T.C.
MİMAR SİNAN GÜZEL SANATLAR ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

ÇOKLU BAĞLANTI DURUMUNDA SIRALI LOJİSTİK REGRESYON MODELLERİNDE YÖNTEMLERİN KARŞILAŞTIRILMASI

DOKTORA TEZİ

Onur BAYRAM

İstatistik Anabilim Dalı

İstatistik Programı

Tez Danışmanı: Prof. Dr. Eylem DENİZ

ARALIK 2022

BEYAN

Mimar Sinan Güzel Sanatlar Üniversitesi Fen Bilimleri Enstitüsü tez yazım klavuzuna uygun olarak hazırladığım bu tez çalışmasında;

- tez içindeki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel etik kurallarına uygun olarak sunduğumu,
- başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel normlara uygun olarak atıfta bulunduğumu,
- atıfta bulunduğum eserlerin tümünü kaynak olarak gösterdiğimi,
- kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- ücret karşılığı başka kişilere yazdırmadığımı (dikte etme dışında), uygulamalarımı yaptırmadığımı,
- ve bu tezin herhangi bir bölümünü bu üniversite veya başka bir üniversitede başka bir tez çalışması olarak sunmadığımı beyan ederim.

Onur BAYRAM

ÖNSÖZ

Bu çalışmanın yürütülmesi sırasında bilgi ve deneyimleri ile yanımda olan, görüşlerime güvenip beni onurlandıran danışmanım Sayın Prof. Dr. Eylem Deniz'e ve tez yazım sürecim ve sonrasında yardımlarını, olumlu yorumlarını esirgemeyen Sayın Doç. Dr. Ayça Çakmak Pehlivanlı'ya, Sayın Dr. Öğr. Üy. Alev Bakır Kayı'ya, Sayın Prof. Dr. Gülay Başarır'a, Sayın Doç. Dr. Atıf A. Evren'e;

İstatistik alanında uzmanlaşmamı ve bu seviyeye ulaşmamı sağlayan Mimar Sinan Güzel Sanatlar Üniversitesi İstatistik Anabilim Dalı öğretim üyesi tüm hocalarıma;

Bana her zaman inanan sevgili annem ve babam ile destekleri, iyi niyet ve enerjileriyle sürekli güç veren eşim ve kardeşime;

Bu tez çalışmasında emeği geçen herkese teşekkür ederim.

İÇİNDEKİLER

Sayfa

ÇİZELGE LİSTESİ

ŞEKİL LİSTESİ

ÇOKLU BAĞLANTI DURUMUNDA SIRALI LOJİSTİK REGRESYON MODELLERİNDE YÖNTEMLERİN KARŞILAŞTIRILMASI

(Doktora Tezi)

ÖZET

Veri bilimi kapsamında son yıllarda oldukça popüler olan makine öğrenmesi, yapay zekâ, derin öğrenme vb. alanlarda oldukça sık karşılaşılan sınıflandırma problemlerine, farklı sorunların dikkate alındığı simülasyon çalışmaları ile doğru sınıflandırma performanslarının düzeltilmesine yönelik araştırmalarla çözüm bulunmaya çalışılmıştır. İstatistik biliminde sınıflandırmada çok tercih edilen yöntemlerden biri olan lojistik regresyon, bağımlı değişkenin iki ya da daha çok düzeyde kategorik olduğu durumlarda kullanılan bir yöntem olup kategorilerin sıralı olduğu durumda sıralı lojistik regresyon adını almaktadır. Sıralı lojistik regresyonda, doğrusal regresyon modelinde olduğu gibi bağımsız değişkenler arasında korelasyonların yüksek olması çoklu bağlantı sorununu ortaya çıkarmaktadır. Bu çalışmada, sıralı lojistik regresyon modelindeki bağımsız değişkenler arasında çoklu bağlantılı olması durumunda klasik yöntem ile alternatif yöntemlerin doğru sınıflandırma performanslarının nasıl değişiklik gösterdiği incelenmiştir. Klasik sıralı lojistik tekniği ve alternatif yöntemlerden sıralı lojistik ridge, sıralı lojistik lasso ve sıralı lojistik elastik-net regresyon teknikleri değişken sayıları, örneklem büyüklüğü, dengeli/dengesiz kategori dağılımı, çoklu bağlantının gücü ve farklı bağlantı fonksiyonlarına göre karşılaştırılmıştır. Bu kapsamda simülasyon çalışması ile üretilmiş veri setleri aracılığıyla belirlenen tüm durumları araştırmak ve karşılaştırmak amacıyla 864 tane farklı deneme elde edilmiş, sıralı lojistik regresyon ve düzenlileştirme yöntemleri uygulanarak analiz sonuçlarına ulaşılmıştır. Ayrıca bu çalışmada ilgili yöntemler gerçek bir veri setiyle yapılan uygulamayla değerlendirilmiş ve yorumlanmıştır. Alternatif sıralı lojistik regresyon yöntemlerinin sınıflandırma çalışmalarında doğru ve daha üstün yöntemler olarak kullanılabileceği ortaya konmaya çalışılmıştır.

Anahtar Kelimeler: Sıralı Lojistik Regresyon Modeli, Doğru Sınıflandırma Oranı, Simülasyon, Çoklu Bağlantı

COMPARISON OF
ORDINAL LOGISTIC REGRESSION MODELS
IN MULTICOLLINEARITY SITUATION

(Ph.D. Thesis)

SUMMARY

Within the scope of data science, machine learning, artificial intelligence, deep learning, etc. which have been very popular in recent years and are quite common in fields, it has been tried to find solutions to the classification problems with simulation studies in which different problems are taken into account, and researches to improve the correct classification performances. Logistic regression which is one of the most preferred methods for classification in statistics, is a method used when the dependent variable is categorical at two or more than two levels, and it is called ordinal logistic regression when the categories are ordered. In ordinal logistic regression, as in linear regression model, high correlations between the independent variables reveals the problem of multicollinearity. In this research, it was analyzed how correct classification rate performances of the classical method and the alternative methods differ in case of multicollinearity between the independent variables in ordinal logistic regression model. Classical method ordinal logistic regression and the alternative methods such as ordinal logistic ridge, ordinal logistic lasso and ordinal logistic elastic-net regression were compared according to number of variables, sample size, balanced/unbalanced category distribution, strength of multicollinearity and different link functions. In this study, 864 different trials were obtained in order to investigate and compare all the situations determined through data sets produced by the simulation study, and the analysis results were obtained by applying ordinal logistic regression and regularization methods. In addition, in this study, the relevant methods were evaluated and interpreted with an application conducted with a real-life data set. It has been tried to demonstrate that alternative ordinal logistic regression methods can be used as accurate and superior methods in classification studies.

Keywords: Ordinal Logistic Regression, Correct Classification Rate, Simulation, Multicollinearity

1. GİRİŞ

Regresyon analizi tüm uygulamalı bilimlerde çok önemli bir yere sahip olan, iki yada ikiden daha çok değişken arasındaki ilişkinin belirlenmesi, değişkenler arasındaki neden-sonuç ilişkisinin araştırılması için kullanılan bir istatistiksel analiz yöntemidir. Bağımlı ve bağımsız değişkenler arasındaki ilişkiyi ortaya çıkaran bir matematiksel model sunan regresyon analizinde, bağımlı/bağımsız değişkenlerin sayısı, türü ve ölçek tipine göre farklı tahmin yöntemleri kullanılmaktadır.

Regresyon analizinde bağımlı değişkenin kategorik olması durumu ile oldukça sık karşılaşılmaktadır. Bağımlı değişken iki ya da ikiden daha fazla kategorili niteliksel veri tipinde olduğunda, bağımsız değişkenlerle sebep-sonuç ilişkisini incelemek için lojistik regresyon analiz teknikleri geliştirilmiştir. Doğrusal regresyon modelinde bağımlı değişkenin değeri için tahmin yapılırken, lojistik regresyon analizinde bağımlı değişkenin alacağı değerlerden birinin gerçekleşme olasılığı için tahmin yapılmaktadır (Bircan, 2004).

Lojistik regresyon analizi bağımlı değişkenin düzey sayısının iki ya da ikiden daha çok olması durumuna göre farklılaşmaktadır (Chen ve Hughes, 2004). Kategorik bağımlı değişkenin düzey sayısının ikiden fazla ve sıralı olması durumunda ise sıralı (ordinal) lojistik regresyon modelinin kullanılması söz konusu olmaktadır.

Çizelge 1.1 incelendiğinde bağımlı değişkenin düzey sayısı, bağımsız değişken sayısı ve bağımsız değişkenlerin veri tipine bağlı olarak lojistik regresyon yöntemleri yer almaktadır.

Çizelge 1.1. Lojistik Regresyon Yöntemlerini Gösteren Şema

Bağımlı Değişken Kategori Sayısı	Bağımsız Değişken Sayısı	Bağımsız Değişken Tipi	Regresyon Yöntemi
İki	**Tek**	**Kesikli ve/veya Sürekli**	**İkili (Binominal) Lojistik Regresyon**
İki	**Çok**	**Kesikli ve/veya Sürekli**	**Çok Değişkenli İkili Lojistik Regresyon**
İkiden Çok (Sırasız)	**Tek/Çok**	**Kesikli ve/veya Sürekli**	**Çok Düzeyli (Multinominal) Lojistik Regresyon**
İkiden Çok (Sıralı)	**Tek/Çok**	**Kesikli ve/veya Sürekli**	**Sıralı (Ordinal) Lojistik Regresyon**

Çoklu bağlantı, lojistik regresyon da dahil olmak üzere, model tahminlerinde genel bir problemdir. Lojistik regresyon bağımsız değişkenler arasında çok az çoklu bağlantı veya hiç olmamasını gerektirmektedir. Bu çalışmada sıralı lojistik regresyonda çoklu bağlantı olması durumunda geliştirilmiş olan farklı tahmin yöntemleri karşılaştırılmıştır.

Bu yöntemlerden ilki Hoerl ve Kennard (1970)'ın geliştirdiği ridge regresyon yöntemidir. Çoklu bağlantı durumu söz konusu olduğunda çok boyutlu verilerde ridge regresyon yönteminin klasik yöntemden daha iyi sonuçlar verebileceği gösterilmiştir. Ridge regresyon, klasik modele eklenen ceza parametresi ile katsayıların sıfıra doğru daraltılması yaklaşımı ile çalışmaktadır (Hoerl & Kennard, 1970).

Lasso regresyon yöntemi, Tibshirani (1996) tarafından ridge regresyona alternatif olarak geliştirilmiştir. Ridge regresyon yönteminde katsayıların sıfıra daraltılması nedeniyle, bu teknik ile elde edilen modelin çok boyutlu verilerde yorumlanması zorlaşmaktadır. Lasso regresyon ise modele katkı sağlamayan değişkenlerin katsayılarını sıfıra eşitleyerek bu problemi çözmektedir. Çoklu bağlantı durumunda ve çok boyutlu verilerde klasik yönteme göre lasso regresyon yönteminin daha doğru sonuçlar verdiği görülmüştür (Tibshirani, 1996).

Elastik-Net regresyon, lasso regresyon tekniğinin kısıtlarının ortadan kaldırılarak çoklu bağlantı sorunu olan durumlarda en iyi modelin belirlenmesi için Zou ve Hastie (2005) tarafından geliştirilmiş bir yöntem olup ridge düzeltmesi ve lasso düzeltmesinin birleşimidir (Zou ve Hastie, 2005).

Bu çalışmada, simülasyon çalışması ile dört kategorili sıralı bağımlı değişken ve aralarında farklı derecelerde (zayıf-orta-yüksek) çoklu bağlantı olan birden çok bağımsız değişken bulunan çok boyutlu veri seti üretilerek, sıralı lojistik regresyon yöntemi ile alternatif yöntemler olan sıralı lojistik ridge regresyon, sıralı lojistik lasso regresyon ve sıralı lojistik elastik-net regresyon için en iyi doğru sınıflandırma performansları karşılaştırılmıştır. Analizler R istatistiksel programlama dili ile R-Studio programında gerçekleştirilmiştir.

Bu çalışma ile sıralı lojistik regresyon modelinin doğru sınıflandırma performansına ilişkin 864 farklı deneme kapsamlı şekilde yorumlanarak literatüre katkı sağlanması amaçlanmaktadır.

Birinci bölümde çalışma ile ilgili temel bilgiler verilerek çalışmaya ilişkin literatür taraması yer almaktadır. İkinci bölümde lojistik regresyon analizi başlığı altında lojistik regresyon modeli tüm yönleriyle ele alınmıştır. Çalışmanın üçüncü bölümünde sıralı

lojistik regresyon modeli anlatılmış ve alternatif düzenlileştirme tahmin yöntemleri olan sıralı lojistik ridge, sıralı lojistik lasso, sıralı lojistik elastik-net regresyon yöntemlerine yer verilmiştir. Dördüncü bölümde simülasyon çalışması ve veri üretimi, model oluşturma süreci yer alıp, analiz sonuçları, tartışma ve öneriler Beşinci bölümde verilmektedir.

1.1 Literatür Taraması

Bu çalışmanın konusu ile paralel olan, sıralı lojistik regresyon ile ilgili literatürde yapılan bazı araştırma sonuçları aşağıda verilmiştir.

McCullagh (1980) çalışmasında, sıralı veri setleri üzerine regresyon modelleri geliştirmiş ve bu regresyon modellerini uygulamalı örneklerle tartışıp sunmuştur.

Anderson ve Philips (1981) çalışmalarında, sıralı kategorik değişkenlerin analizini lojistik model aracılığıyla yapmışlar, aracısız değerlendirme ve ilgili değişkenlere dayanan derecelendirme ölçeklerinin elde edilebilmesi için yeni bir teknik üzerinde çalışmışlardır. Modellerin uygulanmasındaki karşılaşılan zorluklardan ve sıralı bağımlı değişkenler ile oluşturulan modellerin olası kullanımlarından bahsettikleri bir sonuca ulaşmışlardır.

Greenland (1994) çalışmasında, kümülatif olasılık modeli ve süreklilik oranı modeli gibi iki tür sıralı lojistik regresyon modeli için daha önceden bahsedilmemiş modelleri araştırarak daha fazla esneklik yaratan üçüncü bir model önerisi yapmıştır. Model, kömür madencilerinin akciğerde toz birikmesi (pnömokonyoz) hastalığı ile ilişkilerini betimlemek için geliştirilmiştir. Geliştirilen model başka modeller için bir alternatif olabilmiştir.

Tutz ve Hennevogl (1996) çalışmalarında, genelleştirilmiş doğrusal modellerin özgün durumları açısından sıralı regresyon modellerinin doğrusal belirleyicilerini tesadüfi etkiler barındıracak şekilde genişleterek büyütmüşlerdir. Tesadüfi etkiler, kümülatif modeldeki eşiklerin kayması veya kovaryansların konuya özgü ağırlıkları olarak açıklanmıştır.

Ananth ve Kleinbaum (1997) çalışmalarında, sıralı kategoriler için birçok farklı modelin sunulduğunu ancak bu modellerin yeterince kullanılmadığından bahsetmişlerdir. Ayrıca, sıralı kategorik verilerinin modellenebilmesi için istatistiksel teknikler tanıtmışlardır. Bu teknikler, kümülatif lojit modeli, kısıtlı ve kısıtsız kısmi orantısal olasılık modelleri, bitişik kategori lojit modeli, süreklilik oranı modeli, stereotip lojistik modelleri, polimerik lojistik modelleridir.

Chen ve Hughes (2004) çalışmalarında, demografi ile ilgili bağımsız değişkenler ve üniversite deneyimine dair farklı düzeylerde öğrenci memnuniyeti sıralı bağımlı değişken olmak üzere ve çoğunluğu yetişkin olmamış sağlık bilimleri merkezindeki öğrenci ortamı arasındaki ilişkileri incelemek için sıralı regresyon yönteminden yararlanmışlardır.

Zhou ve diğerleri (2008) çalışmalarında, tasarımcılar, müşteriler, duygusal ürün tasarımı alanına kadar sıralı lojistik regresyon analizi kullanmışlardır. Tasarım barındıran unsurlar ve müşterileri etkileyen ihtiyaçlar arasındaki niceliksel ilişkileri ortaya koymak için sıralı lojistik regresyon ve ayrıca ağırlıklı sıralı lojistik regresyon analizi uygulamışlardır.

Emeç (2002), 1317 hane üzerinde hane halkı tüketim harcamalarına yönelik Ege bölgesi için yapılan çalışmasında sıralı lojistik model kullanarak bir analiz yapmıştır. Sıralı lojistik modelin kullanılmasındaki nedenin bireylerin harcama olasılıklarının belirli farklı durumlara göre tahmin edilmesi olduğu belirtilmektedir.

Ayhan (2006), Türkiye'deki hemşirelerin iş bırakma niyeti ile sosyodemografik özellikler, çalışma motivasyonu ile iş memnuniyeti arasındaki ilişkiyi incelemek için oluşturduğu tez çalışmasında sıralı lojistik regresyon tekniğini kullanmıştır. Sağlık sisteminde çalışan hemşirelere kurumları ve idarecileri tarafından gösterilen olumlu davranışların, hemşirelerin iş memnuniyetini arttırdığı ve iş bırakma niyetlerini azaltacağı sonuçlarına ulaşılmıştır.

Nizam ve Akdeniz (2007) çalışmalarında, 2005 yılına ait Sağlık Bakanlığı yataklı tedavi kurumları verileri üzerinden bu kurumların kapasite kullanım oranlarını analiz ederek incelemişlerdir. Kapasite kullanımı az olan hastanelerin daha yüksek kategorilerde yer alan hastanelerde sunulan hizmetleri verebilme durumlarını tahmin etmek için sıralı lojistik regresyon analizi uygulamışlardır.

Şerbetçi (2012), lojistik regresyon modelini detaylı bir şekilde ele aldığı tez çalışması yapmıştır. Uygulama bölümünün sonucunda oluşturduğu sıralı model ile Erzurum Atatürk Üniversitesi iktisat ve ekonometri bölümü öğrencilerinin istatistik ve ekonometri derslerindeki başarılarını etkileyen bileşenler belirlenerek tartışılmıştır.

Akın ve Şentürk (2012) çalışmalarında, 2007 yılında yapılmış Avrupa Yaşam Kalitesi Anketi aracılığıyla edinilen ikincil verilere dayalı sıralı lojistik regresyon yöntemini uygulayarak bireylerin sosyodemografik özelliklerine göre mutluluk düzey durumlarını araştırmışlardır.

Deniz Başar (2012) çalışmasında, İstanbul'daki bir üniversitede öğrenim gören 460 öğrenciye uygulanan anket çalışması ile elde edilen verilerle sıralı lojistik regresyon

analizi gerçekleştirmiştir. Elde edilen sonuçlar ile öğretim üyelerinin unvanını belirleyecek modelin anlamlandırılması amaçlanmıştır.

Yavuz ve diğerleri (2014) çalışmalarında, lojistik regresyon ve sıralı lojistik regresyon yöntemlerini teorik olarak detaylı şekilde anlatmışlardır. Uygulama bölümünde, Erzincan Üniversitesi öğrencilerinin şehir memnuniyetlerini sosyodemografik değişkenler aracılığıyla inceleyerek, öğrencilerin kent memnuniyetini etkileyen faktörleri sıralı lojistik regresyon analizi kullanarak araştırmışlardır.

2. LOJİSTİK REGRESYON ANALİZİ

2.1. Lojistik Regresyon Modelinin Genelleştirilmiş Doğrusal Modeller ile İlişkisi

Regresyon analizi, bir bağımlı değişkenin diğer açıklayıcı değişkenlere bağımlılığını ve bağlanım özelliklerini ortaya çıkarmak için kullanılan istatistiksel bir çözümleme tekniğidir.

Literatürde genellikle "Generalized Linear Models" olarak adlandırılan Genelleştirilmiş Doğrusal Modeller, yapay zekâ, derin öğrenme, makine öğrenmesi dallarında oldukça yaygın olarak kullanılan istatistiksel analizlerdendir. Modellerdeki bağımlı değişkeni açıklayan anlamlı değişkenlerin incelenmesi veya değişkenler arasındaki eğilimle model çıktılarının tahmin edilebildiği doğrusal regresyon modelinin denklemi;

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon \tag{2.1}$$

Burada bağımlı değişken y, bağımsız değişkenler x, regresyon katsayıları β, modeldeki bağımsız değişken sayısı k ve hata terimi ε tarafından temsil edilmektedir.

Doğrusal regresyon modellerinde parametre tahmini için genellikle "En Küçük Kareler" (ordinary least squares) tekniği kullanılmaktadır. Buna göre hata terimlerinin sıfır ortalamalı ve sabit bir varyans ile normal dağıldığı, hata terimlerinin birbirlerinden bağımsız olduğu ve bağımsız değişkenlerin arasında yüksek seviyede çoklu doğrusal bağlantının bulunmadığı şartlarına göre doğrusal regresyon modeli incelenebilmektedir. Doğrusal olmayan regresyon modeli ise, parametrelerin tümünün doğrusal olduğu kabul edilen doğrusal regresyon modeline karşıt olarak, tümü veya bazı parametrelerin doğrusal olmaması durumunda kullanılan modellerdir. Ancak doğrusal regresyon modeline benzer şekilde hata terimlerinin sabit varyans ile ve birbirlerinden bağımsız olduğu durumlarıyla incelenmektedir.

Genelleştirilmiş doğrusal regresyon modelinde modelin doğru sonuçlar vermesi için bulunan koşullar her zaman sağlanamamaktadır. Böyle durumlarda bağımlı değişken için genellikle değişken dönüşümleri yapılabilmektedir. Ancak bağımlı değişkenin kategorik olması durumunda değişken dönüşümleri doğrusal regresyon modelinin şartlarını yerine getirmede yetersiz olabilmektedir. Bağımlı değişkenin kategorik olduğu durumlarda farklı regresyon modelleri önerilmektedir. Uygulamada lojistik regresyon modeli en fazla tercih edilen yöntemdir. Yapısal durumuna bağlı şekilde bağımlı değişken ikili, sıralı, sınıflayıcı olduğu durumlarda çeşitli lojistik regresyon modelleri bulunmaktadır.

Son yılların en çok tercih edilen ve uygulanan yöntemi olan lojistik regresyon analizi; değişken türü ve dağılımına bağlı varsayımlarının az olması, analiz sonuçlarının açıklayıcı şekilde yorumlanıp incelenebilmesi nedeniyle lojistik regresyon modeli sık kullanılan bir teknik haline gelmiştir (Alpar, 2018).

Lojistik regresyonun genelleştirilmiş doğrusal regresyon modeline göre önemli farkı; lojistik regresyon modellerinde bağımlı değişkenin kategorik ve düzeylerinin ikili veya çoklu olabilmesidir. Bu önemli ayrım parametrik model seçimlerine ve dağılımsal varsayımlara etki etmektedir. Bu lojistik regresyon ve genelleştirilmiş doğrusal regresyon arasında üç belirgin fark vardır:

i) Tahmin edilecek bağımlı değişken doğrusal regresyon için sürekli değerler, lojistik regresyon için kesikli değerler alır.
ii) Bağımsız değişkenlerin çok değişkenli normal dağılıma sahip olma şartı doğrusal regresyonda sağlanması gerekirken, lojistik regresyonda bu şartın sağlanması gerekmez.
iii) Doğrusal regresyon yönteminde bağımlı değişkenin alacağı değer kestirilirken, lojistik regresyonda bağımlı değişkenin alabileceği değerlerden birine ait gerçekleşme olasılığı içim kestirim yapılır (Bircan, 2004).

2.2. Lojistik Regresyon Modelin Tanımı ve Matematiksel Gösterimi

Lojistik fonksiyonun kökeni 19. yüzyıla kadar uzanmaktadır (Cramer, 2002). Joseph Berkson tarafından "logit" modeller olarak 1944 yılında tanımlanması ve geliştirilmesiyle de istatistiksel teknik olarak lojistik regresyonun kullanılmaya ve yaygın hale gelmeye başladığı görülmektedir. 1958 yılında yapılmış olan bir çalışmada, gözlem değerleri 0 ve 1 olan regresyon modellerinin analizi hakkında bilgiler sunulmuştur. 1967 yılında, kalp hastalıklarının gözlenmiş olduğu veriler ve büyük verileri analizi üzerine bir yaklaşım hakkında bilgiler verilmiştir (Walker ve Duncan, 1967). 1970 yılında ise birçok çalışma ile logit, probit modeller üzerine geliştirilen araştırmalar çoğalmış ve daha yaygın hale gelmiştir (Cramer, 2002).

1987 yılında Robert ve arkadaşları, hipotez testleri, olabilirlik oranı, ve en çok olabilirlik konularında çalışmalar yapmış ve 1995 yılında Hsu ile Leonard ise Monte Carlo dönüşümünün lojistik regresyon modellerinde uygulanabilirliğini açıklamışlardır (Bircan, 2004).

Lojistik regresyon modelleri, tahmin edilen bağımlı değişkenin, içermiş olduğu kategorik değerlerden birinin gerçekleşme olasılığını kestirmeyi amaçlamaktadır. Genelleştirilmiş doğrusal regresyon modellerine benzer olarak, lojistik dönüşümler bağlantı fonksiyonlarıyla yapıldığından açıklayıcı değişkenler ile kestirilen bağımlı değişken arasında doğrusallık aranmamaktadır. Lojistik regresyon analizinde, bağımsız değişkenlerin kendi aralarında yüksek seviyede bağlantısının bulunmaması ve yeterli örneklem büyüklüğüne sahip olması modellerin daha güvenilir yorumlanmasını sağlamaktadır. Ayrıca, bağımsız değişkenlerin yapısının sürekli ya da kesikli değerler almasının bir koşulu olmamakla birlikte bağımlı değişkenin kategorik bir yapıya sahip olması gerekmektedir.

Bağımlı değişkenin kategorik olduğu regresyon model çıktılarında lojistik regresyon modelleri istatistiksel bir sınıflandırma tekniği olarak kullanılmaktadır. Bağımlı değişkenin hasta ya da sağlıklı, başarılı ya da başarısız gibi iki düzeyli değerlere sahip olan modeller, bağımlı değişken ile bağımsız değişkenler arasındaki ilişkiyi inceleyen ikili lojistik regresyon modelleri olarak bilinmektedir. Bağımlı değişkenin ikiden fazla kategoriye sahip ve belirli bir sırayı takip ettiği durumlarda sıralı lojistik regresyon modelleri ve çok kategorili düzeylerinde çok düzeyli lojistik regresyon modelleri olarak bilinmektedir.

Bağımlı değişkeninin alacağı değerlerin 0 ile 1 aralığında olabilmesi için bir dönüşüm yapılmasına ihtiyaç duyulmaktadır. $\pi(x)/[1-\pi(x)]$ dönüşümü bağımlı değişkenin değer sınırlarını 0 ile ∞ aralığına getirmektedir. Değer sınırlarının $-\infty$ ile $+\infty$ aralığında olabilmesi için ise $\pi(x)/[1-\pi(x)]$ oranının logaritmasının alınmasına ihtiyaç duyulmaktadır. Bağımlı değişken bu dönüşüm yardımıyla, bağımsız değişkenin doğrusal bir fonksiyonu olabilmektedir. $\pi(x)$'i $-\infty$ ile $+\infty$ aralığında tanımlı yapan bu dönüşüm "logit dönüşüm" olarak isimlendirilmiştir (Atabey, 2010).

Logit dönüşümü,

$$\eta_i = x_i'\beta \tag{2.2}$$

$$\eta_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right) \tag{2.3}$$

$$x_i'\beta = \ln\left(\frac{\pi_i}{1-\pi_i}\right) \tag{2.4}$$

(2.4) eşitliğinin her iki tarafının üstel değeri alındığında,

$$\exp\left(x_i'\beta\right) = \exp\left[\ln\left(\frac{\pi_i}{1-\pi_i}\right)\right] \tag{2.5}$$

$$\exp\left(x_i'\beta\right) = \frac{\pi_i}{1-\pi_i} \tag{2.6}$$

$$\exp\left(x_i'\beta\right)(1-\pi_i) = \pi_i \tag{2.7}$$

elde edilir ve parantez içindeki ifadeler dağıtıldığında,

$$\exp\left(x_i'\beta\right) = \pi_i + \pi_i\left(\exp\left(x_i'\beta\right)\right) \tag{2.8}$$

$$\exp\left(x_i'\beta\right) = \pi_i\left(1 + \exp\left(x_i'\beta\right)\right) \tag{2.9}$$

yazılabilir. Buradan π_i değeri,

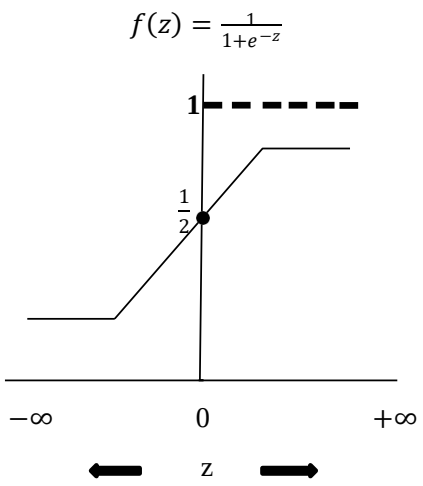
$$\pi_i = \frac{\exp\left(x_i'\beta\right)}{1+\exp\left(x_i'\beta\right)} \tag{2.10}$$

eşitliği ile elde edilir. Diğer bir ifade ile,

$$E(y_i) = \pi_i = \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}} = \frac{1}{1+e^{-x_i'\beta}} \tag{2.11}$$

şeklinde elde edilir. Lojistik regresyon modeli bu şekilde elde edilmektedir.

Lojistik regresyon fonksiyonu Şekil 2.1.'deki gibi bir gösterime sahiptir (Kleinbaum ve Klein, 2010).



Şekil 2.1. Lojistik Regresyon Fonksiyonunun Grafiksel Gösterimi

Lojistik regresyon analizinin bu kadar popüler olmasının nedeni, lojistik regresyona ait fonksiyonun sahip olduğu eğrinin S şeklinde olmasıdır. Bu eğride olasılıklar artan ve azalan z (z; bağımsız değişkenlerin kombinasyonu) değerlerine karşılık fonksiyon değerleri 0 ile 1 arasında kalmaktadır. $z = -\infty$ olduğunda $f(z) \approx 0$ değerini alırken, $z = \infty$ olduğunda $f(z) \approx 1$ değerini almaktadır (Kleinbaum and Klein, 2010).

2.3. Lojistik Regresyon Modelinin Varsayımları

Lojistik regresyon analizi diğer sınıflandırma tekniklerine göre varsayımlarının esnek olmasından dolayı uygulama kolaylığı sağlamaktadır. Ancak lojistik regresyon analizi de istatistiki bir teknik olduğundan varsayımlardan bağımsız olması mümkün değildir.

Lojistik regresyon analizinin doğrusal regresyon analizine göre daha tercih edilebilir olması, yerine gelmesi gerekmeyen bazı varsayımlardan süregelmektedir. Bunlar, bağımsız değişkenler ile bağımlı değişken arasındaki ilişkinin doğrusal olması, hata terimlerinin ve bağımlı değişkenin normal dağılıma sahip olması, bağımlı değişkenin ölçümle belirtilmesi ve homojen varyans yapısında olma gibi varsayımlardır. Bu varsayımların hiçbirini lojistik regresyon modeli yerine getirmek zorunda değildir. Ancak aşağıda verilen noktalar lojistik regresyonda dikkatle incelenmelidir (Sümbüloğlu ve Akdağ, 2007).

Doğrusallık: Lojistik regresyonda, doğrusallık varsayımı aranmamaktadır. Ancak bağımlı değişkenin logiti ile açıklayıcı değişkenler arasında doğrusallık varsayımı gerekmektedir. Eğer logit doğrusallığı olmazsa, lojistik regresyon modeli bağımlı ile bağımsız değişkenler arasındaki ilişkilerin seviyesini düşük düzeyde tahmin edecektir.

Çoklu Bağlantı: Açıklayıcı değişkenlerin kendi aralarında çoklu bağlantı olmamalıdır. Bir açıklayıcı değişken herhangi bir açıklayıcı değişkenin doğrusal fonksiyonu ise lojistik regresyon modelinde çoklu bağlantı problemi ortaya çıkacaktır. Çoklu bağlantı sorunu genelde değişkenler arasındaki yüksek korelasyon nedeniyle ortaya çıkar. Bağımsız değişkenler arasındaki düşük düzeydeki korelasyon sorun yaratmazken, yüksek düzeydeki korelasyon örneğin 0,60'ın üzerinde bir ilişki olması gerçekte etkisi önemli bulunan bir değişkenin modele katkısının istatistiksel olarak önemsiz bulunmasına neden olacağından sorun yaratır (Alpar, 2018).

Lojistik regresyonda çoklu bağlantı için doğrusal regresyonda olduğu gibi tolerans ya da varyans şişme değerlerinden yararlanılması önerilmektedir. 0,10'un altındaki tolerans değeri veya 10'un üzerindeki varyans şişme değeri bulunduğunda modeldeki açıklayıcı değişkenler arasında çoklu bağlantı olduğuna işaret etmektedir (Allison, 2012).

Aşırı Değerler: Lojistik regresyon modelindeki bağımsız değişkenlerde aşırı değerler olması sonuç çıktılarını belirgin derecede etkileyecektir. Aşırı değerler modelden çıkarılması veya ayrı olarak modellenmesi gerekmektedir.

Büyük Örneklemler: Lojistik regresyon analizinde parametre tahmini için genelde en çok olabilirlik (maximum likelihood) tahmincisi kullanılır. Bu yöntem büyük örneklemlerde güvenilir sonuç verdiğinden büyük örneklemler tercih edilmelidir.

2.4. Lojistik Regresyon Modelinde Parametrelerin Kestirilmesi

Lojistik regresyon modelinde parametrelerin kestirilmesinde en çok olabilirlik (maximum likelihood) tekniği kullanılmaktadır (Harrell, 2015). Doğrusal regresyon modellerinde kullanılan en küçük kareler tekniği, lojistik regresyonda kullanılamamaktadır. Bu iki regresyon yöntemi parametre kestiriminde de ayrışmaktadır. Bu yöntemde, gözlem değerlerinin elde edilme olasılıklarını en büyük olmasını sağlayacak değerler belirlenmektedir (Bircan, 2004).

Regresyon katsayılarının tahmin edilmesine dair denklem, aşağıdaki eşitlikte gösterilmiştir.

$$\mathrm{L}(\beta, y_i) = \left(\prod_{i=1}^{n} \pi(X_i)^{y_i}\left(1-\pi(X_i)\right)^{1-y_i}\right) \tag{2.12}$$

Denklemde gözlemlerin birbirlerinden bağımsız olması nedeniyle, olasılık fonksiyonları çarpım şeklinde ifade edilmektedir. Matematiksel olarak kolaylaştırmak adına, aşağıdaki

gibi, yukarıdaki denklemin (Eşitlik 2.12) logaritmasıyla işlem yapılarak, log-likelihood fonksiyonu olarak isimlendirilmektedir.

$$\mathrm{L}(\beta, y_i) = \sum_{i=1}^{n} \left(y_i \log \pi_i + (1 - y_i) \log (1 - \pi_i)\right) \quad (2.13)$$

Eşitlik 2.13 denklemindeki β değerine göre türevleri sıfıra eşitleyerek ve olabilirlik değerlerini elde ederek, denklemi en büyük olmasını sağlayacak katsayıların bulunabilmesini sağlanmıştır.

2.5. Lojistik Regresyon Analizinin Bağımlı Değişkenin Niteliğine Göre Sınıflandırılması

Bağımlı değişkenin kategori sayısına göre ikili (Binomial), çok düzeyli (Multinominal) ve sıralı (Ordinal) lojistik regresyon analizi olarak lojistik regresyon analizleri üç grupta incelenmektedir.

İkili (Binomial) Lojistik Regresyon Analizi: Bağımlı değişkenin iki kategorili olduğu çalışmalarda kullanılmaktadır. Yanıt(Bağımlı) değişken (0,1) olarak kodlanmaktadır. Analizin temel kullanımı, yanıt değişkeni ile bağımsız değişkenler arasındaki nedensellik ilişkilerini araştırmayı amaçlamaktadır (Şahin, 2017). Örnek vermek gerekirse, bir hastanede tedavi gören hastaların İyileşti-İyileşemedi olarak 0 ile 1 şeklinde regresyon modelinin oluşturulduğu çalışmalarda kullanılmaktadır.

Çok Düzeyli (Multinominal) Lojistik Regresyon Analizi: Bağımlı değişkenin ikiden daha fazla kategoride, sırasız olduğu çalışmalarda kullanılmaktadır. Örnek vermek gerekirse; işletme, iktisat ve ekonometri bölümlerinde eğitim görmekte olan üniversite öğrencilerini içeren bir veri setiyle çalışılan bir araştırmada çok düzeyli lojistik regresyon analiz tercih edilmektedir (Çokluk, 2010). Diğer bir ifadeyle, yanıt değişkeninin sınıflayıcı ölçekli, üç veya daha çok kategori barındırdığı durumlarda yanıt değişken ile açıklayıcı değişkenler arasındaki nedensellik ilişkilerinin incelenmesi amacıyla kullanılmaktadır (Zortuk, Koç ve Bayrak, 2014). Kategorilerin ardışık olarak veya sıralı gruplandığı kesin olmadığında çok düzeyli lojistik regresyon analizi uygulanmaktadır. Yanıt değişkeni J adet bir kategoriye sahip olan çok düzeyli lojistik modellerin "$J-1$" adet lojistik regresyon modeline sahip olduğu bilinmektedir (Long ve Freese, 2014).

Sıralı (Ordinal) Lojistik Regresyon Analizi: Bağımlı değişkenin ölçeğinin sıralı veya ardışık gruplanmış olduğu ve bağımlı değişken yapısının birbirlerini etkilediği çalışmalarda kullanılmaktadır (Aktar Demirtaş, Anagün ve Köksal, 2009). Örnek vermek

gerekirse, bir üniversitedeki lisans bölümünde eğitim görmekte olan öğrencilerin “iyi”, “orta” ve “zayıf” olarak gruplanmış olduğu araştırmalarda uygulanmaktadır (Çokluk, 2010).

Lojistik regresyon analizi ikili, çok düzeyli veya sıralı kategorik bağımlı değişken ile sürekli, kesikli veya kategorik bağımsız değişkenler arasındaki ilişkileri analiz etmek için uygulanmaktadır. Bu tekniklere göre, açıklayıcı değişkenlerin yanıt değişkeninin ilgili kategorilere ait olma olasılığı üzerindeki etkileri hesaplanmaktadır. Lojistik regresyon modeli, kategorik bir yanıt değişkeni ile bir veya birden daha çok kategorik veya sürekli/kesikli bağımsız değişkenler arasındaki ilişkileri açıklamak, ilgili hipotezleri oluşturmak, araştırmak ve tüm durumları test etmek için kullanılan bir yöntemdir (Peng, Lee ve Ingersoll, 2002).

3. SIRALI LOJİSTİK REGRESYON ANALİZİ

3.1. Sıralı Lojistik Regresyon Modelinin Tanımı

Doğrusal, lojistik ve sıralı lojistik regresyon modellerinin genellikle bağımlı değişken yapısı ve modele ait varsayımların sağlanmasıyla bağlantılıdır. Sürekli, iki kategorili, çok kategorili veya sıralı çok kategorili ölçülebilen bağımlı değişkenler, yanıt değişkeninin sıralı çok kategorili olarak gözlendiği veya seçildiği araştırmalarda sıralı lojistik regresyon anlamlı sonuçlara ulaşmada en önemli bir tekniktir. Sıralı regresyon modelleri bağımlı değişkenin kategorilerinin düşükten yükseğe şekilde düzenlenmiş olduğu durumları içermektedir (Menard, 2001).

Özellikle sosyal bilimlerde sıralı kategorik veri setleri ile sıkça karşılaşılmaktadır (Long ve Freese, 2014). Birçok çalışmada gözlemlenen birimlerin birçok özelliği, örnek vermek gerekirse gelir durumu, sıralı kategoriler ile ifade edilir veya çalışanların statüleri işçiler en alt düzey, yöneticiler en yüksek düzey olarak hiyerarşik bir sırada incelenmektedir. Ayrıca anket çalışmalarında, likert ölçeği sık olarak kullanılmaktadır. Sıralı lojistik regresyon modelinde, bağımsız değişkenlerin tamamının sürekli veya kategorik olmasına ilişkin herhangi bir şart bulunmamaktadır. Ancak yapılacak olan çalışmalarda bağımsız değişkenlerin sürekli tercih edilmesi önerilmektedir.

Bağımlı değişkenin kategorileri en düşükten en yükseğe doğru sıralanmış olduğu çok kategorili durumlarda çok yakın kategoriler arasındaki gerçek aralıklar genellikle anlaşılamaz. Bu gibi durumlarda eşit aralığa sahip olmayan sıralı kategorilere sahip bağımlı değişkene ait etkilerin ve ilişkilerin belirlenmesinde sıralı lojistik regresyon analizi en uygun yöntem olmaktadır.

Sıralı lojistik regresyon modelinin elde edilmesinde logit, probit, cloglog ve cauchit olarak adlandırılan temel bağlantı fonksiyonları kullanılmaktadır (Long ve Freese, 2014).

Sıralı lojistik regresyon modellerinin genel özellikleri ile şu şekilde özetlenebilir (Chen ve Hughes, 2004; Breslaw ve McIntosh, 1998):

- Sıralı lojistik regresyon modeli, bağımsız değişkenlerin sıralı kategorik değişken üzerindeki etkilerini, normallik ve sabit varyans varsayımlarına gerek olmaksızın, bağlantı fonksiyonu kullanarak açıklar.
- Gözlemlenmemiş sürekli latent (gizli) bir değişkenden düzenlenebilir, gruplanmış ve sıralı bir kategorik değişken olarak tanımlanan bağımlı değişken içermektedir. Fakat kategorilerin eşit aralıklarla oluşup oluşmadığı net olmamaktadır.

• Sıralı lojistik regresyon analizi, regresyon katsayısının değeri kategorik sıralı bağımlı değişkeninin kategorilerine bağlı olmadığından dolayı, bağımsız değişken ile kategorik sıralı bağımlı değişken arasındaki ilişkiyi kategoriden bağımsız olarak varsaymaktadır. Yani logit, probit, cloglog ve cauchit olarak adlandırılan temel bağlantı fonksiyonları ile kestirilen β tahminleri (regresyon katsayıları) her eşik değerinde (kesme noktasında) aynı olmaktadır.

3.2. Sıralı Lojistik Regresyon Modelinin Elde Edilmesi

Peter McCullagh'ın 1980 yılında geliştirmiş ve sunmuş olduğu sıralı lojistik regresyon modeli, gözlemlenebilir bir kategorik değişkenin altında gözlemlenemeyen bir gizli değişkene sahip olduğu varsayımı altında temellendirilmiştir.

Genel şekilde sıralı lojistik regresyon modeli

$$link\left(\gamma_j\right) = \tau_j - \sum \beta_k' x_k \tag{3.1}$$

Eşitlik (3.1) ile gösterilmektedir. Fakat bağımsız değişkenlerin farklı değerleri (bağımsız değişken kategorik ise) sıralı bağımlı değişkenin farklı kategorilerinde daha yüksek oranda bulunuyorsa genelleştirilmiş sıralı lojistik regresyon modeli kullanılır (McCullagh, 1980). Böyle çalışmalarda ölçek bileşeni olarak ifadenin paydasına eklenmektedir. Aşağıda gösterilen bu durum sıralı bağımlı değişkenin kategori sayısı üç ve üçten büyük olur ise gerçekleşebilmektedir (Ishwaran ve Gatsonis, 2000; McCullagh, 1980).

$$link(\gamma_j) = \frac{\tau_j - [\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k]}{\exp\left(\theta_0 + \theta_1 z_1 + \theta_2 z_2 + \cdots + \theta_l z_f\right)} \tag{3.2}$$

Burada, γ_j, j. kategori için birikimli olasılık değeri, τ_j j. kategorinin eşik değeri, $\beta_1 \ldots \beta_k$ regresyon katsayıları, yer parametreleri için $x_1 \ldots x_k$ açıklayıcı değişkenler ve k açıklayıcı değişken sayısı, β ve θ bilinmeyen yer ve ölçek parametreleri vektörü, τ_f bilinmeyen kesme noktaları vektörü ve z_l ölçek parametreleri için bağımsız değişkenleri ifade etmektedir.

Bağlantı fonksiyonları ve özellikleri: Sıralı lojistik regresyon modelin yapılandırılmasında olasılık fonksiyonları olarak tanımlanan bağlantı fonksiyonları kullanılır. Sıralı lojistik regresyon modelinde birçok bağlantı fonksiyonu bulunmaktadır.

Çizelge 3.1'de bu bağlantı fonksiyonları ayrıntılı şekilde gösterilmiştir. Gösterimlerde yer alan "γ" sembolü bir olayın meydana gelme olasılığını temsil etmektedir. Sıralı lojistik regresyon modelinde bir olayın olma olasılığının birikimli olasılıklar aracılığıyla tekrardan tanımlandığının bilinmesi gerekmektedir.

Çizelge 3.1. Bağlantı Fonksiyonları

Fonksiyon	Gösterim	Uygulama Durumu
logit	$ln(\gamma / (1-\gamma))$	Kategori olasılıkları eşit dağıtıldıklarında kullanılır.
probit	$\Phi^{-1}(\gamma)$	Normal dağılıma sahip gizli değişkenin varlığı durumlarında kullanılır.
clog-log	$ln(-ln(1-\gamma))$	Daha yüksek kategorilerde daha yüksek olasılık değerleri için kullanılır.
cauchit	$tan(\pi(\gamma-0.5))$	Çok sayıda uç değerli kategori elde edilmesi durumlarında kullanılır.

Sıralı lojistik modelleri için yapılan çalışmalarda logit ve probit modelleri kullanmak daha uygun tercihler olarak önerilmektedir. Olasılık değerlerinde ani değişim gözleniyorsa diğer bağlantı fonksiyonları denebilmektedir. Birikimli olasılıklar ilk kategoriden itibaren oldukça yavaş bir şekilde artarak, ardından son kategoriye doğru hızlı bir şekilde yakınlaşması durumunda, clog-log bağlantı fonksiyonunun kullanılması önerilmektedir. Kategoriler birçok aşırı değer barındıracak şekilde oluşabiliyorsa cauchit bağlantı fonksiyonunun kullanılması önerilmektedir. Fakat araştırmalardaki veri setlerine en doğru ve uygun bağlantı fonksiyonunun seçilmesine yönelik kesin bir yöntem bulunmamaktadır (Yay ve Akıncı, 2009).

Sıralı lojistik regresyon modellerinin yapılandırılmasında yaygın olarak probit ve logit bağlantı fonksiyonları kullanılmaktadır (Tansel ve Güngör, 2004).

Uygulamalarda logit, probit, clog-log ve cauchit fonksiyonları sıklıkla kullanılarak, modeli en anlamlı yapacak bağlantı fonksiyonu seçilmektedir.

3.3. Sıralı Lojistik Regresyon Modelinin Parametrelerinin Tahmini

Sıralı lojistik regresyon modelinde genelleştirilmiş doğrusal regresyon modelinde yapılan parametre değerlerinin tahmin edilmesi işlemi benzer şekilde uygulanmaktadır. Fakat bu işlem gerçekleştirilirken sıralı lojistik regresyon modeli ile doğrusal regresyon modeli arasında bir teknik farkı vardır. Parametre değerlerinin kestirimi için klasik doğrusal regresyon analizindeki varsayımların karşılanması için en küçük kareler yöntemi uygulanırken, kategorik verilerde parametre değerlerinin kestirimi için sıralı lojistik regresyon analizinde en çok olabilirlik yöntemi uygulanmaktadır.

En çok olabilirlik yöntemi: $x_1, x_2, \dots, x_n$ gözlem değerlerine sahip olan örneklemin, olasılık fonksiyonu $f(x;\theta)$ olan bir anakütleden seçilmiş olduğu varsayıldığında, "θ" kestirimi yapılmak istenen anakütleye ait bilinmeyen bir değer olma özelliği taşımaktadır. Bütün gözlem değerlerinin her biri bağımsız olup her biri için olasılık fonksiyonu $f(x;\theta)$ şeklinde temsil edilmektedir. Bu gözlemler için birleşik olasılık fonksiyonu $f(x_1;\theta)^* f(x_2;\theta)^* \dots .^* f(x_n;\theta)$ şeklinde verilmektedir. Burada gösterilen olasılık fonksiyonu $\theta' nın$ bir fonksiyonu olduğuna göre, $L(\theta; x_1, \mathrm{x}_2, \dots, x_n)$ ve "L" fonksiyonu olabilirlik fonksiyonu şeklinde adlandırılmaktadır.

$$L(\theta; x_1, x_2, \dots, x_n) = \prod_{i=1}^{n} f(x_i;\theta) \tag{3.3}$$

Matematiksel hesaplamaların basitleştirilebilmesi için olabilirlik fonksiyonun logaritması alınmakta, logaritması alınan fonksiyonun bilinen parametrelere göre türevi alınmaktadır. Türevi alınmış olan bu değer sıfıra eşitlenerek sonuç bulunmaktadır. Eşitlikte görüldüğü gibi olabilirlik fonksiyonunu maksimum yapan $\theta' nın$ tahmin değeri, en çok olabilirlik tahmini olarak ifade edilir. Bu parametreler sıralı lojistik regresyon modelinde β değerleri olarak tanımlanmaktadır (Bircan, 2004). İki kategorili bir bağımlı değişkenin bulunduğu veri setlerinde doğrusal lojistik regresyon modelinin oluşturulmasında ilk olarak bilinmeyen $\beta_0, \beta_1, \dots, \beta_k$ parametreleri kestirilmektedir. Parametrelerin tahmini yapılmadan model oluşturmaya geçilmez. Sıralı lojistik modeli oluşturan veri seti için olabilirlik fonksiyonu Eşitlik (3.4)'de gösterilmiştir.

$$L(\beta) = \prod_{i=1}^{n} \binom{n_i}{y_i} p_i^{y_i} (1-p_i)^{n_i - y_i} \tag{3.4}$$

Olabilirlik fonksiyonu β 'nın bir fonksiyonu şeklinde ifade edilmektedir. $L(\beta)$ 'yi en yüksek değere ulaştıracak β değerlerini bulabilmek için olabilirlik fonksiyonunun logaritması Eşitlik (3.5)'de gösterildiği şekilde alınır.

$$\ln L(\beta) = \sum_i \left\{ \ln \binom{n_i}{y_i} + y_i \ln \left(\frac{p_i}{1-p_i}\right) + n_i \ln (1-p_i) \right\}$$
$$= \sum_i \left\{ \ln \binom{n_i}{y_i} + y_i n_i - n_i \ln (1 + e^{[\beta_1 x_1 + \beta_2 x_2 + \cdots \ldots + \beta_k x_k]}) \right\} \quad (3.5)$$

Kısmi türevi değeri bilinmeyen her β parametresine göre alınmış olan Ln olabilirlik fonksiyonunun, oluşan türevlerin sıfıra eşitlenmesi sonucunda değeri bilinmeyen β parametrelerine ilişkin k+1 sayıda doğrusal olmayan denklemler ortaya çıkmaktadır. Ortaya çıkan bu eşitlikler ile her bir parametreye uyumlu şekilde hesaplandığında parametre değerleri elde edilir. Doğrusal tahmin edici olarak isimlendirilen tahmini yapılan regresyon modeli aşağıda gösterilmektedir (Fırat ve Onay, 1999).

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \ldots + \beta_k x_k \quad (3.6)$$

3.4. Sıralı Lojistik Regresyon Modelinin Varsayımı

Sıralı lojistik regresyon modelinde yapısal olarak gerekli olan bağlantı fonksiyonları bulunmaktadır, bu bağlantı fonksiyonları da anlamlı bir paralel eğriler varsayımı altında sıralı lojistik regresyon modellini oluşturmak için gereklidir. Doğru bir lojistik modelin oluşturulabilmesi için paralel eğriler varsayımının anlamlı olması gerekli olmaktadır.

Paralel eğriler varsayımı: Paralel doğrular varsayımı, sonuç değişkeni açıklanırken belirlenmiş olan regresyon katsayılarının, sıralı bağımlı değişkenin tüm kategorilerinde eşit olduğunu varsaymaktadır. Diğer deyişle, açıklayıcı değişkenler ile sonuç değişkeni arasındaki ilişki sonuç değişkenin kategorilerine göre farklılık göstermemelidir. Sıralı lojistik regresyon analizinde, *J* adet kategoriye sahip sonuç değişkeni için *J*-1 adet lojit karşılaştırma ile bu koşul sağlandığında τ_{j-1} tane kesim noktası ve bir adet β parametresi bulunmaktadır (Kleinbaum ve Klein, 2010).

Paralel doğrular varsayımı aracılığıyla bağımlı değişkenin kategorilerinin birbirine paralel oldukları söylenebilmektedir. Bu varsayım sağlanamadığı takdirde kategorilerin paralelliği bozulmaktadır (Fullerton ve Xu, 2012).

Paralel doğrular varsayımını ayrıntılı olarak ifade etmek gerekirse, h'ye eşit ya da daha düşük kategoriye sahip modelin birikimli olasılık modeli ele alınmalıdır.

$$P(y \leq h \mid x) = F(\tau_h - \boldsymbol{x\beta}) \tag{3.7}$$

Eşitlik (3.7)'de verilen birikimli dağılım fonksiyonundaki "$\tau_h - \boldsymbol{x\beta}$" değer birikimli olasılık değeri olup, $\boldsymbol{\beta}$ ise tüm kategoriler için aynı olduğundan Eşitlik (3.7) farklı kesim noktalarıyla iki kategorili model için gösterilmek istenirse,

$$\tau_h - x\boldsymbol{\beta} = (\tau_h - \beta_0) - \sum_{k=1}^{K} \beta_k x_k \tag{3.8}$$

Eşitlik (3.8) şeklinde ifade edilir.

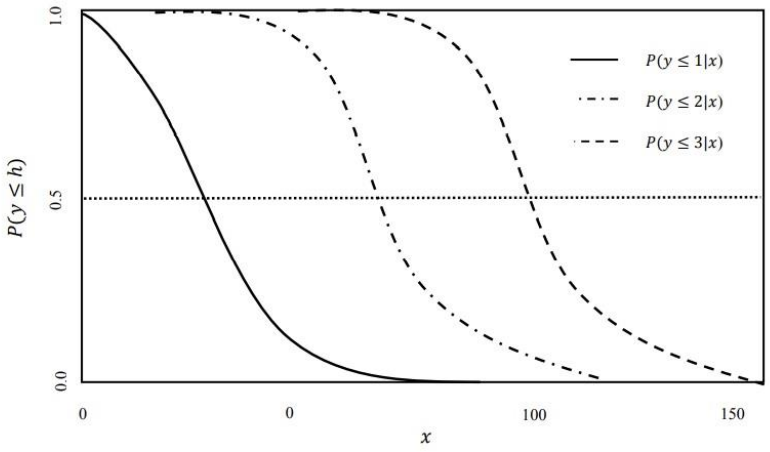
$y \leq 1$ için model, $\tau_1 - \beta_0$ kesim noktasıyla

$$P(y \leq 1 \mid x) = F\left((\tau_1 - \beta_0) - \sum_{k=1}^{K} \beta_k x_k\right) \tag{3.9}$$

ve $y \leq 2$ için model, aşağıdaki gibi elde edilir.

$$P(y \leq 2 \mid x) = F\left((\tau_2 - \beta_0) - \sum_{k=1}^{K} \beta_k x_k\right) \tag{3.10}$$

Eşitlik (3.9) ve Eşitlik (3.10)'da ifade edildiği gibi kesim noktaları değişiklik göstermektedir fakat katsayılar aynı olmaktadır. Kesim noktalarındaki bu farklılık olasılık eğrisinde sola ya da sağa doğru çeşitli farklılıklar neden olur ancak eğrinin şeklini değiştirmez (Long,1997).



Şekil 3.1. Birikimli Olasılık Fonksiyonu

Şekil 3.1.'de dört kategorinin birikimli olasılık değerleri verilmiştir. Sıralı lojistik regresyon analizinde, daha önce değinildiği gibi dört kategorili bağımlı değişken için üç tane birikimli olasılık eğrisi elde edilmektedir. Olasılık değeri olan $0.5'$in, x ekseninde noktalarla temsil edildiği görülmektedir. $0.5'$deki üç olasılık eğrisinin eğim katsayısı Eşitlik (3.11)'de gösterilmektedir (Long, 1997).

$$\frac{\partial P(y \leq 1|x)}{\partial x} = \frac{\partial P(y \leq 2|x)}{\partial x} = \frac{\partial P(y \leq 3|x)}{\partial x} \tag{3.11}$$

Yani bağımlı değişkenin bütün kategorilerinde eğim katsayısı eşit olmaktadır. Böylece eğriler paraleldir yorumu yapılmaktadır.

Paralel doğrular varsayımını test edilmesi: Sıralı lojistik regresyonda modelin güvenilir olabilmesi adına paralel doğrular varsayımının sağlanması gerekmektedir. Bu koşul sağlanamadığı takdirde sonuçlar doğru veya anlamlı olmayabilir. Bu nedenle modelin varsayımının test edilmesi gerekmektedir. Paralel doğrular varsayımını test edip doğrulayabilmek için Olabilirlik Oran Testi ve Wald ki-kare testi uygulanmaktadır (Agresti, 2019). Eğer varsayım sağlanamamış ise önerilen alternatif model yöntemlerinin kullanılması ve modelin sonuçlarının daha dikkatli incelenmesi gerekmektedir (O'Connell, 2006).

Bağımlı değişkeni açıklarken belirlenmiş olan regresyon katsayılarının her bir kategoride birbirine eşit olup olmadığına bakılırken, varsayımın hipotezleri aşağıda verilen şekilde kurulmaktadır.

$H_0: \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 \cdots = \boldsymbol{\beta}_{J-1} = \boldsymbol{\beta}$

H_1: $\boldsymbol{\beta}$ katsayılarından en az biri farklıdır

Yani, modelde bulunan regresyon katsayıları, bağımlı değişkenin tüm kategorilerinde aynıdır hipotezine karşılık, modelde bulunan regresyon katsayıları, bağımlı değişkenin en az bir kategorisinde farklıdır hipotezi test edilmektedir.

3.5. Sıralı Lojistik Regresyon Modelinin Uygunluğunun Test Edilmesi

Sıralı lojistik regresyon modelinin uygunluğunu test etmek, değerlendirmek ve bu modeller içerisinden en uygun modeli seçmek için tüm gözlem değerlerini anlamlı şekilde temsil eden bir istatistiksel ölçü değerine ihtiyaç duyulmaktadır. Bu istatistik ölçüsü

modelin uygunluk değerini ifade ederek tahmin edilen tüm parametrelerin anlamlılığı hakkında bilgi vermeyi amaçlamaktadır.

Regresyonda modelin uygunluğunu test etmek için genellikle Pearson ki-kare istatistiği, sapma ölçüsü ve sözde R^2 kullanılmaktadır. Fakat sözde R^2 değeri ile kategorik bağımlı değişkenlerin olduğu regresyon modellerinde kesin sonuçlara ulaşılamamaktadır (Long, 1997). Sapma ölçüsü de gözlem sayısının yetersiz kaldığı çalışmalarda kesin sonuçlara ulaşmamızı zorlaştırmaktadır (Fujimoto, 2003).

Pearson ki-kare istatistiği,

$$z^2 = \sum_{i=1}^{n} \frac{e_i^2}{\hat{p}_i(1-\hat{p}_i)} \text{ veya } \chi^2 = \sum\sum \frac{(O_{ij}-E_{ij})^2}{E_{ij}} \quad (3.12)$$

şeklinde ifade edilmektedir. Burada, y_i değerleri ile bu değerlerin beklenen değerleri olan $\hat{p}_i$ değerleri arasındaki farkın (e_i^2) , $\hat{p}_i(1-\hat{p}_i)$ 'ye bölünmesiyle bulunan istatistik, model test edilmek istendiğinde, (n-p) serbestlik derecesi ile χ^2 dağılmaktadır. Aynı şekilde sapma ölçüsü ise,

$$D = 2\sum \sum O_{ij} \ln\left(\frac{O_{ij}}{E_{ij}}\right) \quad (3.13)$$

şeklinde ifade edilmektedir. Formülde gözlemlenmiş değerler O_{ij} ve bu değerlerin beklenen değerleri E_{ij}'dir.

Her iki istatistik için de bulunan değer tablo değerinden büyük ve $p < 0{,}01$ ise oluşturulan modelin uygun olmadığı belirtilmektedir.

Modelin uygun olup olmadığını incelemek için kullanılabilecek diğer bir ölçüt sözde R^2'dir. Bu istatistik çoklu belirlilik katsayısı ile büyük ölçüde benzerdir.

$$R^2 = 1 - (\ln L / \ln L_0) \quad (3.14)$$

formülü ile belirlenen bu ölçüt McFadden R^2 istatistiği olarak ifade edilmektedir, L_0, sabit terimin (β_0) tek olarak bulunduğu modellerin en çok olabilirlik değeri. L ise kestirilen tüm parametrelerin bulunduğu modellerin en çok olabilirlik değerini temsil eder (Özdinç, 1999; Fujimoto, 2003).

Modeldeki bağımlı değişken ve bağımsız değişkenler arasındaki ilişkilerin ölçülmesinde kullanılan, değinilmesi gereken diğer sözde R^2 istatistikleri, Cox ve Snell R^2 ve Nagelkerke R^2 istatistikleridir.

$$R_{CS}^2 = 1 - (\ln\ L_0/\ln\ L)^{\frac{2}{n}} \tag{3.15}$$

ve

$$R_N^2 = \frac{R_{CS}^2}{1-L_0^{\frac{2}{n}}} \tag{3.16}$$

şeklinde ifade edilmektedirler.

3.6. Sıralı Lojistik Regresyon Modelinde Parametrelerin Yorumu

Sıralı lojistik regresyon modelinin parametreleri üç değişik şekilde yorumlanabilmektedir (Long, 1997; Tansel ve Güngör, 2004).

Standartlaştırılmış katsayılara göre yapılan yorum: Sıralı lojistik regresyon analizinde $y^* = x\beta + \varepsilon$ olarak ifade edilir ve x_k'lara göre y^* üzerindeki kısmi değişme $\frac{\partial y^*}{\partial x_k} = \beta_k$ olarak gösterilmektedir. Lojistik modeller doğrusal olduğundan, kısmi değişme şöyle yorumlanır: yorumlanan değişken dışındaki tüm değişkenler sabit olduğunda, $x_k$üzerindeki bir birimlik değişim, y^* 'da β_k birim kadar değişime sebep olur.

Eğer gizli y^* için mutlak standart sapma σ_{y^*} ise, x_k için y^* standartlaştırılmış katsayı, $\beta^{s_{y^*}}{}_k = \frac{\beta_k}{\sigma_{y^*}}$ 'dır ve şu şekilde yorumlanır: Kendisi dışında tüm değişkenler sabit olduğunda, x_k üzerindeki bir birimlik değişme, y^* üzerinde, $\beta_k^{s_{y^*}}$ standart sapması kadar değişime sebep olmaktadır.

Eğer x_k 'nın standart sapması σ_k ise, standartlaştırılmış katsayı $\beta_k{}^s = \frac{\sigma_k \beta_k}{\sigma_{y^*}} = \sigma_k \beta_k^{s_{y^*}}$ olarak gösterilmektedir ve bu katsayının yorumu şu şekildedir: Kendisi dışındaki tüm değişkenler sabit olduğunda, X_k üzerindeki bir standart sapması kadar değişim, y^* üzerinde, $\beta_k{}^s$ standart sapma kadar değişime sebep olmaktadır.

Tahmin edilen olasılıklara göre yapılan yorum: Sıralı lojistik regresyon modelinin yorumlanması için değinilen diğer yaklaşım, olasılıklardaki kısmi değişimin belirlenmesi durumudur.

$$\Pr(y_i = m \mid x_i) = F(\tau_m - x_i\beta) - F(\tau_{m-1} - x_i\beta) \quad (3.17)$$

Yukarıda bulunan formülün x_k 'ya göre kısmi türevleri alınmaktadır. Sonuç olarak da

$$\frac{\partial \Pr(y_i=m|x_i)}{\partial x_k} = \beta_k[f(\tau_m - x_i\beta) - f(\tau_{m-1} - x_i\beta)] \quad (3.18)$$

eşitliği elde edilmektedir. Kısmi değişme tüm değişkenlerin düzeylerine bağlı olarak gerçekleşmektedir. Bu yüzden etkiyi belirlerken değişkenlerdeki hangi değerlerinin alınacağına karar verilmesi gerekmektedir. Yukarıda bulunan denklemde tüm gözlemlerin ortalaması alınarak yorumlar çıkan sonuca göre değerlendirilmektedir.

$$\text{ort}\ \frac{\partial \Pr(y_i=m|x_i)}{\partial x_k} = \beta_k[f(\tau_m - \bar{x}\beta) - f(\tau_{m-1} - \bar{x}\beta)] \quad (3.19)$$

Fakat kısmi etki, x_k üzerindeki bir birimlik değişmeye karşılık gözlemlenen olasılıkta meydana gelen değişmeyi tam olarak gösteremez. Tahmin edilmiş olan olasılıkların yorumu şöyle yapılabilir: Bağımsız değişken üzerindeki bir birimlik değişimin sonuç değişkeni kategorisinde meydana gelme olasılığına katkısını ifade etmektedir.

Kesikli değişmeye göre yapılan yorum: Yüksek kategorilerin olasılık değeri daha çok olduğunda, sıralı lojistik model hakkında kısmi etkileri kullanarak değerlendirme yapmak tutarlı olmayacaktır (Long, 1997). Dolayısıyla kesikli değişmeye göre yapılan yorumlama daha güvenilirdir. Açıklayıcı (bağımsız) değişkenin başlangıç değerinden son değerine x_k üzerindeki değişme için kestirilmiş olasılıktaki değişme, kesikli değişme diye adlandırılır. Aşağıda bulunan formül ile ifade edilmektedir.

$$\frac{\Delta \Pr(y_i=m|x_i)}{\Delta x_k} = \Pr(y = m \mid x, x_k = x_{\text{son}}) - \Pr\left(y = m \mid x, x_k = x_{\text{başlangıç}}\right) \quad (3.20)$$

$\Pr(y = m \mid x, x_k)$ ifadesi, x_k üzerinde verilen belirli bir değer için m kategorisinin olasılığına karşılık gelmektedir. Yorumu ise şöyledir: Kendisi dışında tüm değişkenler sabit olduğunda, x_k, $x_{\text{başlangıç}}$ 'dan x_{son} 'a doğru değiştiğinde, m kategorisinin kestirilmiş olasılığı $\frac{\Delta \Pr(y_i=m|x_i)}{\Delta x_k}$ kadar değişecektir.

3.7. Sıralı Lojistik Regresyon Modelinde Çoklu Bağlantı Sorunu

Çoklu bağlantı (multicollinearity), bağımsız değişkenlerin kendi aralarında doğrusal ya da doğrusala çok yakın ilişkiler olması durumu olarak isimlendirilmektedir (Saleh ve diğerleri, 2019).

Regresyon analizinde bağımsız değişkenler her zaman birbirlerinden bağımsız değildir. Birçok çalışmada özellikle sağlık alınanda yapılan araştırmalarda bağımsız değişken bir diğer bağımsız değişkenden etkilenmektedir. Bu durumda uluslararası literatürde "multicollinearity" olarak isimlendirilen çoklu bağlantı sorunu meydana gelmektedir. Çoklu bağlantı problemi; veri setlerindeki verilerin toplanma yöntemi, modelin özellikleri ve kısıtlamaları, modelin eksik veya aşırı tanımlanmış olmasından meydana gelebilecek bir problemdir (Montgomery ve diğerleri, 2012).

Lojistik regresyon modellerinde de açıklayıcı değişkenlerin birbirlerinden bağımsız olması modelin geçerliliği ve yorumların güvenilir olması adına önem taşımaktadır. Bu yüzden çoklu bağlantı problemi olması durumunda kurulan model tutarlı bir sonuç vermeyecektir. Bu tutarsız durum nedeniyle p anlamlılık değerlerine bağlı olarak yanlış anlamlı değişkenlerin seçilmesi, regresyon katsayılarının yanlış kestirilmesi, modelin öngörü gücünün bozulması, regresyon katsayılarının standart hatalarının çok yüksek değerler alabilmesi ve daha geniş güven aralıklarının oluşabilmesi gibi yanlış ve istenmeyen durumlar oluşacaktır (Young, 2017).

Bağımsız değişkenler arasındaki ilişkinin tespit edilebilmesi için, çoklu bağlantı olmasının bir sinyali olarak çeşitli teknikler bulunmaktadır. Aralarındaki korelasyonun belirleneceği değişkenlerin sürekli, kesikli veya sıralı değerler almalarına göre farklı katsayı belirleme yöntemleri mevcuttur. Sürekli değişkenler için Pearson veya Spearman gibi korelasyon katsayıları ve kategorik değişkenler için Phi ve Cramer's V gibi katsayılar kullanılmaktadır. Değişkenlerden birinin sürekli ve diğerinin kategorik yapıda olduğunda ise Nokta Çift Serili (Point Biserial) korelasyon ve VIF (Variance Inflation Factor) tercih edilmektedir.

Pearson Korelasyon Katsayısı: Bağımsız değişkenlerin sürekli veri yapılarına sahip olduklarında, bir diğer bağımsız değişken ile aralarındaki korelasyon katsayıları; korelasyon katsayısı r ve gözlem sayısı n olarak ifade edildiğinde;

$$r = \frac{n(\sum XY) - (\sum X)(\sum Y)}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}} \tag{3.21}$$

olarak hesaplanır ve korelasyon katsayıları -1 ile 1 arasında değer almaktadır (Shrestha, 2020). Hesaplanan korelasyon katsayılarının mutlak 0.6'den az olması genellikle çoklu bağlantı sorunu olmadığını göstermektedir.

Nokta Çift Serili Korelasyon Katsayısı: Pearson Katsayısı'nın özel bir tipi olarak da ele alınan nokta çift serili korelasyon, aralarındaki ilişkinin belirleneceği iki değişkenden birinin kadın - erkek gibi farklı değer aldığı kategorik ve diğerinin sürekli değerler aldığı durumlarda kullanılan yöntemdir. $q = I - p$ ve standart sapmaya bağlı olmak üzere " r " katsayısıyla ifade edilmektedir.

$$r = \frac{(\bar{x}_p - \bar{x}_q)\sqrt{pq}}{ss} \tag{3.22}$$

Phi Katsayısı (Ortalama Kare Kontenjansı Katsayısı- Phi Coefficient): İki değişkenin de iki kategorili (dichotomous) olduğu durumlardaki ilişkinin belirlenmesi için kullanılabilir. 2x2 boyutlarındaki kontenjans tablosu üzerinden işlemler yapılmaktadır. x_{00}, iki değişkenin de 0 değerini aldığı gözlem sayısı; x_{11}, iki değişkenin de 1 değerini aldığı gözlem sayısı ve x_{10} ile x_{01} ise biri 1 diğeri 0 olduğundaki değerler olmak üzere aşağıdaki gibi hesaplanmaktadır ve mutlak 0,7 ile 1 arası değerler şiddetli ilişki olduğunu ifade etmektedir.

$$\Phi = \frac{X^2}{n} \tag{3.23}$$

ve

$$\Phi = \frac{x_{00}x_{11} - x_{01}x_{10}}{\sqrt{x_0.x_1.x_1.x_0}} \tag{3.24}$$

Cramer's V: "Cramers Phi" olarak da bilinen ve değişkenlerin kategorik olduğu (2x2'den fazla, 2x3 veya 2x4 gibi) durumlarda kullanılmaktadır.

Bir diğer yöntem de VIF , Varyans Büyütme Faktörü değerleridir. Toleransa $\left(1 - R_j^2\right)$ bağlı VIF değerleri,

$$VIF_j = \frac{1}{1 - R_j^2} \tag{3.25}$$

formülünden, her bir bağımsız değişken özelinde hesaplanmaktadır. Genellikle, $\text{VIF}_j = 1$ olduğunda çoklu bağlantı probleminin olmadığı, $1 < \text{VIF}_j < 5$ arasındaki değerlerde orta

seviyede bir çoklu bağlantı problemi olabileceği ve $VIF_j \geq 5$ olduğundan yüksek çoklu bağlantı problemi olarak kabul edilmektedir (Young, 2017).

Lojistik regresyon modellerinde aralarında ilişki olan bağımsız değişkenler aşırı öğrenme ve eksik öğrenmeye de sebep olabilmektedir. Çoklu bağlantının probleminden kurtulabilmek için çeşitli yöntemler geliştirilmiştir. En yaygın tercih edileni, modelin yeniden oluşturulması ile ilgili bir yaklaşım olan, değişken seçimidir. Değişken seçimi yaklaşımı, doğrusal ilişkili değişkenlerden birinin veya birkaçının elenmesi şeklinde açıklanmaktadır. Fakat elenen açıklayıcı değişkenlerin modelin açıklanmasına bir katkıları olduğundan bu teknik modelin tahmin gücünü zayıflatmaktadır (Erar, 2013). Değişken seçimi tekniği ya da ek veri toplanarak örneklemin büyütülmesi çoğu durumda çoklu bağlantı sorununun üstesinden gelmekte yetersiz kalır. Bu nedenle, günümüzde özellikle makine öğrenmesi çalışmalarında oldukça fazla yararlanılan alternatif istatistiksel yöntemler tercih edilmektedir.

3.8 Sıralı Lojistik Regresyon Modeli için Alternatif Tahmin Yöntemleri

Çalışmanın bu bölümünde bir önceki başlıkta açıklanan çoklu bağlantı sorununa çözüm olabilecek alternatif istatistiksel yöntemlere değinilmiştir. Çoklu bağlantı sorununun çözülmesinde düzenlileştirme yöntemleri ismi ile de ifade edilen yanlı regresyon modelleri tercih edilmektedir.

3.9. Düzenlileştirme

Bağımsız değişkenler arasında yüksek düzeyde çoklu bağlantı sorunu olduğunda, güvenilir ve tutarlı katsayı kestirimi yapabilmek için büzülme (shrinkage) algoritmaları olarak da ifade edilen, yanlı regresyon teknikleri ridge regresyon, lasso regresyon ve elastik-net regresyon veya genel adıyla düzenlileştirme (regularization) yöntemleri tercih edilmektedir (Chen ve diğerleri, 2019).

Açıklayıcı değişkenler arasındaki çoklu bağlantının sebep olabileceği model ve analiz hatalarını ortadan kaldırmak veya modellerin eksik öğrenme (underfitting) ve aşırı öğrenme (overfitting) durumlarını minimize edebilmek için düzenlileştirme yöntemleri kullanılmaktadır. Lasso regresyon, ridge regresyon ve elastik-net regresyon, farklı tahmin modellerinin kurulmasına ve doğru sınıflandırma problemlerine de bir çözüm olmaktadır. Veri setlerini belirli oranlarda bölerek, eğitim seti ve test seti olarak adlandırılan iki örneklem oluşturabilmektedir. Veri setinin büyüklüğüne göre genellikle bölme oranı %70'e %30 veya %80'e %20 oranlarında olmaktadır. Araştırmacının tercihine göre %60'a %40 oranı da seçilmektedir. Bunun amacı modelin eğitim veri setindeki gözlemler üzerinden eğitilerek bir model oluşturulması ve sonrasında eğitilen modelin test veri

setine ayrılmış gözlemler aracılığıyla sınanmasıdır. Bu işlem ile modelin performansı üzerine fikir edinilmektedir. Aşırı öğrenme, eksik öğrenme yada bağımsız değişkenler arasındaki çoklu bağlantı problemleri, eğitim veri seti ve test veri setinden edinilen performans istatistiklerine yansımaktadır.

Aşırı öğrenme ve eksik öğrenme durumlarından regresyon modellerinin varyans ve yanlılıkları değişmektedir. Bu nedenle aşırı öğrenmede yanlılık düşük ve varyans yüksek olmaktadır. Bağımsız değişkenler arasında çoklu bağlantı sorunu olduğunda da varyans, tahminlerin değişkenliği yüksek olacaktır. Genel bir yaklaşım olarak, esnek modeller oluşturulduğunda varyans yükselirken yanlılık azalacaktır (James ve diğerleri, 2014). Regresyon modellerine bir ceza parametresi uygulanarak ve yanlılıkları arttırılarak, modellerin karmaşıklığının en aza indirilmeye çalışılması Cezalı (penalized) veya Düzenlileştirme olarak ifade edilmektedir.

Düzenlileştirme yöntemleri ile sıralı lojistik regresyon modellerindeki katsayılara, çeşitli ceza terimleri eklenerek regresyon modelleri kurularak kullanılabilmektedir. Bu durumda ceza terimler likelihood denklemine eklenmektedir. Bu çalışmada lojistik regresyon analizinde yaygın olarak uygulanan üç alternatif düzenlileştirme tekniği veya cezalı teknik incelenmiştir: lojistik ridge regresyon, lojistik lasso regresyon ve lojistik elastik-net regresyon. Değinilen üç alternatif düzenlileştirme tekniği de sıralı lojistik regresyon çalışmalarında benzer şekilde uygulanabilmektedir.

3.9.1. Lojistik Ridge Regresyon

Doğrusal regresyonda çoklu bağlantı sorununun giderilmesi için geliştirilen teknikler, lojistik regresyon modeli için de bazı araştırmacılar tarafından uyarlanmıştır. Duffy ve Santer tarafından bağımlı değişkenin iki kategorili olduğu durumlarda kullanılabilen lojistik ridge tahmin edicisi, ilk kez 1989 yılında sunulmuştur, belirtilen lojistik ridge regresyonu Eşitlik (3.12)'de gösterilmiştir.

$$\hat{\beta}_{\text{log-ridge}} = \underset{\beta}{\operatorname{argmin}}\left[-\sum_{i=1}^{n} \{y_i \ln(\pi_i) + (1-y_i)\ln(1-\pi_i)\} + \lambda \sum_{j=1}^{p} \beta_j^2\right] \qquad (3.26)$$

Eşitlik (3.12)' de ifade edilen lojistik ridge tahmin edicisinin log-olabilirlik fonksiyonu ise Eşitlik (3.13)'de gösterilmiştir (Dufy ve Santer, 1989).

$$\ln\, l(\beta) = \sum_{i=1}^{n} y_i \left(\ln\left(\frac{\pi_i}{1-\pi_i}\right)\right) - \sum_{i=1}^{n} \ln\left(1 + e^{x_i{}'\beta}\right) + \lambda \sum_{j=1}^{p} \beta_j^2 \qquad (3.27)$$

Lojistik regresyon modelinde matris şeklinde gösterimiyle ridge regresyonun en çok olabilirlik kestiricisi Eşitlik (3.14)'de verilmiştir.

$$\hat{\beta}_{EÇO} = \left(X'\hat{V}X\right)^{-1}X'\hat{V}\hat{Z} \tag{3.28}$$

$\hat{\beta}_{EÇO,}$ kestiricisi $\hat{V}$ ağırlık matrisi dolayısıyla çoklu bağlantıya hassasiyet göstermektedir.

$$\hat{V} = \text{diag}\,[\hat{\pi}(1-\hat{\pi})] \tag{3.29}$$

Eşitlik (3.14)' de gösterilen $\hat{Z}, i$ 'inci elemanı $\hat{z}_i = \log\left(\hat{\pi}_i\right) + \frac{y_i - \hat{t}_i}{1-\hat{\pi}_i}$ eşitliği ile ifade edilen vektördür. En çok olabilirlik kestiricisinin, beklenen değeri $E\left(\hat{\beta}_{EÇO}\right) = \beta$ ve varyansı $\text{Var}\left(\hat{\beta}_{EÇO}\right) = \left(X'\hat{V}X\right)^{-1}$ olmaktadır.

Lojistik ridge kestiricisinin cezalı regresyon modeli olarak gösterimi ise Eşitlik (3.15)'de ifade edilmektedir.

$$\underset{\beta}{\text{argmin}} \sum_{i=1}^{n} \left\{y_i \log\, \pi(x_i) + (1-y_i)\log\left(1-\pi(x_i)\right)\right\} + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{3.30}$$

Ayar parametresi λ, lojistik ridge regresyon modelinde büzülme miktarını kontrol ederken, hiçbir durumda açıklayıcı değişkenin katsayılarını tümüyle sıfır yapmaz.
Lojistik ridge regresyon analizinde, açıklayıcı değişkenlerin tümü için katsayı kestirimi yapılmaktadır. Bu yaklaşım, modeldeki katsayıların değerlendirilmesini zorlaştırıp, her bir parametre için kestirimi mecbur hale getirdiğinden kestirim değerinin yanlılığını arttırmaktadır.

3.9.2. Lojistik Lasso Regresyon

Cezalı regresyon tekniklerinden biri olan lasso regresyon 1996 yılında ilk kez Tibshirani tarafından sunulmuştur. Yanıt değişkeninin kategorik olduğu çalışmalarda uygulanan lojistik regresyon modeli, lojistik lasso tekniğine uyarlanmıştır (Hastie ve diğerleri, 2005).
Çoklu bağlantı sorunun giderilmesi için değişken seçimi yaklaşımının kullanılması durumunda hangi değişkenlerin regresyon modelinden çıkarılacağı en önemli konulardan biridir. Modelin anlamlılığını arttıran ve model için önemli bir ihtiyaç olan bağımsız değişkenin modelden elenmesi modelin doğruluğuna zarar verecektir. Değişken seçimi ve parametre tahminini aynı zamanda gerçekleştiren düzenlileştirme teknikleri bu amaca

hizmet etmek için önerilmiştir. Lasso regresyon, doğrusal ve lojistik regresyon analizinde bir büzülme ve değişken seçim tekniğidir (Melkumova ve Shatskikh, 2017).

Lojistik regresyon için lasso kestiricileri, negatif log-olabilirlik fonksiyonuna ceza terimini dahil ederek oluşturulmaktadır. Ridge regresyona benzer şekilde, lasso kestiricileri, log-olabilirlik fonksiyonunun maksimizasyonu aracılığıyla elde edilir. Lasso cezalı lojistik regresyon tekniği, Eşitlik (3.30)'da ifade edilmektedir.

$$\hat{\beta}_{\log-LASSO} = \underset{\beta}{\operatorname{argmin}} \left[-\sum_{i=1}^{n} \{y_i \ln(\pi_i) + (1-y_i) \ln(1-\pi_i)\} + \lambda \sum_{j=1}^{p} |\beta_j| \right] \quad (3.31)$$

Lasso regresyon birçok çalışmada yaygın olarak kullanılmasına karşın, dezavantajları da bulunmaktadır. Dezavantajların ilki, bağımsız değişkenler arasındaki yüksek korelasyona karşı tercihini sağlam yapmaması, yani bu değişkenlerden birini rastgele seçip gerisini görmezden gelmesidir. Diğer bir dezavantajı ise, $p > n$ olduğunda yüksek boyut içeren veri setlerinde, son modelde sıfır olmayan değerlere sahip n'den daha fazla bağımsız değişken katsayılarının olabilecekken, en fazla n bağımsız değişken seçebilmesidir. Bir başka dezavantajı ise, lasso kestiricisinin cezalı fonksiyonunun bütün olarak dışbükey olmaması durumudur. Bu şekilde farklı bağımsız değişkenlerin sırasına göre farklı kestirimler yapmasına neden olabilmektedir.

Lojistik lasso regresyon, araştırmalarda yaygın şekilde kullanılması sebebiyle dezavantajlarının giderilmesi üzerine çalışılmıştır. Bu amaçla lasso regresyona yeni bir kısıt eklenerek elastik-net regresyonu sunulmuştur (Zou ve Hastie, 2005). Bir sonraki bölümde lojistik elastik-net tekniği hakkında bilgiler verilmektedir.

3.9.3. Lojistik Elastik-Net Regresyon

Lasso regresyon yöntemi, modeldeki değişken seçimi ve parametre tahminlerinde yüksek boyutlu veriler için oldukça elverişlidir. Fakat lasso regresyonun, veri setinin yapısal durumuna bağlı bazı sorunların üstesinden gelmede dezavantajları vardır (Genç, 2020). Lasso regresyonun önceki bölümde bahsedilen dezavantajlarının azaltılması için 2005 yılında Zou ve Hastie tarafından elastik-net regresyon yöntemi sunulmuştur. Bu yöntem, ridge regresyon ile lasso regresyon arasında bir orta nokta, anlaşma sağlamaktadır (Zou ve Hastie, 2005).

Lojistik elastik-net regresyon, değişken seçimi yanında düzenlileştirme teknikleri adına yeni bir yöntem olarak tanınmaktadır. Varsayımsal veri setleri ve gerçek veri setleri üzerine yapılan birçok uygulamayla, elastik-net regresyonun lasso regresyona benzediği

ifade edilmektedir. Ayrıca elastik-net regresyon tekniğinin bağımsız değişken sayısının oldukça fazla olduğu çalışmalarda daha üstün olabildiği ve doğru sonuçlar verdiği birçok araştırmada belirtilmiştir.

Çalışmada bulunan değişken sayısının veri setindeki gözlem sayısından daha çok olduğu durumlar (p>n), yüksek boyutlu veri setlerini işaret etmektedir. Bu tür veriler geleneksel veri setleri ile farklılaşmaktadır ve istatistiksel analizleri daha önce karşılaşılmamış zorluklar getirmektedir. Yüksek boyutlu sınıflandırma verilerinin de cezalı lojistik regresyon kullanılarak azaltılması hem katsayıları kestirmek hem de değişken seçimini aynı zamanda yapabilmek için, elastik-net tekniği yüksek boyutlu sınıflandırma çalışmalarında elverişli bir şekilde uygulanmaktadır (Algamal ve Lee, 2015).

Değişken seçiminde lasso regresyon yönteminden faydalanırken, yüksek düzeyde iç ilişki, çoklu bağlantı sorununu çözmek için ridge regresyon yöntemini kullanarak denge kurmaya çalışan lojistik elastik-net yöntemi aşağıdaki şekilde ifade edilmektedir.

$$\hat{\beta}_{\text{log-elastic}} = \sum_{i=1}^{n} \{y_i \log \pi(x_i) + (1 - y_i) \log (1 - \pi(x_i))\} + \lambda \left[\frac{1}{2}(1-\alpha) \parallel \beta \parallel_2^2 + \alpha \parallel \beta \parallel_1\right]$$

Ayrıca Zou ve Hastie (2005), β kestiricilerini elde etmek adına elde edilen veri kümesini $(\boldsymbol{y}, \boldsymbol{X})$ artırılmış bir veriye $(\boldsymbol{y}^*, \boldsymbol{X}^*)$ genişletmişler ve $\boldsymbol{X}^*$, Eşitlik (3.17)' de ifade edilmiş olup, $\boldsymbol{y}^*$ ise Eşitlik (3.18)' de gösterilmiştir.

$$X^*_{(n+p,p)} = (1+\lambda_2)^{-\frac{1}{2}} \begin{pmatrix} X \\ \sqrt{\lambda_2 I} \end{pmatrix} \tag{3.32}$$

$$y^*_{(n+p,1)} = \begin{pmatrix} y \\ 0 \end{pmatrix} \tag{3.33}$$

Veri kümesini bu yaklaşım ile artırması sonucunda, elastik-net regresyon bir lasso ceza parametresi şeklinde yazılıp çözülebilir hale gelmektedir. Dolayısıyla elastik-net regresyon, lasso regresyondaki gibi n tane bağımsız değişken yerine, $p > n$ olduğunda yüksek boyutlu verilerdeki tüm p açıklayıcı değişkenlerini aynı zamanda değerlendirebilir hale gelmektedir (Zou ve Hastie, 2005).

4. UYGULAMA

4.1. Simülasyon Çalışması ve Veri Üretimi

Çalışmanın bu bölümünde, sıralı lojistik regresyon klasik yöntemi ile çoklu bağlantı durumu için alternatif yöntemler olan sıralı lojistik ridge regresyon, sıralı lojistik lasso regresyon ve sıralı lojistik elastik-net regresyon yöntemlerinin doğru sınıflandırma performanslarını değerlendirmek amacıyla gerçekleştirilen simülasyon çalışmasının belirlenmesi ve veri üretimi hakkında bilgilere yer verilmiştir. Uygulama sonuçları için dört farklı modelin doğru sınıflandırma oranlarını tahmin etmek için R istatistiksel programlama dili ile R-Studio Version 1.4.1717 programında VGAM ve OrdinalNet R paketleri kullanılarak simülasyon kodları hazırlanmıştır.

Bu çalışmanın analizi kapsamında üretilen her bir birimin sınıflandırılması amacıyla, ilk olarak belirlenmiş olan her kategoriye ait sınıflandırma değeri ya da kategoriye düşme olasılığı hesaplanıp ardından o birimin en yüksek sınıflandırma değerini ya da olasılığını aldığı ilgili kategoriye atanması işlemi yapılmaktadır. Sınıflandırma sürecinde doğru sınıflandırma oranları ne kadar yüksek olursa o modelin sınıflandırma performansının bir o kadar iyi olduğunu göstermektedir. Bu çalışmanın amacı, simülasyon çalışması üzerinden belirlenen tüm denemeler için ilgili analizler yapılarak, değinilen mevcut tekniklerin sınıflandırma performanslarını karşılaştırmak ve yorumlamaktır.

Simülasyon çalışması için gerekli olan algoritma hazırlanırken modellerin varsayımları ve veri setinin özellikleri bakımından belirli koşullara dikkat edilmiştir. Bu koşullar ayrıntılı olarak aşağıda verilmiştir.

- Bağımlı değişkene ait düzey sayısı belirlenmiştir. Sıralı Lojistik Regresyon modellerinde bağımlı değişkenin sahip olması gereken düzey sayısı en az 3 olması gerektiği için, bağımlı değişkenin düzey sayısı 4 kategori olarak seçilmiştir.
- Bağımsız değişken sayılarının sırasıyla 3, 5 ve 7 olarak seçilmiştir. Çalışma için 10 tane bağımsız değişken ($x_{1,\ldots,}\, x_{10}$) üretilmiştir, bunlar arasında çoklu bağlantı durumunda olan değişkenler kullanılmıştır.
- Bağımsız değişkenlerin tamamı sürekli değişken olarak belirlenmiştir.
- *e* hata parametreleri 0 ortalama ve farklı standart sapma ile normal dağıldığı kabul edilerek, 10 tane hata terimi değişkeni oluşturulmuş ve elde edilen hata terimleri bağımsız değişkenlerin üretiminde kullanılmıştır.
- Örneklem büyüklüğünün sırasıyla 400, 4000 ve 16000 olarak belirlenmesine karar verilmiştir.

- Çalışmada kullanılacak olan β_0 , … , β_k regresyon katsayıları bağımsız değişken sayısı ve bağımlı değişkene ait düzey sayısına göre sabitlenmiştir. β_0 ve β_k'lere önceden belirlenmiş olarak atanan sayılarla, model denklemleri oluşturulmuştur.
- Oluşturulan her bir model denklemi için paralel doğrular varsayımı kontrol edilerek, varsayımının uygun olduğu durumda modeller elde edilmiştir.
- Çoklu bağlantıya sahip veri üretebilmek için p ve a=$\sqrt{(1-p^2)}$ değişkenleri üretilmiştir. Bu değişkenler belirlenen bağımsız değişkenlerin denklemlerinde korelasyon düzeyini kontrol etmek için kullanılmıştır (Deniz ve diğerleri, 2011). Veri seti üretilirken p=0.3 olduğunda ilgili değişkenlerin çoklu bağlantı düzeyi zayıf, p=0.6 olduğunda orta, p=0.9 olduğunda yüksek olarak oluşturulmuş ve bu düzeylere uygun şekilde seçilen bağımsız değişken setleri denemelerde kullanılmıştır.
- Eğitim veri seti %60 ve test veri seti %40 olarak bölünmesine karar verilmiştir.
- Eğitim veri setinde en iyi ceza parametresinin değeri model seçim kriterleri ile belirlenip en doğru modelleri oluşturmak amaçlanmıştır. Ceza parametreleri için OrdinalNet paketinde bulunan ordinalNetTune fonksiyonuyla *k*-kat çapraz geçerlilik kullanılarak en uygun değerleri otomatik bularak denemelerin analiz edilmesi sağlanmıştır. Daha sonra test veri seti üzerinde denemelerin doğru sınıflandırma oranları test edilerek sonuçlara ulaşılmıştır.
- Belirlenen her bir simülasyon denemesi için iterasyon sayısı 100 olarak alınmıştır. 100 iterasyon ile yapılan analiz sonuçlarında her denemeye ait doğru sınıflama oranlarına ulaşılmıştır.

Değinilen koşullar ve yöntem için örnek bir denemenin R kodu Ekler'de yer almaktadır.

4.2. Model Oluşturma Süreci

Bu çalışmada, bölüm 4.1'de değinilen koşullarla üretilmiş veri setlerinde mevcut yöntemlerin doğru sınıflandırma performansları, çoklu bağlantının zayıf-orta-yüksek olduğu, gözlem sayılarının bağımlı değişken kategorilerinde farklı büyüklüklerde, yani kategorilerdeki gözlem sayısına göre dengeli (her bir bağımlı değişken kategori için örneklem büyüklüğünün %25'i şeklinde gözlem bulunmaktadır) ve kategorilerdeki gözlem sayısına dengesiz (bağımlı değişkenin ilk kategorisi olan "1" için örneklem büyüklüğünün %55'i ve diğer kategorileri olan "2", "3" ve "4" için örneklem büyüklüğünün %15'i şeklinde gözlem bulunmaktadır) olduğu, bağımsız değişken

sayılarının 3-5-7 olarak bulunduğu ve farklı bağlantı fonksiyonlarının (logit, probit, cloglog, cauchit) kullanıldığı durumlar ile değerlendirilmek üzere denemeler oluşturulmuştur.

Bu simülasyon çalışmasında belirlenen tüm durumları araştırabilmek ve karşılaştırabilmek amacıyla 864 tane farklı deneme elde edilmiş ve analiz sonuçlarına ulaşılmıştır.

Çizelge 4.1, 4.2, 4.3 ve 4.4'te simülasyon çalışmasında kullanılan 864 farklı denemenin tüm versiyonları yer almaktadır.

Çizelge 4.1. Simülasyon Çalışmasında Sıralı Lojistik Regresyon İçin Elde Edilen Denemeleri İçeren Şema

Doğru Sınıflandırma Oranları					
Bağımlı Değişken ve Düzeyi Y; 1-2-3-4; Dengeli ve Dengesiz Kategori Dağılımı; Paralel Doğrular Varsayımı Var					
Çoklu Bağlantı Düzeyi 0.3; 0.6; 0.9	**İterasyon=100**	**Sıralı Lojistik Regresyon**			
Bağımsız Değişken Sayısı	**Örneklem Büyüklüğü**	**logit**	**probit**	**cloglog**	**cauchit**
3; 5; 7	**400; 4000; 16000**	$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \ldots + \beta_k x_k$ model <- vglm(as.ordered(Y) ~ X)			

Çizelge 4.2. Simülasyon Çalışmasında Sıralı Lojistik Ridge Regresyon İçin Elde Edilen Denemeleri İçeren Şema

Doğru Sınıflandırma Oranları					
Bağımlı Değişken ve Düzeyi Y; 1-2-3-4; Dengeli ve Dengesiz Kategori Dağılımı; Paralel Doğrular Varsayımı Var					
Çoklu Bağlantı Düzeyi 0.3; 0.6; 0.9	**İterasyon=100**	**Lojistik Ridge Regresyon**			
Bağımsız Değişken Sayısı	**Örneklem Büyüklüğü**	**logit**	**probit**	**cloglog**	**cauchit**
3; 5; 7	**400; 4000; 16000**	$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \ldots + \beta_k x_k$ modelr <- ordinalNetTune(X, Y, 0)			

Çizelge 4.3. Simülasyon Çalışmasında Sıralı Lojistik Lasso Regresyon İçin Elde Edilen Denemeleri İçeren Şema

Doğru Sınıflandırma Oranları					
Bağımlı Değişken ve Düzeyi Y; 1-2-3-4; Dengeli ve Dengesiz Kategori Dağılımı; Paralel Doğrular Varsayımı Var					
Çoklu Bağlantı Düzeyi 0.3; 0.6; 0.9	İterasyon=100	Lojistik Lasso Regresyon			
Bağımsız Değişken Sayısı	Örneklem Büyüklüğü	logit	probit	cloglog	cauchit
3; 5; 7	400; 4000; 16000	$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \ldots + \beta_k x_k$ modell <- ordinalNetTune(X, Y, 1)			

Çizelge 4.4. Simülasyon Çalışmasında Sıralı Lojistik Elastik-Net Regresyon İçin Elde Edilen Denemeleri İçeren Şema

Doğru Sınıflandırma Oranları					
Bağımlı Değişken ve Düzeyi Y; 1-2-3-4; Dengeli ve Dengesiz Kategori Dağılımı; Paralel Doğrular Varsayımı Var					
Çoklu Bağlantı Düzeyi 0.3; 0.6; 0.9	İterasyon=100	Lojistik Elastik-Net Regresyon			
Bağımsız Değişken Sayısı	Örneklem Büyüklüğü	logit	probit	cloglog	cauchit
3; 5; 7	400; 4000; 16000	$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \ldots + \beta_k x_k$ modelen <- ordinalNetTune(X, Y, 0.5)			

Bu tez çalışmasında yapılan simülasyon çalışmasına ait tüm analiz ve sonuç çıktıları 5. Bölüm olan "Sonuçlar ve Tartışma" kısmında paylaşılmıştır.

5. SONUÇLAR VE TARTIŞMA

5.1 Simülasyon Çalışmasının Sonuçları

Bölüm 3'te detaylı şekilde anlatılan sıralı lojistik regresyon yöntemi ile sıralı lojistik ridge regresyon, sıralı lojistik lasso regresyon ve sıralı lojistik elastik-net regresyon yöntemlerinin doğru sınıflandırma performanslarını değerlendirmek amacıyla uygulanan analizlere ilişkin her denemeye ait doğru sınıflandırma oranlarını içeren sonuçlara bu bölümde değinilecektir.

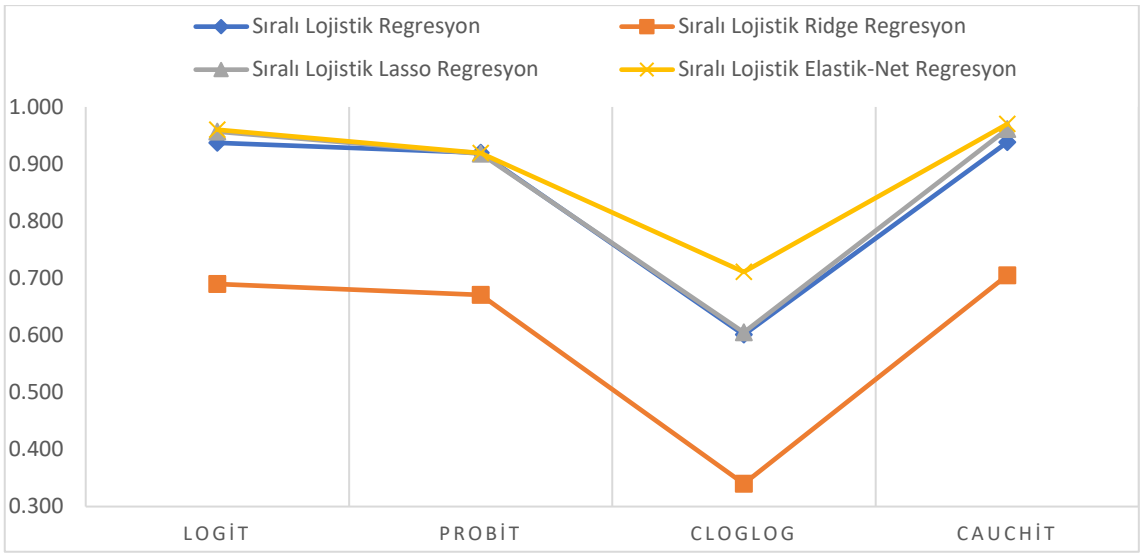
Bu bölümde 864 farklı denemeden 144 tanesi Şekil 5.1'den Şekil 5.9'a kadar grafiklerle verilmiş ve Çizelge 5.1'den Çizelge 5.9'a kadar ayrıntılı şekilde incelenmiş ve yorumlanmıştır. Diğer tüm denemelere ait çizelgeler Ekler'de sunulmuştur.

Çizelge 5.1'de bağımsız değişken sayısı 5, çoklu bağlantı düzeyi yüksek, örneklem büyüklüğünün 16000 olduğu dengeli kategori dağılımına sahip denemeye ait sınıflandırma oranları farklı bağlantı fonksiyonları ve farklı sıralı lojistik modellerine göre verilmiştir.

Çizelge 5.1. 1. Deneme Kümesinin Doğru Sınıflandırma Oranları

Bağımsız Değişken Sayısı	**5**			
Çoklu Bağlantı Düzeyi	**0.9**			
Örneklem Büyüklüğü	**16.000**			
Dengeli Kategorili				
Bağlantı Fonksiyonu	**Sıralı Lojistik Regresyon**	**Sıralı Lojistik Ridge**	**Sıralı Lojistik Lasso**	**Sıralı Lojistik Elastik-Net**
Logit	0.937*	0.690	0.957	0.960
Probit	0.920	0.671	0.918	0.919
Cloglog	0.601	0.340	0.605	0.711
Cauchit	0.938	0.705	0.961	**0.970**

**sınıflandırma matrisi ek 19'da yer almaktadır*



Şekil 5.1. 1. Deneme Kümesinin Doğru Sınıflandırma Oran Grafiği

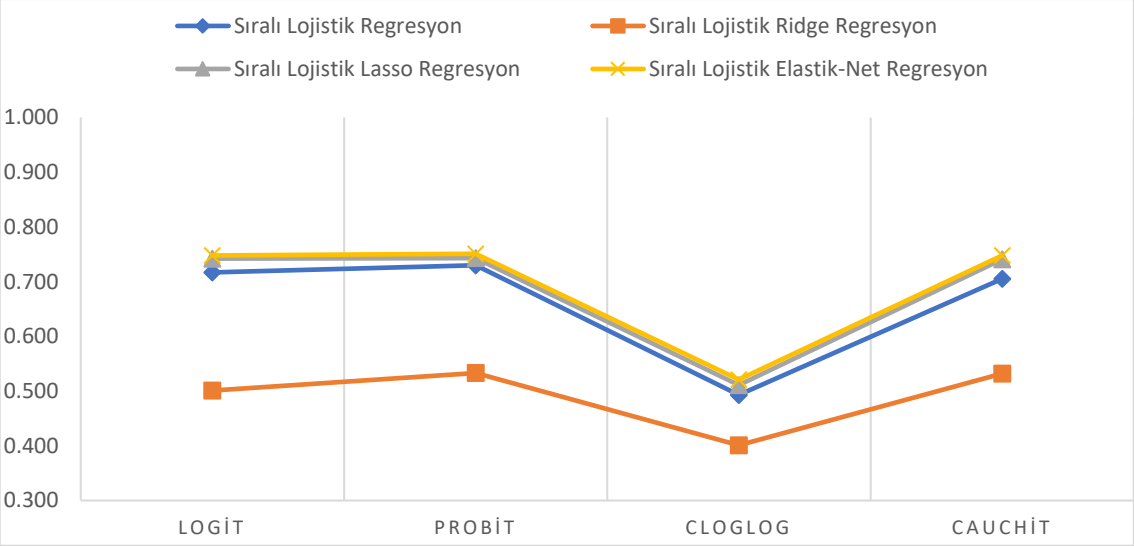
Çizelge 5.1 ve Şekil 5.1 incelendiğinde, sıralı lojistik regresyon logit bağlantı fonksiyonu ile elde edilen denemenin analizinde doğru sınıflandırma oranı %93.7 iken, sıralı lojistik ridge regresyon probit bağlantı fonksiyonunun %67.1, sıralı lojistik lasso regresyon cloglog bağlantı fonksiyonunun % 60.5, sıralı lojistik elastik-net regresyon yönteminin cauchit bağlantı fonksiyonunun %97.0 ile en yüksek yüzdeye sahip olduğu görülmektedir. Simülasyon çalışmasında beklendiği gibi yüksek düzeyde çoklu bağlantı sorununun olduğu denemelerde düzenlileştirme yöntemleri daha iyi sınıflandırma performansı vermiştir. 1. deneme kümesi için analiz sonuçlarına göre sıralı lojistik elastik-net yönteminin en iyi deneme, cauchit bağlantı fonksiyonunun ise en yüksek oranı verdiği görülmektedir.

Çizelge 5.2'de bağımsız değişken sayısı 3, çoklu bağlantı düzeyi zayıf, örneklem büyüklüğünün 400 olduğu dengesiz kategori dağılımına sahip denemeye ait sınıflandırma oranları farklı bağlantı fonksiyonları ve farklı sıralı lojistik modellerine göre verilmiştir.

Çizelge 5.2. 2. Deneme Kümesinin Doğru Sınıflandırma Oranları

Bağımsız Değişken Sayısı	**3**			
Çoklu Bağlantı Düzeyi	**0.3**			
Örneklem Büyüklüğü	**400**			
Dengesiz Kategorili				
Bağlantı Fonksiyonu	**Sıralı Lojistik Regresyon**	**Sıralı Lojistik Ridge**	**Sıralı Lojistik Lasso**	**Sıralı Lojistik Elastik-Net**
Logit	0.717	0.501	0.742	0.748
Probit	0.730	0.530	0.743	**0.751**
Cloglog	0.493	0.401	0.511	0.521
Cauchit	0.705	0.532*	0.741	0.748

**sınıflandırma matrisi ek 20'de yer almaktadır*



Şekil 5.2. 2. Deneme Kümesinin Doğru Sınıflandırma Oran Grafiği

Çizelge 5.2 ve Şekil 5.2 incelendiğinde, sıralı lojistik regresyon logit bağlantı fonksiyonu ile elde edilen denemenin analizinde doğru sınıflandırma oranı %71.7 iken, sıralı lojistik ridge regresyon cloglog bağlantı fonksiyonunun %40.1, sıralı lojistik lasso regresyon cauchit bağlantı fonksiyonunun % 74.1, sıralı lojistik elastik-net regresyon yönteminin probit bağlantı fonksiyonunun %75.1 ile en yüksek yüzdeye sahip olduğu görülmektedir. Simülasyon çalışmasında beklendiği gibi zayıf düzeyde çoklu bağlantı sorununun olduğu denemelerde de doğru sınıflandırma performansları klasik modelin performansı görece daha yakın olsa da düzenlileştirme yöntemleri daha iyi sınıflandırma performansı vermiştir. 2. deneme kümesi için analiz sonuçlarına göre sıralı lojistik elastik-net

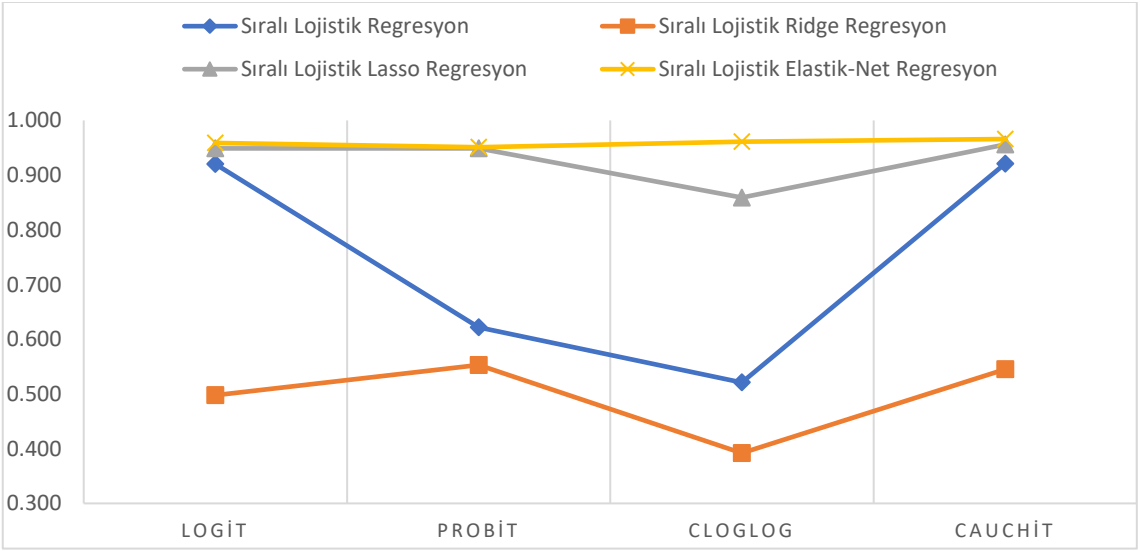
yönteminin en iyi deneme, probit bağlantı fonksiyonunun ise en yüksek oranı verdiği görülmektedir.

Çizelge 5.3'te bağımsız değişken sayısı 5, çoklu bağlantı düzeyi orta, örneklem büyüklüğünün 400 olduğu dengesiz kategori dağılımına sahip denemeye ait sınıflandırma oranları farklı bağlantı fonksiyonları ve farklı sıralı lojistik modellerine göre verilmiştir.

Çizelge 5.3. 3. Deneme Kümesinin Doğru Sınıflandırma Oranları

Bağımsız Değişken Sayısı	**5**			
Çoklu Bağlantı Düzeyi	**0.6**			
Örneklem Büyüklüğü	**400**			
Dengesiz Kategorili				
Bağlantı Fonksiyonu	**Sıralı Lojistik Regresyon**	**Sıralı Lojistik Ridge**	**Sıralı Lojistik Lasso**	**Sıralı Lojistik Elastik-Net**
Logit	0.920	0.498	0.949	0.959
Probit	0.622	0.553	0.949	0.951
Cloglog	0.521	0.392	0.859	0.961*
Cauchit	0.921	0.545	0.956	**0.966**

**sınıflandırma matrisi ek 21'de yer almaktadır*



Şekil 5.3. 3. Deneme Kümesinin Doğru Sınıflandırma Oran Grafiği

Çizelge 5.3 ve Şekil 5.3 incelendiğinde, sıralı lojistik regresyon probit bağlantı fonksiyonu ile oluşturulan denemenin analizinde doğru sınıflandırma oranı %62.2 iken, sıralı lojistik ridge regresyon logit bağlantı fonksiyonunun %49.8, sıralı lojistik lasso

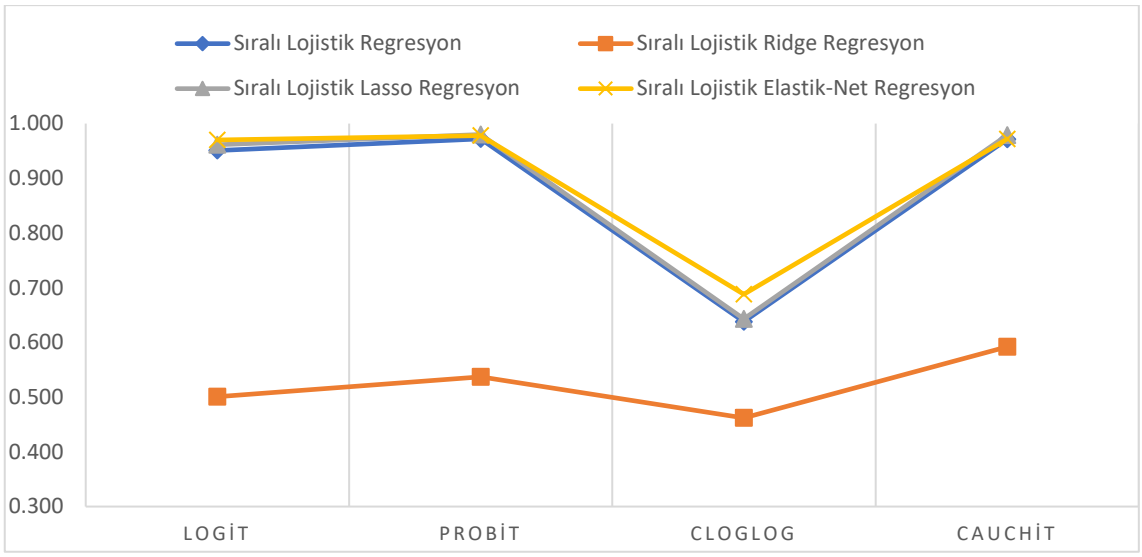
regresyon cloglog bağlantı fonksiyonunun % 85.9, sıralı lojistik elastik-net regresyon yönteminin cauchit bağlantı fonksiyonunun %96.6 ile en yüksek yüzdeye sahip olduğu görülmektedir. Simülasyon çalışmasında beklendiği gibi orta düzeyde çoklu bağlantı sorununun olduğu denemelerde doğru sınıflandırma performansları klasik modelin performansı görece daha yakın olsa da düzenlileştirme yöntemleri daha iyi sınıflandırma performansı vermiştir. 3. deneme kümesi için analiz sonuçlarına göre sıralı lojistik elastik-net yönteminin en iyi deneme, cauchit bağlantı fonksiyonunun ise en yüksek oranı verdiği görülmektedir.

Çizelge 5.4'te bağımsız değişken sayısı 7, çoklu bağlantı düzeyi orta, örneklem büyüklüğünün 4000 olduğu dengeli kategori dağılımına sahip denemeye ait sınıflandırma oranları farklı bağlantı fonksiyonları ve farklı sıralı lojistik modellerine göre verilmiştir.

Çizelge 5.4. 4. Deneme Kümesinin Doğru Sınıflandırma Oranları

Bağımsız Değişken Sayısı	**7**			
Çoklu Bağlantı Düzeyi	**0.6**			
Örneklem Büyüklüğü	**4.000**			
Dengeli Kategorili				
Bağlantı Fonksiyonu	**Sıralı Lojistik Regresyon**	**Sıralı Lojistik Ridge**	**Sıralı Lojistik Lasso**	**Sıralı Lojistik Elastik-Net**
Logit	0.951	0.501	0.961	0.970
Probit	0.972	0.537	**0.980***	0.978
Cloglog	0.638	0.462	0.643	0.688
Cauchit	0.971	0.592	0.979	0.972

**sınıflandırma matrisi ek 22'de yer almaktadır*



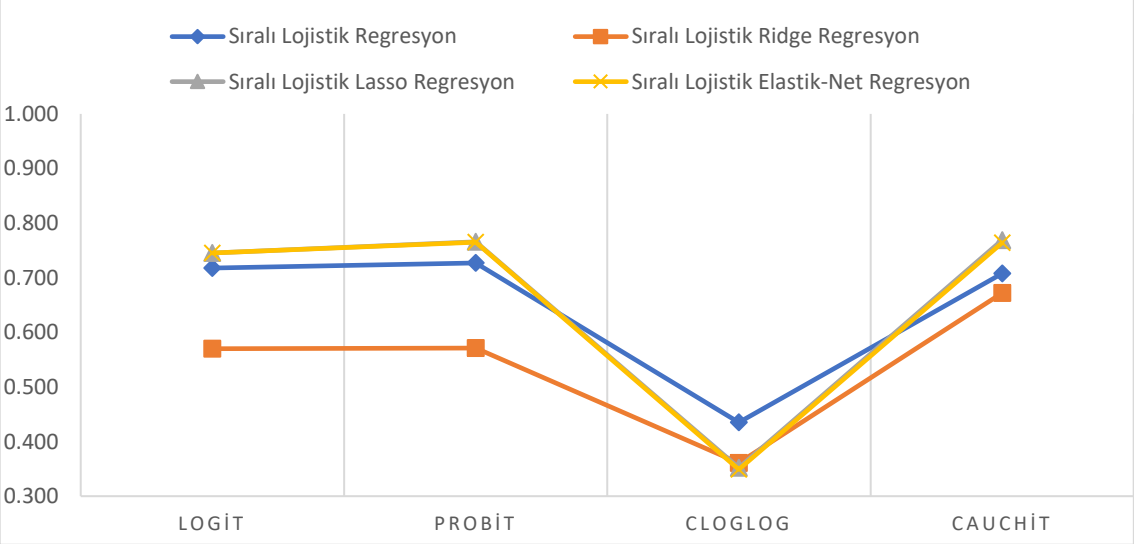
Şekil 5.4. 4. Deneme Kümesinin Doğru Sınıflandırma Oran Grafiği

Çizelge 5.4 ve Şekil 5.4 incelendiğinde, sıralı lojistik regresyon cloglog bağlantı fonksiyonu ile elde edilen denemenin analizinde doğru sınıflandırma oranı %63.8 iken, sıralı lojistik ridge regresyon cauchit bağlantı fonksiyonunun %59.2, sıralı lojistik elastik-net regresyon logit bağlantı fonksiyonunun % 97.0, sıralı lojistik lasso regresyon yönteminin probit bağlantı fonksiyonunun %98.0 ile en yüksek yüzdeye sahip olduğu görülmektedir. Simülasyon çalışmasında beklendiği gibi orta düzeyde çoklu bağlantı sorununun olduğu denemelerde doğru sınıflandırma performansları klasik modelin performansı görece daha yakın olsa da düzenlileştirme yöntemleri daha iyi sınıflandırma performansı vermiştir. 4. deneme kümesi için analiz sonuçlarına göre sıralı lojistik lasso yönteminin en iyi deneme, probit bağlantı fonksiyonunun ise en yüksek oranı verdiği görülmektedir.

Çizelge 5.5'te bağımsız değişken sayısı 3, çoklu bağlantı düzeyi zayıf, örneklem büyüklüğünün 16000 olduğu dengeli kategori dağılımına sahip denemeye ait sınıflandırma oranları farklı bağlantı fonksiyonları ve farklı sıralı lojistik modellerine göre verilmiştir.

Çizelge 5.5. 5. Deneme Kümesinin Doğru Sınıflandırma Oranları

Bağımsız Değişken Sayısı	**3**			
Çoklu Bağlantı Düzeyi	**0.3**			
Örneklem Büyüklüğü	**16.000**			
Dengeli Kategorili				
Bağlantı Fonksiyonu	**Sıralı Lojistik Regresyon**	**Sıralı Lojistik Ridge**	**Sıralı Lojistik Lasso**	**Sıralı Lojistik Elastik-Net**
Logit	0.718	0.570	0.745	0.745
Probit	0.727	0.571	0.766	0.765
Cloglog	0.435	0.361	0.352	0.349
Cauchit	0.708	0.672	**0.769**	0.764



Şekil 5.5. 5. Deneme Kümesinin Doğru Sınıflandırma Oran Grafiği

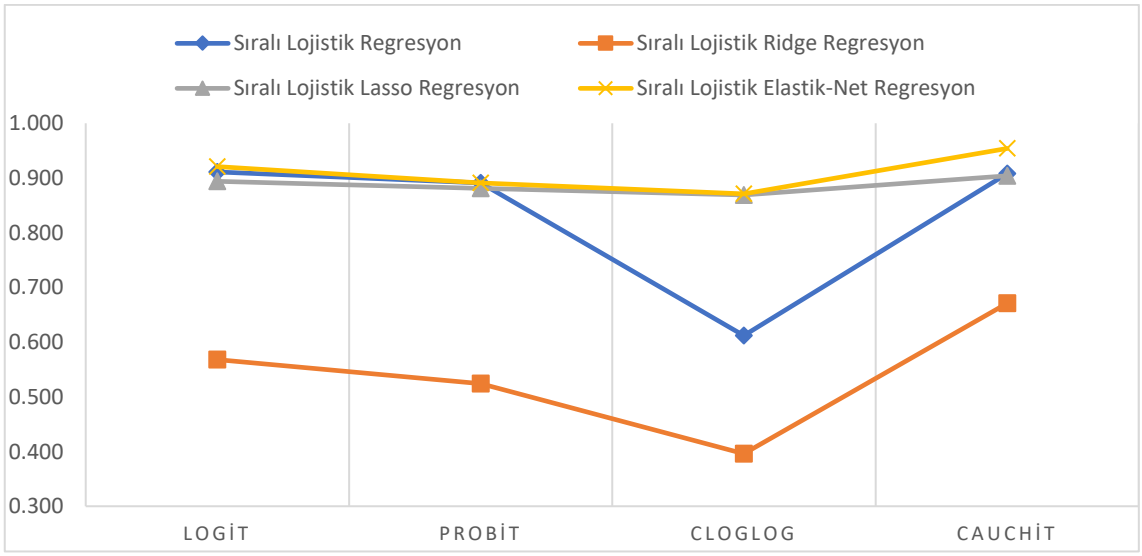
Çizelge 5.5 ve Şekil 5.5 incelendiğinde, sıralı lojistik regresyon cloglog bağlantı fonksiyonu ile elde edilen denemenin analizinde doğru sınıflandırma oranı %43.5 iken, sıralı lojistik ridge regresyon logit bağlantı fonksiyonunun %57.0, sıralı lojistik elastik-net regresyon probit bağlantı fonksiyonunun % 76.5, sıralı lojistik lasso regresyon yönteminin cauchit bağlantı fonksiyonunun %76.9 ile en yüksek yüzdeye sahip olduğu görülmektedir. Simülasyon çalışmasında beklendiği gibi zayıf düzeyde çoklu bağlantı

sorununun olduğu denemelerde doğru sınıflandırma performansları klasik modelin performansı daha yakın olsa da düzenlileştirme yöntemleri daha iyi sınıflandırma performansı vermiştir. 5. deneme kümesi için analiz sonuçlarına göre sıralı lojistik lasso yönteminin en iyi deneme, cauchit bağlantı fonksiyonunun ise en yüksek oranı verdiği görülmektedir.

Çizelge 5.6'de bağımsız değişken sayısı 7, çoklu bağlantı düzeyi yüksek, örneklem büyüklüğünün 400 olduğu dengeli kategori dağılımına sahip denemeye ait sınıflandırma oranları farklı bağlantı fonksiyonları ve farklı sıralı lojistik modellerine göre verilmiştir.

Çizelge 5.6. 6. Deneme Kümesinin Doğru Sınıflandırma Oranları

Bağımsız Değişken Sayısı	**7**			
Çoklu Bağlantı Düzeyi	**0.9**			
Örneklem Büyüklüğü	**400**			
Dengeli Kategorili				
Bağlantı Fonksiyonu	**Sıralı Lojistik Regresyon**	**Sıralı Lojistik Ridge**	**Sıralı Lojistik Lasso**	**Sıralı Lojistik Elastik-Net**
Logit	0.911	0.568	0.894	0.921
Probit	0.891	0.524	0.881	0.891
Cloglog	0.612	0.396	0.869	0.871
Cauchit	0.908	0.671	0.904	**0.954**



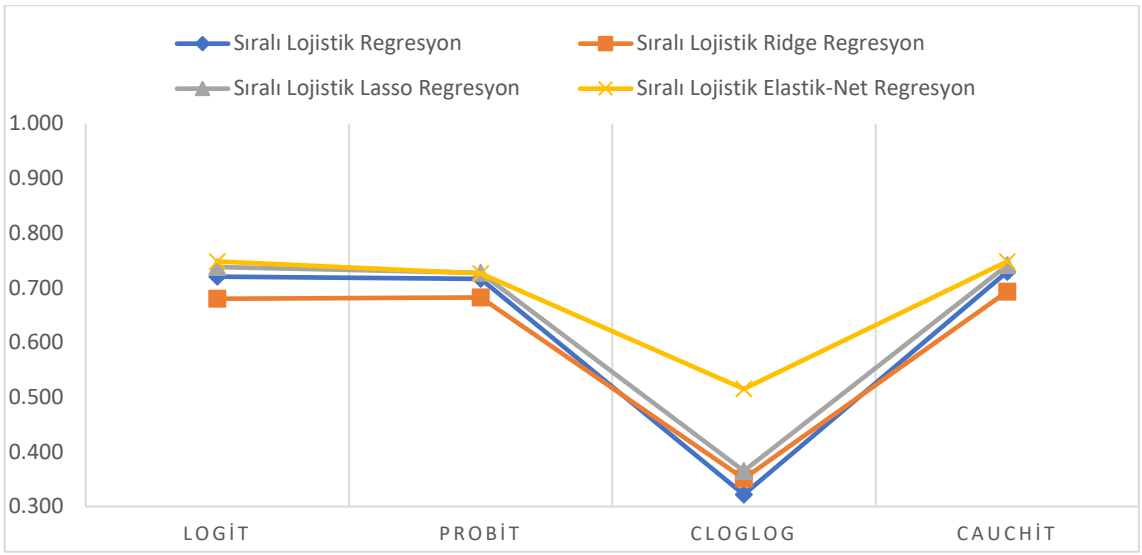
Şekil 5.6. 6. Deneme Kümesinin Doğru Sınıflandırma Oran Grafiği

Çizelge 5.6 ve Şekil 5.6 incelendiğinde, sıralı lojistik regresyon logit bağlantı fonksiyonu ile elde edilen denemenin analizinde doğru sınıflandırma oranı %91.1 iken, sıralı lojistik ridge regresyon probit bağlantı fonksiyonunun %52.4, sıralı lojistik lasso regresyon cloglog bağlantı fonksiyonunun % 86.9, sıralı lojistik elastik-net regresyon yönteminin cauchit bağlantı fonksiyonunun %95.4 ile en yüksek yüzdeye sahip olduğu görülmektedir. Simülasyon çalışmasında beklendiği gibi yüksek düzeyde çoklu bağlantı sorununun olduğu denemelerde düzenlileştirme yöntemleri daha iyi sınıflandırma performansı vermiştir. 6. deneme kümesi için analiz sonuçlarına göre sıralı lojistik elastik-net yönteminin en iyi deneme, cauchit bağlantı fonksiyonunun ise en yüksek oranı verdiği görülmektedir.

Çizelge 5.7'de bağımsız değişken sayısı 3, çoklu bağlantı düzeyi yüksek, örneklem büyüklüğünün 16000 olduğu dengeli kategori dağılımına sahip denemeye ait sınıflandırma oranları farklı bağlantı fonksiyonları ve farklı sıralı lojistik modellerine göre verilmiştir.

Çizelge 5.7. 7. Deneme Kümesinin Doğru Sınıflandırma Oranları

Bağımsız Değişken Sayısı	**3**			
Çoklu Bağlantı Düzeyi	**0.9**			
Örneklem Büyüklüğü	**16.000**			
Dengeli Kategorili				
Bağlantı Fonksiyonu	**Sıralı Lojistik Regresyon**	**Sıralı Lojistik Ridge**	**Sıralı Lojistik Lasso**	**Sıralı Lojistik Elastik-Net**
Logit	0.720	0.680	0.738	**0.748**
Probit	0.716	0.682	0.727	0.726
Cloglog	0.322	0.351	0.365	0.715
Cauchit	0.729	0.693	0.741	**0.748**



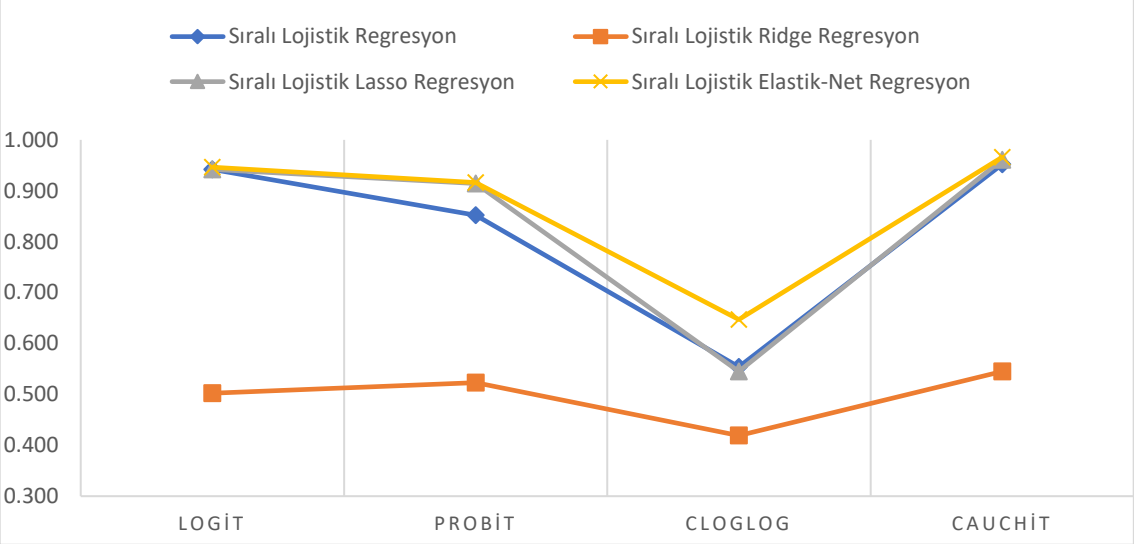
Şekil 5.7. 7. Deneme Kümesinin Doğru Sınıflandırma Oran Grafiği

Çizelge 5.7 ve Şekil 5.7 incelendiğinde, sıralı lojistik regresyon probit bağlantı fonksiyonu ile oluşturulan denemenin analizinde doğru sınıflandırma oranı %71.6 iken, sıralı lojistik ridge regresyon cloglog bağlantı fonksiyonunun %35.1, sıralı lojistik lasso regresyon cauchit bağlantı fonksiyonunun %74.1, sıralı lojistik elastik-net regresyon yönteminin logit bağlantı fonksiyonunun %74.8 ile en yüksek yüzdeye sahip olduğu görülmektedir. Simülasyon çalışmasında beklendiği gibi yüksek düzeyde çoklu bağlantı sorununun olduğu denemelerde düzenlileştirme yöntemleri daha iyi sınıflandırma performansı vermiştir. 7. deneme kümesi için analiz sonuçlarına göre sıralı lojistik elastik-net yönteminin en iyi deneme, logit ve cauchit bağlantı fonksiyonlarının ise en yüksek oranı verdiği görülmektedir.

Çizelge 5.8'de bağımsız değişken sayısı 5, çoklu bağlantı düzeyi orta, örneklem büyüklüğünün 4000 olduğu dengeli kategori dağılımına sahip denemeye ait sınıflandırma oranları farklı bağlantı fonksiyonları ve farklı sıralı lojistik modellerine göre verilmiştir.

Çizelge 5.8. 8. Deneme Kümesinin Doğru Sınıflandırma Oranları

Bağımsız Değişken Sayısı	**5**			
Çoklu Bağlantı Düzeyi	**0.6**			
Örneklem Büyüklüğü	**4.000**			
Dengeli Kategorili				
Bağlantı Fonksiyonu	**Sıralı Lojistik Regresyon**	**Sıralı Lojistik Ridge**	**Sıralı Lojistik Lasso**	**Sıralı Lojistik Elastik-Net**
Logit	0.942	0.502	0.942	0.946
Probit	0.852	0.523	0.914	0.916
Cloglog	0.554	0.419	0.545	0.647
Cauchit	0.952	0.545	0.961	**0.966**



Şekil 5.8. 8. Deneme Kümesinin Doğru Sınıflandırma Oran Grafiği

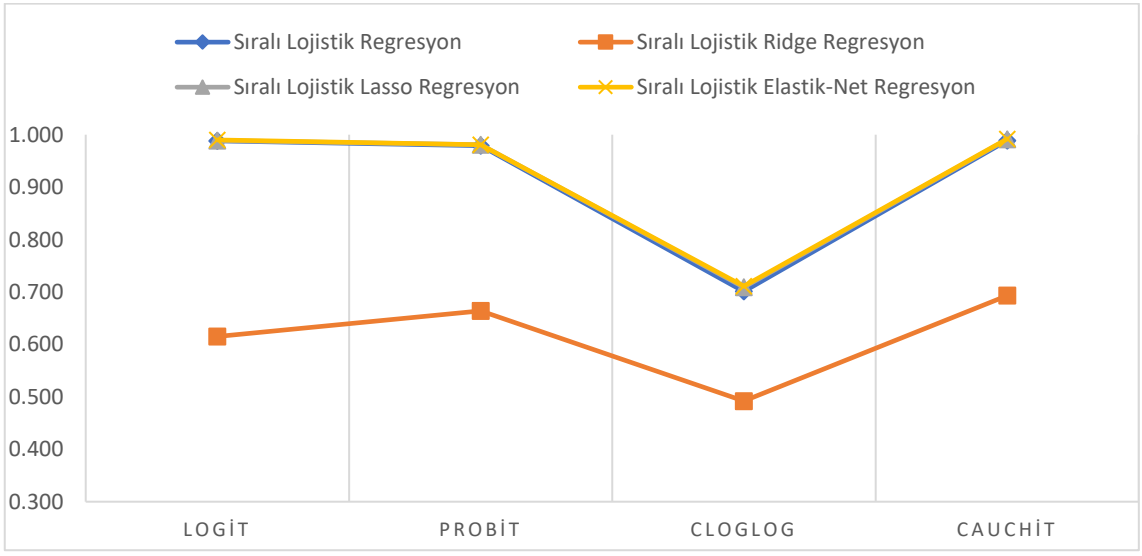
Çizelge 5.8 ve Şekil 5.8 incelendiğinde, sıralı lojistik regresyon cloglog bağlantı fonksiyonu ile elde edilen denemenin analizinde doğru sınıflandırma oranı %55.4 iken, sıralı lojistik ridge regresyon logit bağlantı fonksiyonunun %50.2, sıralı lojistik lasso regresyon probit bağlantı fonksiyonunun %91.4, sıralı lojistik elastik-net regresyon yönteminin cauchit bağlantı fonksiyonunun %96.6 ile en yüksek yüzdeye sahip olduğu görülmektedir. Simülasyon çalışmasında beklendiği gibi orta düzeyde çoklu bağlantı

sorununun olduğu denemelerde düzenlileştirme yöntemleri daha iyi sınıflandırma performansı vermiştir. 8. deneme kümesi için analiz sonuçlarına göre sıralı lojistik elastik-net yönteminin en iyi deneme, cauchit bağlantı fonksiyonunun ise en yüksek oranı verdiği görülmektedir.

Çizelge 5.9'de bağımsız değişken sayısı 7, çoklu bağlantı düzeyi zayıf, örneklem büyüklüğünün 400 olduğu dengeli kategori dağılımına sahip denemeye ait sınıflandırma oranları farklı bağlantı fonksiyonları ve farklı sıralı lojistik modellerine göre verilmiştir.

Çizelge 5.9. 9. Deneme Kümesinin Doğru Sınıflandırma Oranları

Bağımsız Değişken Sayısı	**7**			
Çoklu Bağlantı Düzeyi	**0.3**			
Örneklem Büyüklüğü	**400**			
Dengeli Kategorili				
Bağlantı Fonksiyonu	**Sıralı Lojistik Regresyon**	**Sıralı Lojistik Ridge**	**Sıralı Lojistik Lasso**	**Sıralı Lojistik Elastik-Net**
Logit	0.988	0.615	0.989	0.990
Probit	0.979	0.664	0.981	0.981
Cloglog	0.701	0.492	0.709	0.711
Cauchit	0.989	0.693	0.991	**0.992**



Şekil 5.9. 9. Deneme Kümesinin Doğru Sınıflandırma Oran Grafiği

Çizelge 5.9 ve Şekil 5.9 incelendiğinde, sıralı lojistik regresyon probit bağlantı fonksiyonu ile elde edilen denemenin analizinde doğru sınıflandırma oranı %97.9 iken, sıralı lojistik ridge regresyon logit bağlantı fonksiyonunun %61.5, sıralı lojistik lasso regresyon cloglog bağlantı fonksiyonunun %70.9, sıralı lojistik elastik-net regresyon yönteminin cauchit bağlantı fonksiyonunun %99.2 ile en yüksek yüzdeye sahip olduğu görülmektedir. Simülasyon çalışmasında beklendiği gibi zayıf düzeyde çoklu bağlantı sorununun olduğu denemelerde düzenlileştirme yöntemleri daha iyi sınıflandırma performansı vermiştir. 9. deneme kümesi için analiz sonuçlarına göre sıralı lojistik elastik-net yönteminin en iyi deneme, cauchit bağlantı fonksiyonunun ise en yüksek oranı verdiği görülmektedir.

Tüm analiz sonuçları doğruluk, yani doğru sınıflandırma oranı üzerinden incelenmiştir, bu değerlere ek olarak tüm denemelere ait duyarlılık (recall), kesinlik (precision), F1-Score, Cohen'in Kappa ve AUC değerleri incelenerek de modellerin karşılaştırılması yapılabilir. Doğruluk (doğru sınıflandırma oranı) değerleri birçok çalışmada modellerin sonuç çıktısı için tercih edilmektedir fakat modelleri birbirlerinden ayırmada kesinlik ve duyarlılık da anlamlı performans ölçütleri olarak bilinmektedir. Ayrıca bu iki ölçütün birlikte değerlendirildiği F1-Score, Cohen'in Kappa ve AUC değerleri de veri setlerinin analiz sonuçlarında tercih edilmektedir. Bu çalışmada sadece doğru sınıflandırma oranları sunulmuş olsa da 4 deneme için sınıflandırma matrisi örnekleri Ekler'de verilmiştir.

5.2 Gerçek Bir Veri Seti ile Analiz

Simülasyon çalışmasında üretilen veriler ile elde edilen sonuçlar, kaggle.com isimli internet sitesinin veri tabanında paylaşılan "*Regression on Diamonds Dataset"* (Kaggle, 2022) isimli gerçek veri seti üzerinde uygulanarak desteklenmiştir. Veri seti 53940 gözlem ve 11 adet değişken içermekte olup ünlü bir mücevher markasının ürün özelliklerinin olduğu fiyat listesinden oluşmuştur.

Veri setinde elmasların dolar para biriminden fiyat değerleri, sıralı lojistik regresyon analizine uygun şekilde 4 kategorili bağımlı değişkene dönüştürülerek kullanılmıştır. Fiyat değişkeni 1242 dolar altındaki 18002 adet ürün "1" uygun fiyatlı elmas olarak, 1242 ve 3932 dolar aralığındaki 16282 adet ürün "2" orta fiyatlı elmas olarak, 3933 ve 6623 dolar altındaki 9633 adet ürün "3" yüksek fiyatlı elmas olarak, 6623 dolar üzerindeki 10024 adet ürün "4" çok yüksek fiyatlı elmas olarak kategorize edilmiştir. Bağımsız değişkenler sürekli degişken olup elmasların karat ağırlığı, üst tabla genişliği, derinlik toplam yüzdesi, x uzunluk mm, y genişlik mm, z derinlik mm değişkenlerinden oluşmaktadır.

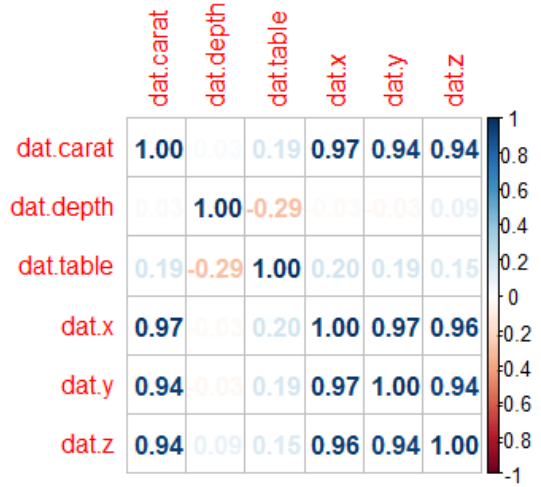
Bu değişkenlere ait özellikler, veri tipleri ve değer aralıkları kategorik değişkenler için Çizelge 5.10'de verilmiştir.

Çizelge 5.10. Diamonds Veri Seti Değişken Özellikleri

Attributes	Değişkenler	Veri Tipi ve Değerler
Price	Fiyat	Sıralı, Kategorik [1, 2, 3, 4]
Carat	Karat Ağırlığı	Sürekli [0.20, …, 5.01]
Table	Üst Tabla Genişliği	Sürekli [43, …, 95]
Depth Total Percentage	Derinlik Toplam Yüzdesi	Sürekli [43, …, 79]
X Length	X Uzunluk	Sürekli [0.01, …, 10.74]
Y Width	Y Genişlik	Sürekli [0.01, …, 58.90]
Z Depth	Z Derinlik	Sürekli [0.01, …, 31.80]

Eğitim veri seti raslantısal olarak %60 oranında ve test veri seti için %40 oranında gözlem sayıları ayrılmıştır. Sonuçta 32466 gözlem değeri eğitim setinde, 21474 gözlem değeri test setinde yer almaktadır.

Bağımsız değişkenler arasında ilişki Pearson'ın korelasyon katsayıları ile incelenmiştir, Bu çalışmada çoklu bağlantı durumunda sıralı lojistik regresyon modellerinin doğru sınıflandırma oranları karşılaştırıldığı için bağımsız değişkenler arasında belirli oranlarda çoklu bağlantı olması tercih edilmiştir. Bu tercihe göre seçilen sürekli yapıdaki değişkenler arasındaki ilişki dikkate alınmıştır. Bağımsız değişkenlere ilişkin korelasyon matrisi Şekil 5.10'da verilmiştir.



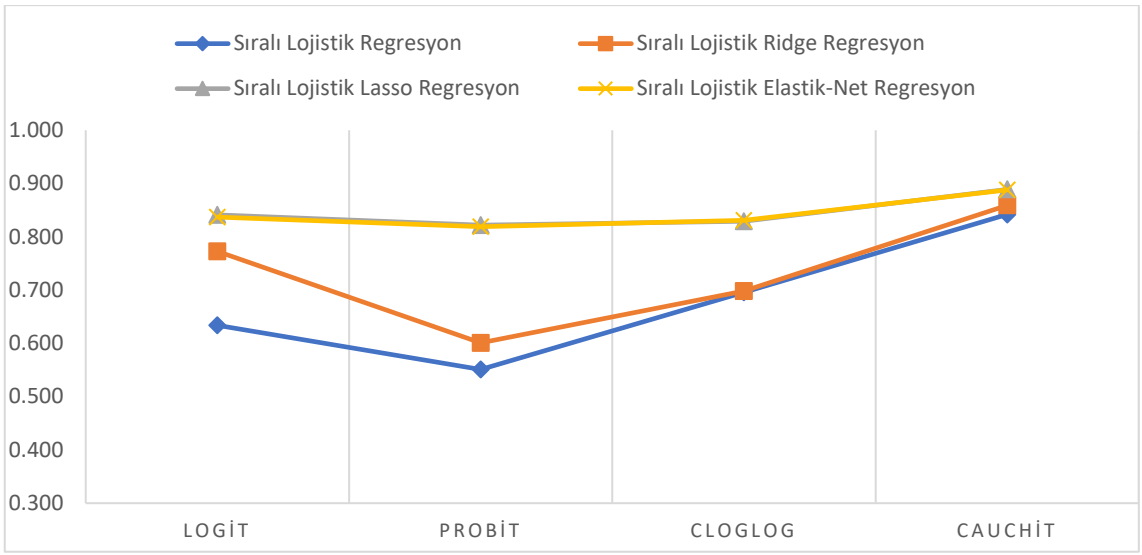
Şekil 5.10. Diamonds Veri Seti Bağımsız Değişkenler İçin Korelasyon Matrisi

Şekil 5.10'daki gösterilen matris incelendiğinde, bağımsız değişkenlerin bir diğeriyle arasındaki korelasyon katsayısının eşik değeri olarak belirlenen 0.6'dan yüksek olduğu durumlar görülmektedir. Bu durumda analizde kullanılacak olan veri setimizde çoklu bağlantı sorunu olduğu için sıralı lojistik regresyon modeli ile düzenlileştirme modellerini karşılaştıracağımız analize uygun bir şekilde seçim yapılmıştır.

Bağımlı değişken olan *Fiyat* sıralı kategorik değişken olarak seçildiği için uygulamaya uygun şekilde öncelikle sıralı lojistik regresyon yöntemi uygulanmıştır. Bunun için R Studio'da kullanılan VGAM paketi ve fonksiyonda vglm, paralel doğrular varsayımı altında, model çıktı ailesi "cumulative" ve bağlantı fonksiyonları "logit", "probit", "cloglog", "cauchit" olarak belirlenmiş ve bu kodlar ile sıralı lojistik regresyon modeli incelenmiştir. Daha sonra karşılaştırılacak olan düzenlileştirme yöntemleri için R Studio'da kullanılan ordinalNet paketi ve fonksiyonda ordinalNet, paralel doğrular varsayımı altında, model çıktı ailesi "cumulative" ve bağlantı fonksiyonları "logit", "probit", "cloglog", "cauchit" olarak belirlenmiştir, bu kodlar ile düzenlileştirme modelleri incelenmiştir. Model önce eğitim seti üzerinde çalıştırılmış ve ilgili modelin test seti üzerinde çalıştırılmış doğru sınıflandırma performansları özeti Çizelge 5.11 ve Şekil 5.11'de gösterilmiştir.

Çizelge 5.11. Diamonds Veri Seti Doğru Sınıflandırma Oranları

Bağlantı Fonksiyonu	Sıralı Lojistik Regresyon	Sıralı Lojistik Ridge	Sıralı Lojistik Lasso	Sıralı Lojistik Elastik-Net
Logit	0.634	0.773	**0.841**	0.837
Probit	0.551	0.601	**0.822**	0.819
Cloglog	0.696	0.684	0.829	**0.831**
Cauchit	0.842	**0.859**	**0.889**	0.888



Şekil 5.11. Diamonds Veri Seti Doğru Sınıflandırma Oran Grafiği

Kriterlere göre sıralı lojistik lasso regresyon ve sıralı lojistik elastik-net regresyon, klasik yöntem olan sıralı lojistik regresyon modeline göre daha iyi sınıflandırma performansı göstermiştir. Sıralı lojistik ridge regresyon yöntemi de görece daha yüksek bir doğru sınıflandırma oranı vermektedir, özellikle lojistik lasso regresyon ve sıralı lojistik elastik-net regresyon yöntemleri "cauchit" bağlantı fonksiyonu ile sırasıyla %88.9 ve %88.8 doğru sınıflandırma oranlarıyla oldukça tatmin edici sonuçlar vermektedir. Bu iki alternatif regresyon yönteminin de daha iyi performans göstermeleri ve önemi az olan değişkenleri modelden eleyerek çıkartma özelliklerine sahip olması, regresyon analizinde kurulacak olan modellerdeki tüm değişkenlerin anlamlı olup olmadığı ve aralarındaki ilişki düzeylerinin incelenmesi gerektiği yorumu çıkarılabilir.

Sonuç olarak, alternatif yöntemler olan düzenlileştirme modelleri sıralı lojistik lasso regresyon, sıralı lojistik elastik-net regresyon ve sıralı lojistik ridge regresyon ile sıralı lojistik regresyona göre daha iyi doğru sınıflandırma oranları yakalanabileceği anlaşılmaktadır. Bu bağlamda gerçek veri setiyle yapılan bu örnek çalışmada benzer bir sonuca ulaşılmıştır.

5.3 Tartışma ve Öneriler

Bu çalışmada, bağımsız değişkenler arasında yüksek, orta ve düşük korelasyonların olduğu çoklu bağlantı durumundaki senaryolarda, sıralı lojistik regresyon ile sıralı lojistik ridge, sıralı lojistik lasso ve sıralı lojistik elastik-net regresyonun doğru sınıflandırma performanslarının karşılaştırılması ve en iyi sınıflama performansı gösteren yöntemin

tespit edilmesi amaçlanmıştır. Bu amaç doğrultusunda, simülasyon çalışması ile belirtilen tüm senaryolar için çeşitli denemeler oluşturulmuş, analiz edilmiş, sonuçları verilmiş ve yorumlanmıştır.

Veri bilimindeki gelişmeler, farklı senaryolar içeren durumlarda regresyon modelleri ile yapılan tahminlerin güvenilirliğin ve doğruluğunun artırılmasının gerekli olduğuna işaret etmektedir. Bağımsız değişkenler arasındaki çoklu bağlantı sorunu, regresyon yönteminin tahmin doğruluğunu doğrudan etkilemektedir. Çoklu bağlantı sorununun çözümüne yönelik yanlı ve boyut indirgeyen alternatif, düzenlileştirme yöntemleri uygulanmaktadır. Sıralı lojistik ridge regresyon ele alınan tüm değişkenlerin katsayılarını tahmin etmeye çalışan yanlı bir regresyon tekniğidir. Sıralı lojistik lasso regresyon ve sıralı lojistik elastik-net regresyon ise aynı anda hem boyut indirgeyen hem de model parametrelerini tahmin eden tekniklerdir. Son yıllarda yaygın olarak kullanılan bu tekniklerin bu tezdeki simülasyon çalışması ile klasik yönteme göre doğru sınıflandırma güçlerinin karşılaştırılması sağlanmıştır.

Bu çalışmada yapılan uygulamalar sonucunda, simülasyon çalışmasının sonuç çıktıları incelendiğinde, doğru sınıflandırma çalışmalarında, sıralı lojistik regresyona göre sıralı lojistik elastik-net regresyon ve sıralı lojistik lasso regresyon yöntemleri ile daha yüksek ve doğru sınıflama oranları elde edilebileceği görülmektedir. Çoklu bağlantıdan etkilenmeden, doğrusal ve lojistik modeller için tahminlemede etkili bir performansa sahip olan sıralı lojistik elastik-net regresyon yönteminin çok boyutlu veri setlerinin analizi için uygun ve önerilebilecek bir teknik olduğu görülmektedir. Tüm denemeler için elde edilen doğru sınıflandırma oranları incelendiğinde, bağımsız değişken sayısının (3, 5, 7) artması diğer tüm durumlar sabit olduğunda, genel olarak daha yüksek bir doğru sınıflandırma oranı bulunmasını sağlamıştır. Ayrıca çalışmada tüm denemeler için elde edilen doğru sınıflandırma oranları incelendiğinde, örneklem büyüklüğünün (400, 4000, 16000) artması diğer tüm durumlar sabit olduğunda, genel olarak daha yüksek bir doğru sınıflandırma oranı bulunmasını sağlamıştır. Tüm denemeler için elde edilen doğru sınıflandırma oranları incelendiğinde, çoklu bağlantı düzeyinin (zayıf, orta, yüksek) artması diğer tüm durumlar sabit olduğunda, düzenlileştirme teknikleri olan sıralı lojistik lasso regresyon ve sıralı elastik-net regresyon yöntemleri ile daha yüksek bir doğru sınıflandırma oranı bulunmasını sağlamıştır, çoklu bağlantı düzeyi azaldıkça ise sıralı lojistik regresyon ile düzenlileştirme yöntemleri arasındaki doğru sınıflandırma oran farklarının azaldığı gözlemlenmektedir. Bu çalışmada tüm denemeler için elde edilen doğru sınıflandırma oranları incelendiğinde, dengeli kategori dağılımına sahip denemeler

dengesiz kategori dağılımına sahip denemelere göre diğer tüm durumlar sabit olduğunda, genel olarak daha yüksek bir doğru sınıflandırma oranı bulunmasını sağlamıştır. Ayrıca bu çalışmada tüm denemeler için elde edilen doğru sınıflandırma oranları incelendiğinde, bağlantı fonksiyonları (logit, probit, cloglog, cauchit) açısından cloglog bağlantı fonksiyonu en düşük doğru sınıflandırma gücüne sahipken, cauchit ve probit bağlantı fonksiyonları en yüksek doğru sınıflandırma gücüne sahip oldukları gözlemlenmiştir.

İstatistik literatüründe gözlenen birimleri özelliklerine göre tahmin ederek sınıflara ayırmak, elde edilecek bilgilerin özetlenmesi ve yorumlanması açısından çok önemli bir yere sahiptir. Sonuç olarak, istatistiksel düzenlileştirme yöntemleri olan ridge regresyon, lasso regresyon ve elastik-net regresyon yöntemlerinin sıralı lojistik modellerin sınıflandırma çalışmalarında doğru ve güvenilir yöntemler olarak kullanılabileceği görülmektedir. Sıralı lojistik modellerin uygulanması öncesinde mutlaka veri setinin ve tüm değişkenlerin yapısı incelenmeli, çalışmanın amacına uygun olabilecek tüm alternatif yöntemler denenmeli, en yüksek anlamlılık ve doğru sınıflandırma oranının elde edileceği modelin seçilmesi tavsiye edilebilir. Bu tez çalışmasının amaçlarına benzer şekilde, sıralı lojistik regresyon modelinin birçok farklı simülasyon çalışması ve gerçek veri setleri ile detaylı bir biçimde incelenmesi, ileride gerçekleştirilecek çalışmalara yol ve yön göstermesi açısından önemli bir katkı sağlayacaktır. Özellikle bağımlı değişkenin kategori sayısının çok fazla olduğu veri setlerindeki durumlara ilişkin veya bu çalışmanın kapsamına benzer şekilde bağımlı değişkenin kategori sayılarının denemeler boyunca değiştiği, veri setlerindeki etkileşimlerinin de göz önünde bulundurularak incelenebileceği yeni çalışmaların gerçekleştirilmesi önerilmektedir.

6. KAYNAKLAR

Agresti, A. (2019). An Introduction to Categorical Data Analysis, Third Edition. Wiley Series in Probability and Statistics, printed in USA.

Akın, H. B. ve Şentürk, E. (2012). Bireylerin Mutluluk Düzeylerinin Ordinal Lojistik Regresyon Analizi ile İncelenmesi. Öneri Dergisi, 10 (37), 183-193.

Aktar Demirtaş, E., Anagün, A. S., Köksal, G. (2009). Determination of Optimal Product Styles by Ordinal Logistic Regression Versus Conjoint Analysis for Kitchen Faucets. International Journal of Industrial Ergonomics, 39 (5), 866-875.

Algamal, Z. ve Lee, M. H. (2015). High Dimensional Logistic Regression Model using Adjusted Elastic Net Penalty. Pakistan Journal of Statistics and Operation Research, 11, 667-676.

Allison, P. (2012). When Can You Safely Ignore Multicollinearity? Statistical Horizons: Erişim Adresi (04.06.2022): (http://www.statisticalhorizons.com/multicollinearity).

Alpar, R. (2018). Uygulamalı Çok Değişkenli İstatistiksel Yöntemler. Ankara: Detay Yayıncılık.

Ananth, C. V. ve Kleinbaum, D. G. (1997). Regression Models for Ordinal Responses: A Review of Methods and Applications. International Journal of Epidemiology, 26 (6), 1323-1333.

Anderson, J. A. ve Philips, P. R. (1981). Regression, Discrimination and Measurement Models for Ordered Categorical Variables. Journal of the Royal Statistical Society. Series C (Applied Statistics), 30 (1), 22-31.

Arı, E. (2013). Sıralı Lojistik Regresyonda Paralel Doğrular Varsayımı ve Çözümleme Yaklaşımları. Doktora Tezi. Eskişehir: Eskişehir Osmangazi Üniversitesi Fen Bilimleri Enstitüsü.

Atabey, Ö. (2010). Lojistik regresyon modeli ve geriye doğru eliminasyon yöntemiyle değişken seçiminin hipertansiyon riski üzerine uygulamasında Bootstrap yöntemi. Yüksek Lisans tezi, GÜ, 122.

Ayhan, S. (2006). Sıralı Lojistik Regresyon Analiziyle Türkiye'deki Hemşirelerin İş Bırakma Niyetini Etkileyen Faktörlerin Belirlenmesi. Yüksek Lisans Tezi. Eskişehir: Osmangazi Üniversitesi Fen Bilimleri Enstitüsü.

Bircan, H. (2004). Lojistik Regresyon Analizi: Tıp Verileri Üzerine Bir Uygulama. Kocaeli Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 2, 185-208.

Breslaw, J.A. ve McIntosh, J. (1998). Simulated latent variable estimation of models with ordered categorical data. Journal of Econometrics, 87, 25- 47.

Chen C.K. ve Hughes J. (2004). Using ordinal regression model to analyze student satisfaction questionnaires. Association for Institutional Research, Volume1.

Chen, J., Hoogh, K., Gulliver, J., Hoffmann, B., Hertel, O., Ketzel, M., Bauwelinck, M., Donkelaar, A., Hvidtfeldt, U., Katsouyanni, K., Janssen, N., Martin, R., Samoli, E., Schwartz, P., Stafoggia, M., Bellander, T., Strak, M., Wolf, K., Vienneau, D., Vermeulen, R., Brunekreef, B. ve Hoek, G. (2019). A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide. Environment International, 130.

Cramer, J. (2002). The Origins of Logistic Regression. Tinbergen Institute Discussion Papers, 119, 4.

Çokluk, Ö. (2010). Lojistik Regresyon Analizi: Kavram ve Uygulama. Kuram ve Uygulamada Eğitim Bilimleri 10 (3), 1357-1407.

Deniz Başar, Ö. (2012). Öğretim Üyelerinin Unvanları ile Öğrencilerin Öğretim Üyelerini Değerlendirmeleri Arasındaki İlişkinin Sıralı Lojistik Regresyon Analizi ile İncelenmesi. Trakya Üniversitesi Sosyal Bilimler Dergisi, 14 (1), 121- 136.

Deniz E., Akbilgic O. ve Howe J. A. (2011). Model selection using information criteria under a new estimation method: least squares ratio. Journal of Applied Statistics, 38:9, 2043-2050

Duffy, D. E. ve Santner, T. J. (1989). On the Small Sample Properties of Norm-Restricted Maximum Likelihood Estimators for Logistic Regression Models. Communs Statist, Theory Meth., 18, 959-980.

Emeç, H. (2002). Ege Bölgesi Tüketim Harcamaları için Sıralı Logit Tahminleri ve Senaryo Sonuçları. Dokuz Eylül Üniversitesi Sosyal Bilimler Enstitüsü Dergisi 4(2), 13-29.

Erar, A. (2013). Doğrusal Regresyon Analizine Giriş. Nobel Yayınları.

Fırat, M. ve Onay, A. (1999). Bitki Doku Kültürü Çalışmalarından Elde Edilen Binom

Verilerinin Genelleştirilmiş Lineer Modeller Kullanılarak Analizi. Tr. J. Of Biology, 23 (1999), 261-237.

Fujimoto, K. (2003). Application of Multinomial and Ordinal Regressions to data of Japanese female labor market. M.S. thesis. University of Pittsburgh.

Fullerton, A. S. ve Xu, J. (2012). The Proportional Odds with Partial Proportionality Constraints Model for Ordinal Response Variables. Social Science Research, 41(1), 182-198.

Greenland, S. (1994). Alternative Models for Ordinal Logistic Regression. Statistics in Medicine, 13 (16), 1665-1677.

Harrell, F. (2015). Regression Modeling Strategies (Second Edition). Springer.

Hastie, T., Tibshirani, R. ve Wainwright, M. (2015). Statistical Learning with Sparsity. Taylor & Francis Group, LLC, p.335.

Hoerl, A.E. ve Kennard, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12 (1), 55–67.

Ishwaran, H. ve Gatsonis, C.A. (2000). A general class of hierarchical ordinal regression models with applications to correlated ROC analysis. The Canadian Journal of Statistics, 28.

James, G., Witten, D., Hastie, T. ve Tibshirani, R. (2014). An Introduction to Statistical Learning: with Applications in R. Springer.

Kaggle (2022). Regression on Diamonds Dataset. https://www.kaggle.com/code/heeraldedhia/regression-on-diamonds-dataset-95 score/data

Kleinbaum, D.G. ve Klein M. (2010). Logistic Regression. A Self- Learning Text, Third Edition, Springer, 590.

Long, J. S. (1997). Regression Models for Categorical and Limited Dependent Variables (Vol. 7). Advanced Quantitative Techniques in The Social Sciences. Sage Publications.

Long, J. S. ve Freese, J. (2014). Regression Models for Categorical Dependent Variables Using Stata, Second Edition (Vol. 3.). Texas: Taylor & Francis.

McCullagh, P. (1980). Regression models for ordinal data. Journal of the Royal Statistical Society, 42, 109–142.

Melkumova, L.E. ve Shatskikh, S. Ya. (2017). Comparing Ridge And Lasso Estimators For Data Analysis. Procedia Engineering, 201,746-755.

Menard, S. (2001). Applied Logistic Regresssion Analysis Second Edition. London: Sage Publications.

Montgomery, D., Peck, E. ve Vining, G. (2012). Introduction to Linear Regression Analysis. New Jersey: John Wiley & Sons, Inc.

Nizam, D. K. ve Akdeniz, H. A. (2007). Türkiye'de Yataklı Tedavi Kurumlarının Kapasite Kullanım Oranlarının Sıralı Logistik Regreasyon Analizi. Dokuz Eylül Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 9 (4), 1-14.

O'Connell, A. A. (2006). Logistic Regression Models for Ordinal Response Variables, Quatitative Applications in the Social Sciences (Vol. 146). Sage Publications.

Özdinç, Ö. (1999). Derecelendirme sürecinde ekonometrik bir değerlendirme. Sermaye Piyasası Kurulu.

Peng, C. Y. J., Lee, K. L., Ingersoll, G. M. (2002). An Introduction to Logistic Regression Analysis and Reporting. The Journal of Educational Research, 96 (1), 3-14.

Saleh, E., Arashi, M. ve Kibria, G. (2019). Theory of Ridge Regression Estimation with Applications. John Wiley & Sons, Inc.

Shrestha, N. (2020). Detecting Multicollinearity in Regression Analysis. American Journal of Applied Mathematics and Statistics, 8, 39-42.

Sümbüloğlu, K. ve Akdağ, B. (2007). Regresyon Yöntemleri ve Korelasyon Analizi. Ankara: Hatiboğlu Yayınevi, 139.

Şahin, O. (2017). Lojistik Regresyon Yöntemi ile Ayvalığı Turizm Amaçlı Tercih Etmede Önemli Risk Faktörlerinin Belirlenmesi. Elektronik Sosyal Bilimler Dergisi, 16 (61), 647-660.

Şerbetçi, A. (2012). Sıralı Lojistik Regresyon Analizi ile İstatistik ve Ekonometri Derslerinde Başarıyı Etkileyen Faktörlerin Belirlenmesi: Atatürk Üniversitesi İktisadi ve İdari Bilimler Fakültesi Öğrecileri Üzerine Bir Uygulama. Yüksek Lisans Tezi. Erzurum Atatürk Üniversitesi Sosyal Bilimler Enstitüsü İşletme Anabilimdalı.

Tansel, A. ve Güngör, N. D. (2004). Türkiye'den Yurt Dışına Beyin Göçü: Ampirik Bir Uygulama. ERC Working Paper in Economic 4, 2.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society, 267-288.

Tutz, G., Hennevogl, W. (1996). Random Effects in Ordinal Regression Models. Computational Statistics & Data Analysis, 22 (5), 537-557.

Walker, S. ve Duncan, D. (1967). Estimation of the probability of an event as a function of several independent variables. Biometrika, 54, 167–179.

Yavuz, S., Deveci, M., Karabulut, T., Şentürk, E. (2014). Sıralı Lojistik Regresyon Analiziyle Üniversite Öğrencilerinin Kent Memnuniyetini Etkileyen Faktörlerin Belirlenmesi: Erzincan Üniversitesi Örneği. Cumhuriyet Üniversitesi İktisadi ve İdari Bilimler Dergisi, 15 (1), 95-114.

Young, D. (2017). Handbook of Regression Methods. CRC Press.

Zortuk, M., Koç, E., Bayrak, S. (2014). Hane Halkları Satın Alma Kriterlerinin Analizi: Multinominal Lojistik Regresyon Yaklaşımı. Dumlupınar Üniversitesi Sosyal Bilimler Dergisi, 163-176.

Zou, H. ve Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology), 301-320.

Zhou, F., Wu, D., Yang, X., Jiao, J. (2008). Ordinal Logistic Regression for Affective Product Design. Paper presented at the Industrial Engineering and Engineering Management, 2008. IEEM 2008.

7. EKLER

Ek 1

Bağımlı Değişken ve Düzeyi Y; 1-2-3-4; Dengeli Kategori Dağılımı; %25 1-2-3-4																	
Çoklu Bağlantı Düzeyi 0.9	İterasyon 100	Sıralı Lojistik Regresyon				Ridge Düzenlemeli				Lasso Düzenlemeli				Elastik Net Düzenlemeli			
Bağımsız Değişken Sayısı	Örneklem Büyüklüğü	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit
3	16000	0.710	0.706	0.302	0.709	0.670	0.671	0.331	0.683	0.708	0.707	0.335	0.711	0.708	0.706	0.700	0.710
5	16000	0.937	0.920	0.601	0.938	0.690	0.671	0.340	0.705	0.957	0.918	0.605	0.961	0.960	0.919	0.711	0.970
7	16000	0.958	0.948	0.605	0.958	0.691	0.674	0.340	0.709	0.959	0.949	0.608	0.968	0.965	0.951	0.714	0.979

Ek 2

Bağımlı Değişken ve Düzeyi Y; 1-2-3-4; Dengeli Kategori Dağılımı; %25 1-2-3-4																	
Çoklu Bağlantı Düzeyi 0.6	İterasyon 100	Sıralı Lojistik Regresyon				Ridge Düzenlemeli				Lasso Düzenlemeli				Elastik Net Düzenlemeli			
Bağımsız Değişken Sayısı	Örneklem Büyüklüğü	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit
3	16000	0.700	0.700	0.401	0.701	0.571	0.571	0.341	0.683	0.697	0.700	0.591	0.701	0.701	0.702	0.601	0.701
5	16000	0.947	0.899	0.451	0.951	0.580	0.574	0.341	0.703	0.939	0.901	0.589	0.951	0.950	0.909	0.611	0.951
7	16000	0.952	0.929	0.512	0.959	0.592	0.574	0.347	0.708	0.948	0.911	0.592	0.959	0.955	0.931	0.612	0.959

Ek 3

Bağımlı Değişken ve Düzeyi Y; 1-2-3-4; Dengeli Kategori Dağılımı; %25 1-2-3-4																	
Çoklu Bağlantı Düzeyi 0.3	İterasyon 100	Sıralı Lojistik Regresyon				Ridge Düzenlemeli				Lasso Düzenlemeli				Elastik Net Düzenlemeli			
Bağımsız Değişken Sayısı	Örneklem Büyüklüğü	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit
3	16000	0.718	0.727	0.435	0.708	0.570	0.571	0.361	0.672	0.745	0.766	0.352	0.769	0.745	0.765	0.349	0.764
5	16000	0.957	0.952	0.614	0.958	0.579	0.578	0.370	0.695	0.947	0.938	0.665	0.949	0.955	0.952	0.609	0.957
7	16000	0.968	0.962	0.615	0.969	0.591	0.577	0.364	0.701	0.949	0.948	0.668	0.957	0.969	0.967	0.611	0.969

Ek 4

Bağımlı Değişken ve Düzeyi Y; 1-2-3-4; Dengeli Kategori Dağılımı; %25 1-2-3-4																	
Çoklu Bağlantı Düzeyi 0.9	İterasyon 100	Sıralı Lojistik Regresyon				Ridge Düzenlemeli				Lasso Düzenlemeli				Elastik Net Düzenlemeli			
Bağımsız Değişken Sayısı	Örneklem Büyüklüğü	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit
3	4000	0.670	0.689	0.269	0.671	0.481	0.486	0.228	0.485	0.680	0.712	0.299	0.698	0.684	0.724	0.296	0.731
5	4000	0.911	0.901	0.274	0.918	0.497	0.501	0.249	0.502	0.921	0.911	0.314	0.926	0.928	0.922	0.323	0.928
7	4000	0.928	0.928	0.301	0.929	0.502	0.508	0.297	0.506	0.934	0.939	0.321	0.933	0.941	0.938	0.334	0.932

Ek 5

Bağımlı Değişken ve Düzeyi Y; 1-2-3-4; Dengeli Kategori Dağılımı; %25 1-2-3-4																	
Çoklu Bağlantı Düzeyi 0.6	İterasyon 100	Sıralı Lojistik Regresyon				Ridge Düzenlemeli				Lasso Düzenlemeli				Elastik Net Düzenlemeli			
Bağımsız Değişken Sayısı	Örneklem Büyüklüğü	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit
3	4000	0.925	0.932	0.554	0.934	0.468	0.518	0.401	0.561	0.949	0.968	0.599	0.964	0.951	0.953	0.621	0.959
5	4000	0.942	0.954	0.579	0.952	0.482	0.521	0.420	0.565	0.952	0.974	0.625	0.971	0.961	0.976	0.642	0.968
7	4000	0.951	0.972	0.638	0.971	0.501	0.537	0.462	0.592	0.961	0.980	0.643	0.979	0.970	0.978	0.688	0.972

Ek 6

Bağımlı Değişken ve Düzeyi Y; 1-2-3-4; Dengeli Kategori Dağılımı; %25 1-2-3-4																	
Çoklu Bağlantı Düzeyi 0.3	İterasyon 100	Sıralı Lojistik Regresyon				Ridge Düzenlemeli				Lasso Düzenlemeli				Elastik Net Düzenlemeli			
Bağımsız Değişken Sayısı	Örneklem Büyüklüğü	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit
3	4000	0.931	0.928	0.492	0.935	0.357	0.428	0.293	0.466	0.933	0.931	0.522	0.938	0.929	0.936	0.528	0.946
5	4000	0.944	0.939	0.505	0.958	0.369	0.421	0.306	0.465	0.949	0.942	0.555	0.959	0.946	0.949	0.559	0.961
7	4000	0.953	0.964	0.516	0.965	0.402	0.432	0.389	0.469	0.959	0.964	0.561	0.969	0.957	0.966	0.561	0.969

Ek 7

Bağımlı Değişken ve Düzeyi Y; 1-2-3-4; Dengeli Kategori Dağılımı; %25 1-2-3-4																	
Çoklu Bağlantı Düzeyi 0.9	İterasyon 100	Sıralı Lojistik Regresyon				Ridge Düzenlemeli				Lasso Düzenlemeli				Elastik Net Düzenlemeli			
Bağımsız Değişken Sayısı	Örneklem Büyüklüğü	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit
3	400	0.899	0.895	0.711	0.889	0.580	0.524	0.391	0.601	0.901	0.899	0.811	0.889	0.902	0.910	0.815	0.897
5	400	0.955	0.948	0.724	0.948	0.559	0.567	0.399	0.599	0.961	0.950	0.828	0.967	0.959	0.956	0.831	0.971
7	400	0.961	0.951	0.772	0.954	0.571	0.569	0.400	0.603	0.966	0.955	0.829	0.968	0.967	0.959	0.831	0.971

Ek 8

Bağımlı Değişken ve Düzeyi Y; 1-2-3-4; Dengeli Kategori Dağılımı; %25 1-2-3-4																	
Çoklu Bağlantı Düzeyi 0.6	İterasyon 100	Sıralı Lojistik Regresyon				Ridge Düzenlemeli				Lasso Düzenlemeli				Elastik Net Düzenlemeli			
Bağımsız Değişken Sayısı	Örneklem Büyüklüğü	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit
3	400	0.890	0.851	0.710	0.899	0.592	0.589	0.399	0.619	0.891	0.855	0.722	0.903	0.897	0.856	0.711	0.901
5	400	0.961	0.955	0.737	0.960	0.598	0.653	0.412	0.665	0.971	0.956	0.743	0.965	0.969	0.959	0.741	0.967
7	400	0.967	0.953	0.765	0.964	0.613	0.661	0.457	0.689	0.969	0.961	0.769	0.969	0.969	0.958	0.767	0.970

Ek 9

Bağımlı Değişken ve Düzeyi Y; 1-2-3-4; Dengeli Kategori Dağılımı; %25 1-2-3-4																	
Çoklu Bağlantı Düzeyi 0.3	İterasyon 100	Sıralı Lojistik Regresyon				Ridge Düzenlemeli				Lasso Düzenlemeli				Elastik Net Düzenlemeli			
Bağımsız Değişken Sayısı	Örneklem Büyüklüğü	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit
3	400	0.917	0.913	0.793	0.895	0.595	0.593	0.411	0.632	0.917	0.913	0.795	0.896	0.918	0.915	0.796	0.897
5	400	0.987	0.971	0.791	0.988	0.601	0.642	0.429	0.667	0.988	0.972	0.792	0.989	0.989	0.974	0.792	0.991
7	400	0.988	0.979	0.701	0.989	0.615	0.664	0.492	0.693	0.989	0.981	0.709	0.991	0.990	0.981	0.711	0.992

Ek 10

Bağımlı Değişken ve Düzeyi Y; 1-2-3-4; Dengesiz Kategori Dağılımı; %55 1; %15 2-3-4																	
Çoklu Bağlantı Düzeyi 0.9	İterasyon 100	Sıralı Lojistik Regresyon				Ridge Düzenlemeli				Lasso Düzenlemeli				Elastik Net Düzenlemeli			
Bağımsız Değişken Sayısı	Örneklem Büyüklüğü	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit
3	16000	0.720	0.716	0.322	0.729	0.680	0.682	0.351	0.693	0.738	0.727	0.365	0.741	0.748	0.726	0.715	0.748
5	16000	0.940	0.929	0.641	0.948	0.699	0.689	0.360	0.711	0.967	0.928	0.645	0.971	0.970	0.929	0.721	0.972
7	16000	0.969	0.959	0.645	0.968	0.712	0.699	0.362	0.712	0.973	0.969	0.648	0.978	0.974	0.971	0.744	0.979

Ek 11

Bağımlı Değişken ve Düzeyi Y; 1-2-3-4; Dengesiz Kategori Dağılımı; %55 1; %15 2-3-4																	
Çoklu Bağlantı Düzeyi 0.6	İterasyon 100	Sıralı Lojistik Regresyon				Ridge Düzenlemeli				Lasso Düzenlemeli				Elastik Net Düzenlemeli			
Bağımsız Değişken Sayısı	Örneklem Büyüklüğü	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit
3	16000	0.734	0.730	0.381	0.741	0.681	0.688	0.372	0.699	0.797	0.730	0.591	0.801	0.741	0.732	0.751	0.791
5	16000	0.966	0.932	0.651	0.961	0.702	0.689	0.381	0.723	0.969	0.931	0.689	0.981	0.971	0.929	0.761	0.981
7	16000	0.972	0.969	0.662	0.971	0.719	0.718	0.387	0.728	0.978	0.971	0.742	0.988	0.975	0.974	0.783	0.989

Ek 12

Bağımlı Değişken ve Düzeyi Y; 1-2-3-4; Dengesiz Kategori Dağılımı; %55 1; %15 2-3-4																	
Çoklu Bağlantı Düzeyi 0.3	İterasyon 100	Sıralı Lojistik Regresyon				Ridge Düzenlemeli				Lasso Düzenlemeli				Elastik Net Düzenlemeli			
Bağımsız Değişken Sayısı	Örneklem Büyüklüğü	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit
3	16000	0.818	0.737	0.434	0.748	0.680	0.771	0.391	0.722	0.815	0.796	0.552	0.829	0.755	0.774	0.809	0.804
5	16000	0.972	0.952	0.664	0.968	0.779	0.778	0.410	0.795	0.977	0.968	0.695	0.983	0.976	0.942	0.828	0.987
7	16000	0.988	0.972	0.685	0.979	0.811	0.817	0.464	0.801	0.989	0.988	0.738	0.987	0.979	0.979	0.841	0.988

Ek 13

Bağımlı Değişken ve Düzeyi Y; 1-2-3-4; Dengesiz Kategori Dağılımı; %55 1; %15 2-3-4																	
Çoklu Bağlantı Düzeyi 0.9	İterasyon 100	Sıralı Lojistik Regresyon				Ridge Düzenlemeli				Lasso Düzenlemeli				Elastik Net Düzenlemeli			
Bağımsız Değişken Sayısı	Örneklem Büyüklüğü	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit
3	4000	0.912	0.809	0.498	0.901	0.491	0.494	0.328	0.495	0.919	0.812	0.499	0.908	0.921	0.842	0.464	0.918
5	4000	0.930	0.841	0.545	0.918	0.499	0.522	0.359	0.514	0.931	0.891	0.514	0.926	0.936	0.881	0.501	0.934
7	4000	0.958	0.868	0.551	0.926	0.512	0.528	0.395	0.516	0.959	0.909	0.551	0.938	0.966	0.905	0.562	0.939

Ek 14

Bağımlı Değişken ve Düzeyi Y; 1-2-3-4; Dengesiz Kategori Dağılımı; %55 1; %15 2-3-4																	
Çoklu Bağlantı Düzeyi 0.6	İterasyon 100	Sıralı Lojistik Regresyon				Ridge Düzenlemeli				Lasso Düzenlemeli				Elastik Net Düzenlemeli			
Bağımsız Değişken Sayısı	Örneklem Büyüklüğü	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit
3	4000	0.925	0.821	0.524	0.934	0.498	0.508	0.411	0.542	0.939	0.908	0.519	0.944	0.941	0.913	0.621	0.949
5	4000	0.942	0.852	0.554	0.952	0.502	0.523	0.419	0.545	0.942	0.914	0.545	0.961	0.946	0.916	0.647	0.966
7	4000	0.951	0.874	0.568	0.971	0.507	0.531	0.432	0.552	0.961	0.919	0.603	0.979	0.970	0.928	0.680	0.979

Ek 15

Bağımlı Değişken ve Düzeyi Y; 1-2-3-4; Dengesiz Kategori Dağılımı; %55 1; %15 2-3-4																	
Çoklu Bağlantı Düzeyi 0.3	İterasyon 100	Sıralı Lojistik Regresyon				Ridge Düzenlemeli				Lasso Düzenlemeli				Elastik Net Düzenlemeli			
Bağımsız Değişken Sayısı	Örneklem Büyüklüğü	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit
3	4000	0.931	0.828	0.522	0.945	0.557	0.526	0.413	0.553	0.943	0.911	0.525	0.948	0.943	0.916	0.628	0.951
5	4000	0.968	0.871	0.565	0.969	0.563	0.571	0.466	0.568	0.958	0.972	0.575	0.979	0.959	0.979	0.679	0.980
7	4000	0.973	0.894	0.576	0.972	0.602	0.572	0.489	0.569	0.967	0.979	0.582	0.980	0.970	0.980	0.691	0.981

Ek 16

Bağımlı Değişken ve Düzeyi Y; 1-2-3-4; Dengesiz Kategori Dağılımı; %55 1; %15 2-3-4																	
Çoklu Bağlantı Düzeyi 0.9	İterasyon 100	Sıralı Lojistik Regresyon				Ridge Düzenlemeli				Lasso Düzenlemeli				Elastik Net Düzenlemeli			
Bağımsız Değişken Sayısı	Örneklem Büyüklüğü	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit
3	400	0.879	0.865	0.411	0.869	0.570	0.516	0.390	0.598	0.862	0.872	0.837	0.900	0.899	0.886	0.867	0.901
5	400	0.898	0.890	0.500	0.902	0.548	0.543	0.392	0.549	0.890	0.880	0.861	0.901	0.902	0.890	0.871	0.907
7	400	0.911	0.891	0.612	0.954	0.568	0.524	0.396	0.671	0.894	0.881	0.869	0.904	0.921	0.891	0.871	0.908

Ek 17

Bağımlı Değişken ve Düzeyi Y; 1-2-3-4; Dengesiz Kategori Dağılımı; %55 1; %15 2-3-4																	
Çoklu Bağlantı Düzeyi 0.6	İterasyon 100	Sıralı Lojistik Regresyon				Ridge Düzenlemeli				Lasso Düzenlemeli				Elastik Net Düzenlemeli			
Bağımsız Değişken Sayısı	Örneklem Büyüklüğü	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit
3	400	0.894	0.551	0.410	0.899	0.512	0.519	0.391	0.519	0.889	0.901	0.842	0.899	0.901	0.903	0.897	0.901
5	400	0.920	0.622	0.521	0.921	0.498	0.553	0.392	0.545	0.949	0.949	0.859	0.956	0.959	0.951	0.961	0.966
7	400	0.967	0.653	0.615	0.974	0.511	0.561	0.397	0.649	0.968	0.950	0.859	0.966	0.971	0.952	0.962	0.977

Ek 18

Bağımlı Değişken ve Düzeyi Y; 1-2-3-4; Dengesiz Kategori Dağılımı; %55 1; %15 2-3-4																	
Çoklu Bağlantı Düzeyi 0.3	İterasyon 100	Sıralı Lojistik Regresyon				Ridge Düzenlemeli				Lasso Düzenlemeli				Elastik Net Düzenlemeli			
Bağımsız Değişken Sayısı	Örneklem Büyüklüğü	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit	logit	probit	cloglog	cauchit
3	400	0.717	0.730	0.493	0.705	0.501	0.533	0.401	0.532	0.742	0.743	0.511	0.741	0.748	0.751	0.521	0.748
5	400	0.723	0.733	0.500	0.715	0.495	0.542	0.424	0.541	0.745	0.749	0.542	0.743	0.745	0.754	0.552	0.749
7	400	0.725	0.739	0.501	0.719	0.497	0.544	0.482	0.543	0.764	0.751	0.543	0.747	0.769	0.771	0.553	0.767

Ek 19

	categHat				
testy$Yo	1	2	3	4	Sum
1	1616	55	0	0	1671
2	72	1590	80	0	1742
3	0	76	1459	69	1604
4	0	0	60	1515	1575
Sum	1688	1721	1599	1584	6592

Ek 20

	Preds			
testy$Yo	1	3	4	Sum
1	82	0	0	82
2	22	0	0	22
3	17	13	1	31
4	2	11	8	21
Sum	123	24	9	156

Ek 21

	Preds				
testy$Yo	1	2	3	4	Sum
1	81	1	0	0	82
2	2	19	1	0	22
3	0	0	27	0	27
4	0	0	2	23	25
Sum	83	20	30	23	156

Ek 22

	Preds				
testy$Yo	1	2	3	4	Sum
1	417	7	0	0	424
2	4	413	4	0	421
3	0	5	352	3	360
4	0	0	8	392	400
Sum	421	425	364	395	1605

Ek 23

```
library(VGAM)
set.seed(123)
reps <- 100
n <- 16000
p=0.9
a=sqrt(1-p^2)
e1  <- rnorm(n, 0, 1.0)
e2  <- rnorm(n, 0, 1.1)
e3  <- rnorm(n, 0, 12.3)
e4  <- rnorm(n, 0, 5.4)
e5  <- rnorm(n, 0, 10.5)
e6  <- rnorm(n, 0, 12.6)
e7  <- rnorm(n, 0, 1.7)
e8  <- rnorm(n, 0, 1.8)
e9  <- rnorm(n, 0, 0.9)
e10 <- rnorm(n, 0, 19.99)
X1=10+e1
X2=10+p*e1+a*e2
X3=1+e7+a*e3+0.4*p*e2
X4=-8+X1+0.5*X2+p*X3+e4
X5=5+0.5*X1+X2+e5
X6= 9+e7+a*e10+0.5*p*e8
X7=5+p*e6+a*e7
X8=-3+0.1*X1+X2+e8
X9=9+p*e2+0.3*a*e9+0.6*a*e4
X10=8+e7+a*e9+0.4*p*e2
allvariables = data.frame(X1,X2,X3,X4,X5,X6,X7,X8,X9,X10)
cor(allvariables)
b0 <- 8
b1 <- 1.5
b2 <- 0.3
b3 <- 1.2
```

```
b4 <- 10.9
b5 <- 2.65
b6 <- 1.4
b7 <- 0.91
b8 <- -10.1
b9 <- 1.1
b10 <- 1.2
XB <- b0 + b1*X1 + b2*X2 + b9*X9 + b6*X6 + b10*X10   sd.error <- 1
tau1 <- qnorm(.25, mean = mean(XB), sd = sqrt(var(XB) + sd.error^2))
tau2 <- qnorm(.50, mean = mean(XB), sd = sqrt(var(XB) + sd.error^2))
tau3 <- qnorm(.75, mean = mean(XB), sd = sqrt(var(XB) + sd.error^2))
for(i in 1:reps)
{
  Y.star <- rnorm(n, XB, sd.error)
  Yo <- rep(NA, n)
  Yo[Y.star < tau1] <- 1
  Yo[Y.star >= tau1 & Y.star < tau2] <- 2
  Yo[Y.star >= tau2 & Y.star < tau3] <- 3
  Yo[Y.star >= tau3] <- 4
  fulldata = data.frame(Yo,X1,X2,X9,X6,X10)
  ind <- sample(2, nrow(fulldata), replace=TRUE, prob = c(0.60,0.40))
  trainy <- fulldata[ind==1,]
  testy <- fulldata[ind==2,]
  modelr <- vglm(as.ordered(Yo)~ X1+X2+X6+X9+X10,
                  family=cumulative(link = "logitlink",
                 parallel = TRUE,  reverse = TRUE),
                data= trainy, model = TRUE)
 cat("just completed", i, "\n")
}
PhatCateg <- predict(modelr, testy, type="response")
categHat <- max.col(PhatCateg)
cTab   <- xtabs(~  testy$Yo + categHat, data= testy)
addmargins(cTab)
(CCR <- sum(diag(cTab)) / sum(cTab)
```