede doktorların hastaları ziyaret sayılarının modellenmesi amaçlanmıştır. Bu amaçla klasik poisson, negatif binom ve bu iki modelin sıfır yoğunluklu verileri için önerilen versiyonları olan bol sıfırlı poisson ve bol sıfırlı negatif binom regresyon modellerinin yanı sıra ridge ve Liu tahmin edicilerine dayalı yanlı tahmin versiyonları incelenmiştir. Bu süreçte verinin hazırlanması, düzenlenmesi ve modellenmesi amacıyla R, IBM SPSS ve Microsoft Office Excel programları kullanılmıştır.

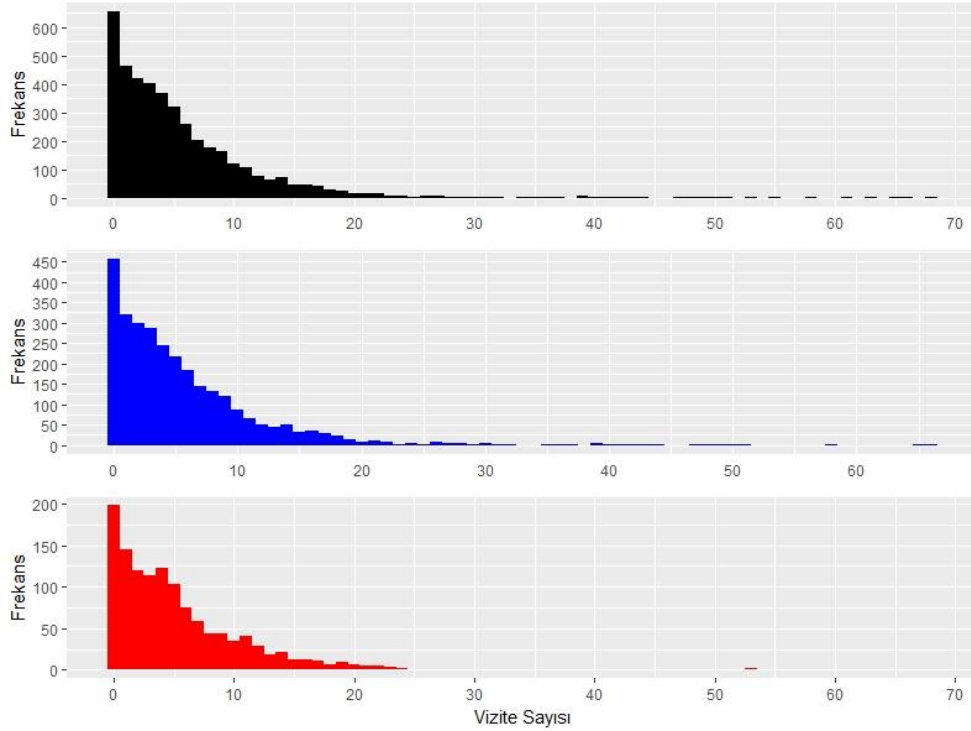
7.1. Verinin Yapısı ve Kökeni

Çalışma kapsamında kullanılan veri Deb ve Trivedi (1997) çalışmasından alınmıştır. Bu veri setine <https://www.jstatsoft.org/article/view/v016i09> adresinden erişilebilir. Söz konusu veri seti hastanede kalınan gün sayısı, bireyin kendi hakkındaki sağlık algısı, kronik rahatsızlık sayısı, günlük yaşamda ki kısıtlılık durumu, yaş, cinsiyet, medeni hal, eğitim durumu, ailenin geliri, iş durumu, özel sağlık sigortasının varlığı gibi açıklayıcı değişkenleri ve yanıt değişkeni olarak doktorların ziyaret (vizite) sayısı değişkeninden oluşmaktadır. Veri setinde toplam 4406 birey vardır.

Regresyon modelleri uydurulmadan önce veri üzerinden birtakım ön işleme süreçleri yürütülmüştür. Veri seti kayıp değer içermemektedir. Tüm bireylere ait ölçümler üzerinden aykırı değerlerin varlığı araştırılmıştır. Her bir açıklayıcı değişken kendi içerisinde değerlendirilip söz konusu değişkenin ortalamasının ± 3 standart sapma aralığının dışında kalan ölçümleri içeren birimler aykırı değer olarak işaretlenmiştir. Sürekli değişkenler üzerinde yürütülen bu işlem sonucunda 224 birimin aykırı değer içerdiği görülmüş olup bu birimler işlem dışı tutulmuştur. Nihai olarak kalan 4182 birim üzerinden modelleme yapılmıştır.

Tez kapsamında değinilen regresyon modellerinin performanslarını daha etkin karşılaştırabilmek amacıyla veri seti %70 ve %30 oranında sırasıyla eğitim ve test verisi olarak ayrılmıştır. Modeller eğitim kümesi üzerinden modellenip, eğitim süresince modellerin karşılaşmadığı yeni verilerle, test verileriyle performansları ölçülmüştür.

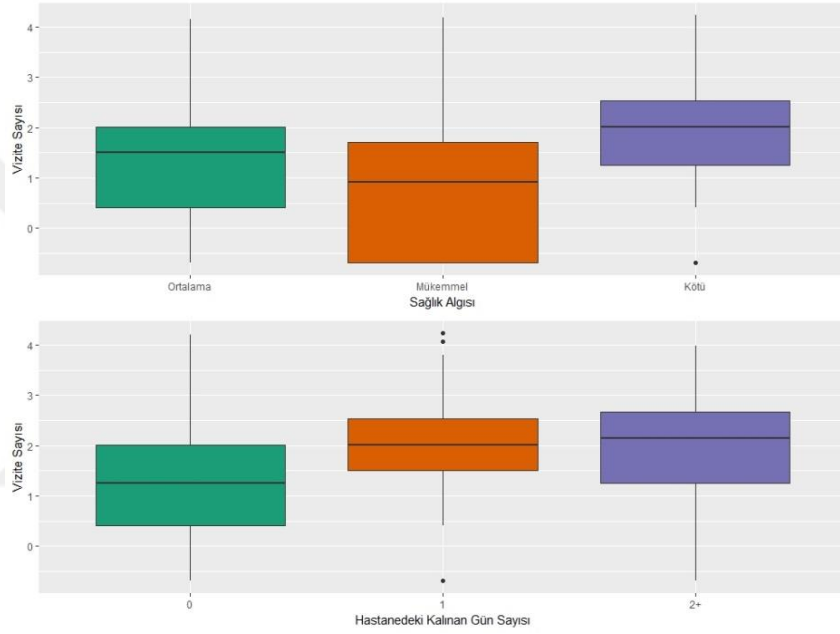
Bir bütün olarak veri setinin yanı sıra eğitim ve test veri setleri için vizite sayılarının frekanslarının dağılımı Şekil 7.1’de verilmiştir. Şekil 7.1’e göre her üç ayrı durumda da sıfır değerinin frekansının ağırlığı görülmektedir.



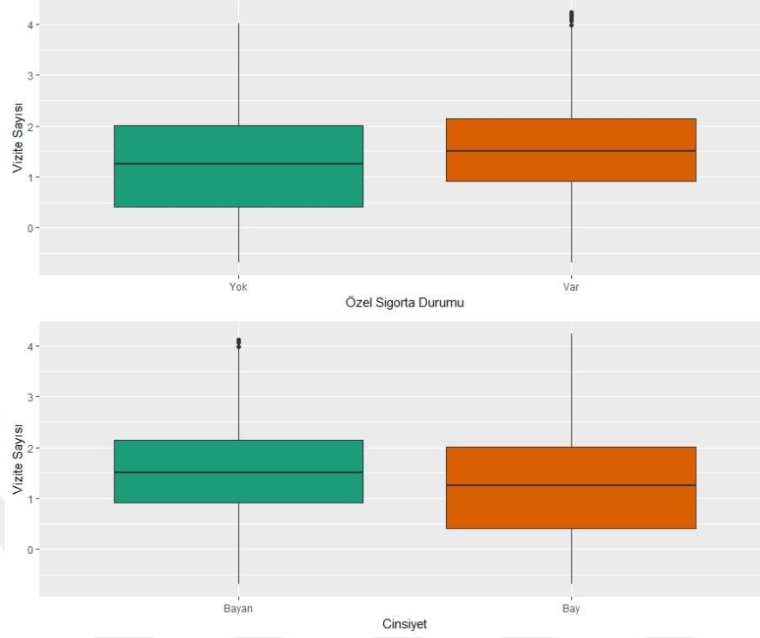
Şekil 7.1. Bütün, eğitim ve test verilerinde vizite sayılarının frekans grafikleri

Veri setinde bireylerin sağlık algılarına ve hastanede kaldıkları gün sayısına göre doktorların vizite sayılarının grafikleri Şekil 7.2’de, özel sigorta durumu ve cinsiyet değişkenlerine bağlı vizite sayılarının grafiği ise Şekil 7.3’de verilmiştir.

Şekil 7.2'ye göre sağlık algısının kötü olduğunu düşünen bireylerde doktorların vizite sayıları artış göstermiştir. Ayrıca hastanede kalınan gün sayısı ile vizite sayısının değeri artış eğilimindedir. Şekil 7.3'e göre bayan hastalarda ve özel sigortası olan bireylerde vizite sayısı, baylara ve özel sigortası olmayan bireylere oranla daha yüksektir.



Şekil 7.2. Bireylerin sağlık algılarına ve hastanede kalınan gün sayılarına göre vizite sayılarının frekans grafikleri



Şekil 7.3. Bireylerin özel sigorta durumları ve cinsiyetlerine göre vizite sayılarının frekans grafikleri

7.2. Eğitim Verisinin Modellenmesi

Ridge ve Liu tahmin edicilerine dayanan poisson ve negatif binom regresyon modellerinin dışında kalan altı model (PR, NBR, ZIPR, ZINBR, PHR ve NBHR) eğitim verisine uygulanmıştır. Eğitim verisine en iyi uyum gösteren modeli belirlemek amacıyla Akaike bilgi kriteri (AIC) ve log olabilirlik değeri (LL) hesaplanmıştır. Elde edilen sonuçlar Çizelge 7.1’de özetlenmiştir. Çizelge 7.1’e göre eğitim verinde en iyi uyum gösteren, diğer bir deyişle en küçük AIC ve LL değerlerini veren modeller NBR, ZINBR ve NBHR modelleri olmuştur. Bu üç modelin performans değerleri birbirine yakın olsa da en küçük AIC ve LL değeri NBHR modelinde gözlenmiştir. Eğitim verisine en zayıf uyum gösteren model ise en yüksek AIC ve LL değerlerine sahip olan PR modeli olmuştur.

Çizelge 7.1. Eğitim modellerinin AIC ve LL değerleri

Model	AIC	LL
Poisson (PR)	23239.538	-11606.769
Negatif binom (NBR)	16091.514	-8031.757
Bol sıfırlı poisson (ZIPR)	20968.239	-10458.119
Bol sıfırlı negatif binom (ZINBR)	16003.561	-7974.781
Poisson hurdle (PHR)	20968.497	-10458.249
Negatif binom hurdle (NBHR)	15993.192	-7969.596

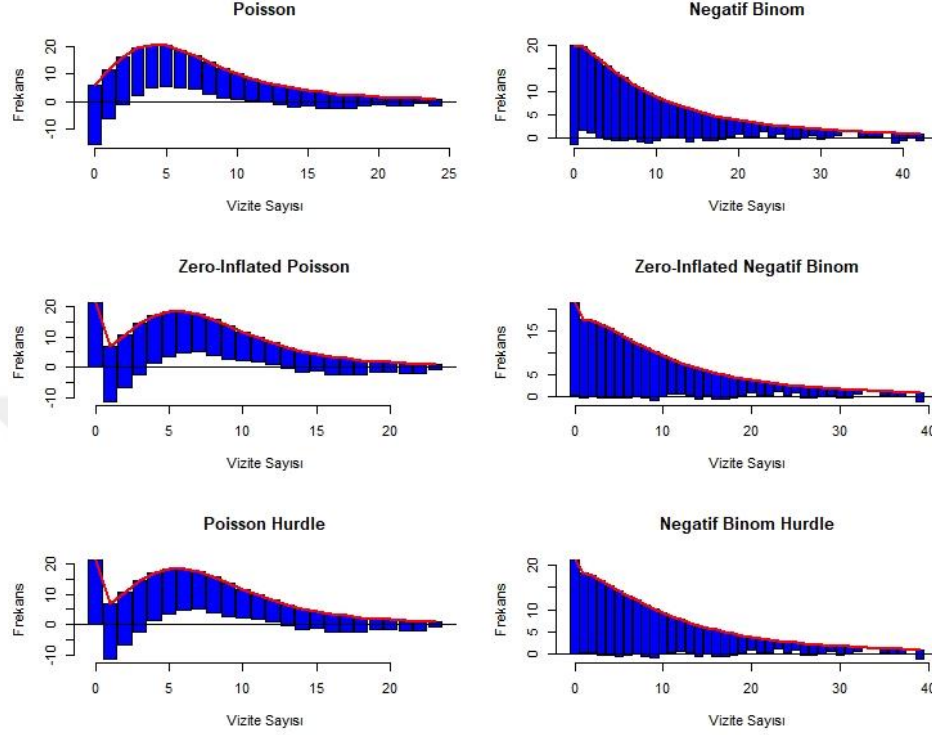
Eğitim verisine uydurulan altı model arasında negatif binom regresyona dayalı modellerde aşırı yayılım parametresinin gerekli olup olmadığını belirlemek amacıyla iç içe modeller arasında log olabilirlik oran testi yürütülmüştür. Bu amaçla temel poisson ve negatif binom regresyon modelleri kendi arasında, bu iki modelin bol sıfırlı ve hurdle versiyonları kendi aralarında log olabilirlik oran testi ile karşılaştırılmıştır. Elde edilen sonuçlar Çizelge 7.2’de verilmiştir. Çizelge 7.2’ye göre üç karşılaştırma da anlamlı bulunmuştur. Buna göre negatif binom regresyona dayalı modeller eğitim verisinde aşırı yayılım durumuyla baş edebilme noktasında poisson modellerine oranla daha uygundur.

İç içe model karşılaştırmalarının dışında olası ikili modeller de Vuong testiyle karşılaştırılmıştır. Bu sayede aşırı sıfır değerinin tahmin edilmesinde hangi modellerin uygun olabileceği belirlenebilmektedir. Vuong testinde pozitif test istatistiği birinci modelin ikinci modele göre daha tercih edilebilir olduğunu, negatif olması durumunda ise ikinci modelin birinci modele tercih edilebileceği yorumu yapılır. Elde edilen test istatistikleri ve anlamlılık sonuçları yine Çizelge 7.2’de özetlenmiştir. Çizelge 7.2’de ki Vuong test sonuçlarına göre PHR modeli PR modeline göre, ZIPR modeli PR modeline göre, ZINBR modeli NBR modeline göre, NBHR modeli NBR modeline göre daha tercih edilebilirdir ve aralarında ki fark istatistiksel olarak anlamlıdır. Sadece ZINBR ile NBHR modelleri ve ZIPR ile PHR modelleri arasında farklılık anlamlı değildir.

Çizelge 7.2. Eğitim modellerinde yürütülen ikili karşılaştırmaların LRT ve Vuong test sonuçları

LRT	NBR – PR		ZINBR – ZIPR		NBHR – PHR	
	7150 (p<0.001)		4966.7 (p<0.001)		4977.3 (p<0.001)	
Vuong (z-ist.)	PR - PHR	ZINBR - NBHR	PR - ZIPR	NBR - ZINBR	ZIPR - PHR	NBR - NBHR
	-14.183 (p<0.001)	-1.037 (p=0.1499)	-14.186 (p<0.001)	-5.008 (p<0.001)	0.745 (p=0.2508)	-5.352 (p<0.001)

Eğitim verisinde model sonuçlarını görsel olarak karşılaştırmak için rootogram grafikleri kullanılmıştır. Rootogram grafiğinde x ekseninde bulunan vizite sayısının frekansına karşılık gelen barların eksene yaklaşması uyumun iyileştiği şeklinde yorumlanır. Kullanılan altı modele ait rootogram grafikleri Şekil 7.4’te verilmiştir. Şekil 7.4 incelendiğinde poisson temelli modellerin (poisson, bol sıfırlı poisson, poisson hurdle), negatif binom temelli modellere (negatif binom, bol sıfırlı negatif binom, negatif binom hurdle) oranla çok daha zayıf uyum sergilediği söylenebilir. Bol sıfırlı negatif binom ve negatif binom hurdle regresyon modelleri negatif binom regresyona oranla biraz daha iyi uyum göstermiştir. Aşırı sıfır durumunda ve aşırı yayılım durumunda basit modellerden negatif binom modeli tercih edilebilir fakat daha etkin bir tahmin için bol sıfırlı versiyonları kullanılmalıdır.



Şekil 7.4. Eğitim verisinde kullanılan modellerin rootogram grafikleri

7.3. Test Verisi Kullanılarak Performans Karşılaştırması

Eğitim verisinde elde edilen sonuçları doğrulamanın yanı sıra ridge ve Liu tahmin edicilerine dayalı poisson ve negatif binom regresyon modelleri eğitim sırasında kullanılmayan test verisinde kullanılmıştır. Eğitim verisinde kullanılan altı modelin yanı sıra poisson ridge (PRR), poisson Liu (PLR), Negatif binom ridge (NBRR) ve Negatif binom Liu (NBLR) modelleri de değerlendirmeye alınmıştır. Söz konusu on modelin test verisindeki performansını karşılaştırmak amacıyla rezidü kareler ortalaması (MSE), rezidü kareler ortalamasının karekökü (RMSE) ve rezidülerin ortalama mutlak hata değerleri (MAE) kullanılmıştır. Rezidü olarak gerçek frekans değeri ile modelin tahmin değeri arasındaki fark olarak tanımlanan standartlaştırılmamış rezidü değeri kullanılmıştır.

Test verisinde elde edilen performans sonuçları Çizelge 7.3'te verilmiştir. Çizelge 7.3'e göre test performansı en yüksek olan modeller en düşük MSE, RMSE ve MAE değerlerini sağlayan ZIPR, ZINBR ve PHR modelleri olmuştur. Aralarındaki fark az da olsa en zayıf performanslar ridge ve Liu temelli poisson ve negatif binom regresyon modellerinde elde edilmiştir. Genel olarak en başarılı model çok az farkla PHR modeli olurken, en zayıf model NBLR modeli olmuştur.

Çizelge 7.3. Test verisinde modellerin performans sonuçları

Model	MSE	RMSE	MAE
Poisson (PR)	39.8572	6.31325	3.9699
Negatif binom (NBR)	40.7568	6.38410	3.9847
Bol sıfırlı poisson (ZIPR)	39.6489	6.29673	3.9449
Bol sıfırlı negatif binom (ZINBR)	39.8083	6.30938	3.9318
Poisson hurdle (PHR)	39.6447	6.29640	3.9450
Negatif binom hurdle (NBHR)	40.0699	6.33007	3.9500
Poisson ridge (PRR)	40.8191	6.38898	3.9842
Poisson Liu (PLR)	40.8198	6.38903	3.9844
Negatif binom ridge (NBRR)	41.0495	6.40698	3.9919
Negatif binom Liu (NBLR)	41.0499	6.40701	3.9920

7.4. En İyi Modelin Belirlenmesi ve Yorumlanması

Eğitim verisinde elde edilen sonuçlar, rootogram grafikleri ve test verisinde elde edilen sonuçlar birlikte değerlendirildiğinde tez kapsamında kullanılan veri seti için en uygun modelin bol sıfırlı negatif binom regresyon modeli (ZINBR) olduğu söylenebilir. Eğitim verisinde NBR, ZINBR ve NBHR en iyi performansı ortaya koymuş ve birbirine çok yakın AIC ve LL değerleri elde edilmişti. Rootogram grafiklerinde de negatif binom temelli regresyon modelleri (NBR, ZINBR, NBHR) görsel olarak başarılı bulunmuştu. Tüm modeller göz önüne alındığında test verisinde en başarılı modeller ise ZIPR, ZINBR ve PHR modelleri bulunmuştur. Ayrıca ZIPR ile ZINBR arasında log olabilirlik oran testi anlamlı

bulunmuştur. Tüm sonuçlar göz önüne alındığında her üç durumda da başarılı performans sergileyen ZINBR modeli en ideal model olarak öne çıkmaktadır.

En iyi modelin ZINBR olduğunun belirlenmesi sonrası tüm veri seti bu model ile modellenmiştir. Elde edilen modelin katsayıları ve standart hataları Çizelge 7.4'te verilmiştir.

Çizelge 7.4. Bol sıfırlı negatif binom regresyon (ZINBR) uyumu sonuçları

Değişken	Count Part			Zero Part		
	β	$\exp(\beta)$	SH	β	$\exp(\beta)$	SH
(Intercept)	1.593*	4.920	0.2158	4.574*	96.93	1.4508
			0		5	2
Kalınan gün sayısı	0.280*	1.324	0.0299	-0.533	0.587	0.3417
			5			1
Sağlık algısı (Mükemmel)	-	0.759	0.0644	0.406	1.500	0.2806
	0.276*		5			1
Sağlık algısı (Kötü)	0.264*	1.303	0.0494	-0.505	0.603	0.5874
			9			6
Kronik rahatsızlık sayısı	0.137*	1.147	0.0130	-	0.324	0.1690
			8	1.127*		8
Günlük yaşam kısıtlılığı (Evet)	0.117*	1.124	0.0418	0.039	1.040	0.3431
			0			0
Yaş (/10yıl)	-	0.945	0.0273	-	0.549	0.1899
	0.057*		0	0.600*		9
Cinsiyet (Bay)	-0.026	0.974	0.0355	1.044*	2.840	0.2417
			2			6
Medeni hal (Evli)	-	0.895	0.0369	-	0.474	0.2582
	0.111*		9	0.747*		3
Eğitim durumu	0.020*	1.020	0.0047	-	0.908	0.0304
			1	0.097*		1
Aile geliri (x10.000\$)	0.007	1.007	0.0088	-0.022	0.978	0.0675

			6			2
İş durumu (Çalışan)	-0.022	0.978	0.0531	-0.334	0.716	0.3729
			4			8
Özel sigorta durumu (Var)	0.137*	1.147	0.0429	-	0.310	0.2291
			2	1.172*		8

* 0.05 düzeyinde anlamlı bulunmuştur.

Çizelge 7.4'te verilen sonuçlara göre ZINBR modelinin sıfırdan farklı frekansları tahmin etmek için oluşturulan katsayı değerlerine (count part) göre doktorların vizite sayıları üzerinde hastanede kalınan gün sayısı, sağlık algısı (mükemmel ve kötü gruplarının norma gruplara oranla), kronik rahatsızlık sayısı, günlük yaşamda kısıtlılık durumu, yaş, medeni hal, eğitim durumu ve özel sigorta durumu anlamlı etkiye sahip olan değişkenler olarak bulunmuştur. Sadece cinsiyet ve aile gelirinin yanıt değişken üzerinde anlamlı olmadığı söylenebilir. Günlük yaşamda kısıtlılık yaşayan bireylerde doktorların viziteye (en az bir veya daha fazla sayıda) gitme durumu, yaşamayanlara oranla yaklaşık %15 daha olasıdır. Benzer şekilde sağlık algısı kötü hissedenlerde doktorun viziteye (en az bir veya daha fazla sayıda) gitme durumu diğer gruplara oranla %30 daha olasıdır. Doktorun viziteye gitme (en az bir veya daha fazla sayıda) durumu evli bireylerde evli olmayanlara göre yaklaşık %11 daha az, özel sigortası olanlarda olmayanlara göre yaklaşık %14 daha fazla olasıdır.

Diğer taraftan ZINBR modelinin sıfır değerli frekansları tahmin etmek için oluşturulan katsayı değerlerine (zero part) göre kronik rahatsızlık sayısının, yaşın, medeni halin, eğitim durumunun, cinsiyet ve özel sigorta durumunun anlamlı etkiye sahip olan değişkenler olarak elde edilmişken kalan değişkenlerin anlamlı olmadığı görülmüştür. Sağlık algısını mükemmel olarak hisseden bireylerde doktorun hiç viziteye gelmeme (sıfır frekans) durumu diğer gruplara oranla %50 daha olasıdır. Doktorların evli bireylerde hiç viziteye gitmeme (sıfır frekans) durumu bekârlara oranla yaklaşık %53 daha az olasıdır. Bay hastalarda doktorların

viziteye hiç gitmeme durumu (sıfır frekans) bayanlarakıyasla%184 daha olasıdır. Doktorun en az bir veya daha fazla kez viziteye gitme durumunun daha olası olması, sıfır değerli frekansların görülmesinin yani doktorun viziteye hiç gitmeme durumunun daha az olası olacağı çıkarımı yapılabilir. Bu bağlamda ele alındığında sıfır frekanslarını tahmin için elde edilen sonuçların, sıfırdan farklı frekanslar için bulunan sonuçların birbirini desteklediği, tutarlı ve uyumlu oldukları görülmektedir.



8. SONUÇ VE ÖNERİLER

Bu tez çalışmasında, poisson ve negatif binom regresyona dayanan on farklı regresyon modelinin performansı gerçek bir veri seti kullanılarak karşılaştırılmıştır. Kullanılan veri setinde doktorların hastalarını ziyaret sayıları, diğer bir deyişle vizite sayılarının tahmin edilmesi amaçlanmıştır. Bu kapsamda poisson, negatif binom, bol sıfırlı poisson, bol sıfırlı negatif binom, poisson hurdle, negatif binom hurdle, poisson ridge, negatif binom ridge, poisson Liu ve negatif binom Liu regresyon modelleri incelenmiştir. Yapılan analizler sonucunda bol sıfırlı negatif binom regresyon modelinin veriye en uygun model olduğu belirlenmiş olup, bu model için katsayı tahminleri yapılmış ve yorumlanmıştır. Sonuç olarak sayım verileri, özellikle de bol sıfırlı sayım verilerinde tez kapsamında değinilen modellerin başarılı birer tahmin modeli olabileceği görülmüştür.

Çalışma kapsamında değinilen modellerin farklı disiplinlerde daha farklı uygulama problemlerinde kullanılabileceği düşünülmektedir. Poisson ve negatif binom regresyon modellerinin ridge ve Liu gibi yanlı tahmin edicilerinin dışında alternatif istatistiksel tahmin edicilerle daha fazla geliştirilmesi ve farklı disiplinlerde özellikle makine öğrenmesi gibi son dönemlerde ilgi gören alanlara başarılı bir şekilde dâhil edilmesi gelecekte odaklanılabilecek araştırma konuları olarak öngörülmektedir.



KAYNAKLAR

- Agresti, A. (2015). Foundations of linear and generalized linear models. John Wiley & Sons, 472p.
- Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. 2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR, September 2-8, 1971, Budapest: Akadémiai Kiadó, pp. 267–281.
- Alkhamisi, M. A., and Shukur, G., 2008. Developing ridge parameters for SUR model. Communications in Statistics-Theory and Methods, 37(4), 544-564.
- Alkhamisi, M., Khalaf, G., and Shukur, G., 2006. Some modifications for choosing ridge parameters. Communications in Statistics-Theory and Methods, 35(11), 2005-2020.
- Andersson, B., 2001. Scandinavian evidence on growth and age structure. Regional Studies, 35(5), 377-390.
- 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.
- Avşar, V., 2014. Türkiye'nin antidamping soruşturmalarını etkileyen faktörler: sanayi verileri ile ekonometrik bir analiz. Eskişehir Osmangazi Üniversitesi İktisadi ve İdari Bilimler Dergisi, 9(1), 41-54.
- Bandyopadhyay, D., DeSantis, S. M., Korte, J. E., and Brady, K. T., 2011. Some considerations for excess zeroes in substance abuse research. The American Journal of Drug and Alcohol Abuse, 37(5), 376-382.
- Batah, F. S. M., Ramanathan, T.V., and Gore, S.D., 2008. The efficiency of modified Jackknife and ridge type regression estimators: A comparison. Surv. Math., pp. 111–122.
- Breslow, N. E., 1984. Extra-Poisson variation in log-linear models. Journal of the Royal Statistical Society: Series C (Applied Statistics), 33(1), 38-44.

- Buu, A., Johnson, N. J., Li, R., and Tan, X., 2011. New variable selection methods for zero-inflated count data with applications to the substance abuse field. *Statistics in Medicine*, 30(18), 2326-2340.
- Cameron, A. C., and Trivedi, P. K., 2013. *Regression analysis of count data* (Vol. 53). Cambridge university press, 598p.
- Chai, T., Xiong, D. Q., and Weng, J., 2018. A zero-inflated negative binomial regression model to evaluate ship sinking accident mortalities. *Transportation Research Record*, 2672(11), 65-72.
- Chatterjee, S., Hadi, A., and Price, B., 2006. *Regression analysis by example*. John Wiley & Sons, New York, 424p.
- Consul, P., & Famoye, F., 1992. Generalized Poisson regression model. *Communications in Statistics-Theory and Methods*, 21(1), 89-109.
- Cragg, J. G., 1971. Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica* (pre-1986), 39(5), 829.
- Deb, P., and Trivedi, P. K., 1997. Demand for medical care by the elderly: a finite mixture approach. *Journal of Applied Econometrics*, 12(3), 313-336.
- Deniz, Ö., 2005. Poisson regresyon analizi. *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*, 4(7), 59-72.
- Dobson, A. J., and Barnett, A. G., 2008. *An introduction to generalized linear models*. Chapman and Hall/CRC, 316p.
- Escario, J. J., and Wilkinson, A. V., 2015. The intergenerational transmission of smoking across three cohabitant generations: A count data approach. *Journal of Community Health*, 40(5), 912-919.
- Fallah, N., Gu, H., Mohammad, K., Seyyedsalehi, S. A., Nourijelyani, K., and Eshraghian, M. R., 2009. Nonlinear Poisson regression using neural networks: a simulation study. *Neural Computing and Applications*, 18(8), 939.

- 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(1), 117-130.
- Faroughi, P., and Ismail, N., 2017. Bivariate zero-inflated negative binomial regression model with applications. *Journal of Statistical Computation and Simulation*, 87(3), 457-477.
- Ghitany, M. E., Karlis, D., Al-Mutairi, D. K., and Al-Awadhi, F. A., 2012. An EM algorithm for multivariate mixed poisson regression models and its application. *Applied Mathematical Sciences*, 6(137), 6843-6856.
- Heinzl, H., and Mittlböck, M., 2003. Pseudo R-squared measures for poisson regression models with over-or underdispersion. *Computational Statistics & Data Analysis*, 44(1-2), 253-271.
- Hinkley, D. V., 1977. Jackknifing in unbalanced situations. *Technometrics*, 19(3), 285-292.
- Hoerl, A. E., and Kennard, R. W., 1970a. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55-67.
- Hoerl, A. E., and Kennard, R. W., 1970b. Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12(1), 69-82.
- Hoerl, A. E., Kennard, R. W., and Baldwin, K. F., 1975. Ridge regression: some simulations. *Communications in Statistics-Theory and Methods*, 4(2), 105-123.
- Huang, J., and Yang, H., 2014. A two-parameter estimator in the negative binomial regression model. *Journal of Statistical Computation and Simulation*, 84(1), 124-134.
- Kibria, B. G., 2003. Performance of some new ridge regression estimators. *Communications in Statistics-Simulation and Computation*, 32(2), 419-435.

- Kibria, B. G., Månsson, K., and Shukur, G., 2011. A ridge regression estimator for the zero-inflated poisson model. Royal Institute of Technology, CESIS-Centre of Excellence for Science and Innovation Studies, Stockholm.
- Kibria, B. G., Månsson, K., and Shukur, G., 2012. Performance of some logistic ridge regression estimators. *Computational economics*, 40(4), 401-414.
- Kibria, B. G., Månsson, K., and Shukur, G., 2013. Some ridge regression estimators for the zero-inflated poisson model. *Journal of Applied Statistics*, 40(4), 721-735.
- Köleoğlu, N., 2006. Olay zamanı analizinde tesadüfi etkiler poisson regresyon modeli ile gözlemlenemeyen heterojenliğin incelenmesi, Marmara Üniversitesi, Sosyal Bilimler Enstitüsü (Doktora Tezi).
- Liu, K., 1993. A new class of biased estimate in linear regression. *Communications in Statistics-Theory and Methods*, 22(2), 393-402.
- Liyanage, T., Ninomiya, T., Jha, V., Neal, B., Patrice, H. M., Okpechi, I., and Rodgers, A., 2015. Worldwide access to treatment for end-stage kidney disease: a systematic review. *The Lancet*, 385(9981), 1975-1982.
- Månsson, K., 2012. On ridge estimators for the negative binomial regression model. *Economic Modelling*, 29(2), 178-184.
- Månsson, K., 2013. Developing a Liu estimator for the negative binomial regression model: method and application. *Journal of Statistical Computation and Simulation*, 83(9), 1773-1780.
- Månsson, K., and Shukur, G., 2011. A poisson ridge regression estimator. *Economic Modelling*, 28(4), 1475-1481.
- Månsson, K., Kibria, B. G., and Shukur, G., 2012. On Liu estimators for the logit regression model. *Economic Modelling*, 29(4), 1483-1488.
- Månsson, K., Shukur, G., and Kibria, B. G., 2010a. A simulation study of some ridge regression estimators under different distributional assumptions. *Communications in Statistics-Simulation and Computation*, 39(8), 1639-1670.

- Mansson, K., Shukur, G., & Kibria, B. G., 2010b. On some ridge regression estimators: A Monte Carlo simulation study under different error variances. *Journal of Statistics*, 17(1), 1-22.
- Mansson, K., Kibria, B. G., Sjolander, P., and Shukur, G., 2012. Improved Liu estimators for the Poisson regression model. *International Journal of Statistics and Probability*, 1(1), 2.
- Marquardt, D. W., and Snee, R. D., 1975. Ridge regression in practice. *The American Statistician*, 29(1), 3-20.
- Martinez, F. J., Calverley, P. M., Goehring, U. M., Brose, M., Fabbri, L. M., and Rabe, K. F., 2015. Effect of roflumilast on exacerbations in patients with severe chronic obstructive pulmonary disease uncontrolled by combination therapy (REACT): a multicentre randomised controlled trial. *The Lancet*, 385(9971), 857-866.
- McDonald, G. C., and Galarneau, D. I., 1975. A Monte Carlo evaluation of some ridge-type estimators. *Journal of the American Statistical Association*, 70(350), 407-416.
- McDowell, A., 2003. From the help desk: hurdle models. *The Stata Journal*, 3(2), 178-184.
- Montgomery, D. C., Peck, E. A., and Vining, G. G., 2012. *Introduction to linear regression analysis* (Vol. 821). John Wiley & Sons, 672p.
- Mouatassim, Y., and Ezzahid, E. H., 2012. Poisson regression and zero-inflated poisson regression: application to private health insurance data. *European Actuarial Journal*, 2(2), 187-204.
- Mullahy, J., 1986. Specification and testing of some modified count data models. *Journal of econometrics*, 33(3), 341-365.
- Muniz, G., and Kibria, B. G., 2009. On some ridge regression estimators: An empirical comparisons. *Communications in Statistics—Simulation and Computation*, 38(3), 621-630.

- Nelder, J. A., and Mead, R., 1965. A simplex method for function minimization. *The Computer Journal*, 7(4), 308-313.
- Oliveira, M., Einbeck, J., Higuera, M., Ainsbury, E., Puig, P., and Rothkamm, K., 2016. Zero-inflated regression models for radiation-induced chromosome aberration data: A comparative study. *Biometrical Journal*, 58(2), 259-279.
- Saffari, S. E., Adnan, R., and Greene, W., 2012. Hurdle negative binomial regression model with right censored count data. *SORT: statistics and operations research transactions*, 36(2), 0181-194.
- Schaefer, R. L., 1986. Alternative estimators in logistic regression when the data are collinear. *Journal of Statistical Computation and Simulation*, 25(1-2), 75-91.
- Schaefer, R. L., Roi, L. D., and Wolfe, R. A., 1984. A ridge logistic estimator. *Communications in Statistics-Theory and Methods*, 13(1), 99-113.
- Selim, S., ve Üçdoğruk, Ş., 2003. Sayma veri modelleri ile çocuk sayısı belirleyicileri: Türkiye'deki seçilmiş iller için sosyoekonomik analizler. *Dokuz Eylül Üniversitesi İktisadi İdari Bilimler Fakültesi Dergisi*, 18(2), 13-31.
- 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.
- Singh, B., Chaubey, Y. P., and Dwivedi, T. D., 1986. An almost unbiased ridge estimator. *Sankhyā: The Indian Journal of Statistics, Series B*, 342-346.
- Tang, Y., Liu, W., and Xu, A., 2017. Statistical inference for zero-and-one-inflated poisson models. *Statistical Theory and Related Fields*, 1(2), 216-226.
- Türkan, S., and Özel, G., 2016. A new modified Jackknifed estimator for the poisson regression model. *Journal of Applied Statistics*, 43(10), 1892-1905.
- Wang W, and Famoye F., 1997. Modeling household fertility decisions with generalized poisson regression. *J Popul Econ*, 10: 273-283.

- Winkelmann, R., 2008. Econometric analysis of count data. Springer Science & Business Media, 320p.
- Xie, H., Tao, J., McHugo, G. J., and Drake, R. E., 2013. Comparing statistical methods for analyzing skewed longitudinal count data with many zeros: An example of smoking cessation. *Journal of Substance Abuse Treatment*, 45(1), 99-108.
- Yeşilova, A., ve Atlıhan, R., 2007. Farklı sıcaklıkların *Scymnus Subvillosus*'un bıraktığı yumurta sayıları üzerine etkilerinin karışımı poisson regresyon ile analiz edilmesi. *Yüzüncü Yıl Üniversitesi Tarım Bilimleri Dergisi*, 17(2), 73-79.
- Yeşilova, A., Özgökçe, M., S., Atlıhan R., Karaca, İ., Özgökçe, F., Yıldız, Ş., Kaya, Y., 2011. Sıfır değer ağırlıklı genelleştirilmiş poisson regresyonu yardımıyla Van Gölü'nde *Notonecta Viridis* Delcourt, 1909 (Hemiptera: Notonectidae)'in popülasyon değişimi üzerinde fiziko-kimyasal çevresel koşulların etkilerinin araştırılması. *Türk Entomo. Dergisi*, 35(2), 325-338.
- Yeşilyurt, H., 2005. Poisson regresyon modeli ve Türkiye'deki boşanma istatistiklerine uygulanması, Yüksek lisans tezi, Yıldız Teknik Üniversitesi.
- Zhou, M., Li, L., Dunson, D., and Carin, L., 2012. Lognormal and gamma mixed negative binomial regression. In *Proceedings of the International Conference on Machine Learning*. International Conference on Machine Learning (Vol. 2012, p. 1343). NIH Public Access.
- en.wikipedia.org/wiki/Regression_analysis (Erişim tarihi: 24.08.2019)
- en.wikipedia.org/wiki/Poisson_regression (Erişim tarihi: 24.08.2019)
- en.wikipedia.org/wiki/Negative_binomial_distribution (Erişim tarihi: 24.08.2019)
- en.wikipedia.org/wiki/Zero-inflated_model (Erişim tarihi: 24.08.2019)



ÖZGEÇMİŞ

Gizem Yıldırım 1988’de Adana’da doğdu. İlk ve ortaöğretimini Adana’da tamamladı. Ankara Üniversitesi Fen Fakültesi İstatistik Bölümü’nde başladığı lisans eğitiminden 2011 yılında mezun oldu. 2013 yılında Sosyal Güvenlik Kurumu’nda memur olarak atandı. İki yıl memuriyet hayatının akabinde aynı kurumda Sosyal Güvenlik Denetmen Yardımcısı olarak çalışmaya başladı. Hala aynı kurumda çalışma hayatına devam etmektedir. Evli ve bir çocuk annesidir.

