

**T.C.
ONDOKUZ MAYIS ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

YÜKSEK LİSANS TEZİ

**ÇOK DEĞİŞKENLİ BERNOULLİ LOJİSTİK MODELLERDE LASSO
TAHMİNCİLERİYLE MODEL SEÇİMİ**

ASLI YAMAN

İSTATİSTİK ANABİLİM DALI

**SAMSUN
2017**

Her hakkı saklıdır.

TEZ ONAYI

Aslı YAMAN tarafından hazırlanan “ÇOK DEĞİŞKENLİ BERNOULLI LOJİSTİK MODELLERDE LASSO TAHMİNCİLERİYLE MODEL SEÇİMİ” adlı tez çalışması 19/07/2017 tarihinde aşağıdaki jüri tarafından Ondokuz Mayıs Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı’nda **Yüksek Lisans Tezi** olarak kabul edilmiştir.

Danışman Prof. Dr. Mehmet Ali CENGİZ
İstatistik Anabilim Dalı

Jüri Üyeleri

Başkan Prof. Dr. Mehmet Ali CENGİZ
Ondokuz Mayıs Üniversitesi
İstatistik Anabilim Dalı

Üye Doç. Dr. Yüksel TERZİ
Ondokuz Mayıs Üniversitesi
İstatistik Anabilim Dalı

Üye Doç. Dr. Nesrin ALKAN
Sinop Üniversitesi
İstatistik Anabilim Dalı

Yukarıdaki sonucu onaylarım. .../.../20..

.....

Prof. Dr. Bahtiyar ÖZTÜRK
Enstitü Müdürü

ETİK BEYAN

Ondokuz Mayıs Üniversitesi Fen Bilimleri Enstitüsü tez yazım kurallarına uygun olarak hazırladığım bu tez içindeki bütün bilgilerin doğru ve tam olduğunu, bilgilerin üretilmesi aşamasında bilimsel etiğe uygun davrandığımı, yararlandığım bütün kaynakları atıf yaparak belirttiğimi beyan ederim.

19.06.2017

Aslı YAMAN

ÖZET

Yüksek Lisans Tezi

ÇOK DEĞİŞKENLİ BERNOULLI LOJİSTİK MODELLERDE LASSO TAHMİNCİLERİYLE MODEL SEÇİMİ

Aslı Yaman

Ondokuz Mayıs Üniversitesi

Fen Bilimleri Enstitüsü

İstatistik Anabilim Dalı

Danışman: Prof. Dr. Mehmet Ali Cengiz

Bu çalışmada, çok değişkenli Bernoulli lojistik modellerde LASSO tahmincilerinin incelenmesi ve model seçiminde kullanılan Akaike Bilgi Kriteri (AIC), Bayesci Bilgi Kriteri (BIC), Genelleştirilmiş çapraz doğrulama yaklaşım kriteri (GACV) ve Bayesci genelleştirilmiş çapraz doğrulama yaklaşım kriteri (BGACV) ile elde edilen tahmin modellerinin karşılaştırılması amaçlanmıştır.

Çalışmada ilk olarak LASSO tahmincileri ve çok değişkenli Bernoulli lojistik modelleri hakkında bilgi verilmiştir. Daha sonra R programında MVB paketi kullanılarak, tarafımızdan belirlenen başlangıç beta değerleri için model seçiminde kullanılan dört farklı bilgi kriteri ile LASSO tahmin değerleri elde edilmiştir. Farklı gözlem değerleri ve farklı bağımlı değişken sayıları için simülasyonlar yapılmıştır. Simülasyon sonuçları kriterlere göre karşılaştırılmıştır.

Haziran 2017, 45 sayfa

Anahtar Kelimeler: LASSO, Bernoulli dağılımı, lojistik model, Genelleştirilmiş çapraz doğrulama yaklaşım kriteri

ABSTRACT

Master's Thesis

LASSO FEATURE SELECTION IN MULTIVARIATE BERNOULLI LOGISTIC MODELS

Aslı Yaman

Ondokuz Mayıs University

Graduate School of Science

Department of Statistics

Supervisor: Prof. Dr. Mehmet Ali Cengiz

In study, it is purpose of examine the LASSO estimators in multivariate Bernoulli logistic models and compare the estimation models obtained with Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), Generalized approximate cross validation (GACV) and Bayesian generalized approximate cross validation (BGACV) used in model selection.

Firstly, LASSO estimators and multivariate Bernoulli logistic models were given. Then, using the MVB package in the R program, LASSO estimates were obtained with four different information criterions used in the model selection for the initial beta values determined by us. Simulations have been done for different observation values and different dependent variable numbers. Simulation results are compared according to the criteria.

June 2017, 45 pages

Keywords: LASSO, Bernoulli distribution, logistic model, Generalized approximate cross validation

ÖNSÖZ VE TEŞEKKÜR

Akademik eğitim sürecimin bir üst noktası olan yüksek lisans tez çalışmalarım boyunca yardım ve desteğini benden esirgemeyen danışman hocam Sayın Prof. Dr. Mehmet Ali Cengiz' e ve çalışmalarımda bana destek veren Sayın Araş. Gör. Emre Dünder' e teşekkürü bir borç bilirim.

Maddi ve manevi olarak her şekilde beni destekleyen ve hiçbir fedakârlıktan kaçınmayan değerli anneme ve babama sonsuz teşekkür ederim.

Bu tez çalışması, PYO. FEN. 1904. 17. 002 nolu Bilimsel Araştırma Projesi olarak Ondokuz Mayıs Üniversitesi tarafından desteklenmektedir.

Haziran 2017, Samsun

Aslı Yaman

İÇİNDEKİLER DİZİNİ

ÖZET.....	i
ABSTRACT.....	ii
ÖNSÖZ VE TEŞEKKÜR	iii
İÇİNDEKİLER DİZİNİ	iv
SİMGELER VE KISALTMALAR.....	v
ŞEKİLLER DİZİNİ.....	vi
ÇİZELGELER DİZİNİ	vii
1. GİRİŞ	1
2. KURAMSAL TEMELLER VE KAYNAK ÖZETLERİ.....	4
2.1. Model Tahmini	5
2.1.1. En küçük kareler yöntemi	5
2.2.2. En çok olabilirlik tahmin yöntemi	5
3. ÇOKLU DOĞRUSAL BAĞLANTI GİDERME YÖNTEMLERİ	7
3.1. Ridge Regresyonu	7
3.2. LASSO Tahmincisi	8
3.2.1. LASSO algoritması çözümü	10
3.2.2. LASSO ayar parametre seçimi	11
3.2.2.1. K katlı çapraz doğrulama	11
3.2.2.2. Birini dışarıda bırakarak çapraz doğrulama	12
3.2.2.3. Genelleştirilmiş çapraz doğrulama.....	12
4. ÇOK DEĞİŞKENLİ BERNOULLİ LOJİSTİK MODEL	13
4.1. Logaritmik Doğrusal ve Lojistik Modeller	13
4.2. Bernoulli Dağılımı.....	14
4.3. İki Değişkenli Bernoulli Lojistik LASSO Model.....	15
4.4. Model Uyum İstatistikleri	16
4.4.1. Akaike bilgi kriteri.....	16
4.4.2. Bayesci bilgi kriteri.....	16
4.4.3. Genelleştirilmiş çapraz doğrulama yaklaşım kriteri	17
4.4.4. Bayesci genelleştirilmiş çapraz doğrulama yaklaşım kriteri	18
5. BULGULAR VE TARTIŞMA	20
6. SONUÇ VE ÖNERİLER	39
KAYNAKLAR	42
ÖZGEÇMİŞ	46

SİMGELER VE KISALTMALAR

AIC	Akaike Bilgi Kriteri
BGACV	Bayesci Genelleştirilmiş Çapraz Doğrulama Yaklaşımı
BIC	Bayesci Bilgi Kriteri
ÇDB	Çoklu Doğrusal Bağlantı
EÇOT	En Çok Olabilirlik Tahmini
EKK	En Küçük Kareler
GACV	Genelleştirilmiş Çapraz Doğrulama Yaklaşımı
HKT	Hata Kareler Toplamı
KL	Kullback-Leibler
LARS	En Küçük Açık Regresyonu ve Daraltıcısı
LASSO	En Küçük Mutlak Daralma ve Seçim Operatörü
MVB	Çok Değişkenli Bernoulli

ŞEKİLLER DİZİNİ

Şekil 3.1. l-q normu için farklı q değerlerinin grafiksel gösterimi(Friedman vd, 2001).....	9
Şekil 3.2. LASSO ve Ridge regresyonu tahminleri (Hastie vd, 2015)	9
Şekil 5.1. $p=5$ ve $k=2$ değerleri sabit iken $n=50, 100, 200, 250$ ve 300 için kriterler.....	25
Şekil 5.2. $p=5$ ve $k=3$ değerleri sabit iken $n=50, 100, 200, 250$ ve 300 için kriterler.....	26
Şekil 5.3. $k=2$ ve $n=50$ iken $p=5, 15, 20, 25, 50$ ve 55 için kriterler.....	28
Şekil 5.4. $k=2$ ve $n=40$ iken $p=5, 10, 20, 40, 45, 50$ ve 60 için kriterler.....	30
Şekil 5.5. $k=2$ ve $n=30$ iken $p=5, 10, 15, 30, 35, 40$ ve 45 için kriterler.....	32
Şekil 5.6. $k=2$ ve $n=50$ iken $p=5, 15, 20, 25, 50$ ve 55 için farklar.....	34
Şekil 5.7. $k=2$ ve $n=40$ iken $p=5, 10, 20, 40, 45, 50$ ve 60 için farklar.....	36
Şekil 5.8. $k=2$ ve $n=30$ iken $p=5, 10, 15, 30, 35, 40$ ve 45 için farklar.....	38

ÇİZELGELER DİZİNİ

Çizelge 5.1. Bağımlı değişken sayısı 2 iken başlangıç beta değerleri	20
Çizelge 5.2. $n=50$, $p=5$ ve $k=2$ için uyum iyiliği ölçütleri ile elde edilen beta tahminleri.....	21
Çizelge 5.3. $n=300$, $p=5$ ve $k=2$ için uyum iyiliği ölçütleri ile elde edilen beta tahminleri.	22
Çizelge 5.4. $p=5$, $k=2, 3$, $n=50, 100, 200, 250$ ve 300 iken elde edilen sonuçlar	23
Çizelge 5.5. $k=2$, $n=50$ ve $p=5, 15, 20, 25, 50$ ve 55 iken kriterler	27
Çizelge 5.6. $k=2$, $n=40$ ve $p=5, 10, 20, 40, 45, 50$ ve 60 iken kriterler.....	29
Çizelge 5.7. $k=2$, $n=30$ ve $p=5, 10, 15, 30, 35, 40$ ve 45 iken kriterler.....	31
Çizelge 5.8. $k=2$, $n=50$ ve $p=5, 15, 20, 25, 50$ ve 55 iken fark değerleri.....	33
Çizelge 5.9. $k=2$, $n=40$ ve $p=5, 10, 20, 40, 45, 50$ ve 60 iken fark değerleri.....	35
Çizelge 5.10. $k=2$, $n=30$ ve $p=5, 10, 15, 30, 35, 40$ ve 45 iken fark değerleri.....	37

1. GİRİŞ

İstatistik farklı birçok tanımı olması ile birlikte; farklı bilim dallarında veya alanlarda ortaya çıkan/ çıkabilecek sorunlarda ya da araştırmalarda yol gösterecek verilerin doğru bir şekilde toplanarak yorumlanmasını yani analiz edilerek anlamlı bilgilere ulaşılmasını sağlayan bir bilimdir. Toplanan verilerin analiz edilmesi için birçok yöntem bulunmaktadır. İstatistiksel yöntemlerden regresyon analizi ile iki ya da daha fazla değişken arasındaki ilişki açıklanabilmektedir. Bu ilişkiyi gösterebilmek ve formülize edebilmek için değişkenler için en uygun modeli belirlemek amaçlanmaktadır. Model belirlenirken, oluşturulan modelin veriyi en doğru şekilde yansıtması beklenmektedir. Yani oluşturulan modele göre elde edilen tahmin değerleri olası gerçek değerlere yakın olmalıdır. Bunun için, veri setlerinden çıkarım yapmayı sağlamak ve tahmin modelleri oluşturabilmek için farklı yöntemler geliştirilmiştir. Yaygın olarak kullanılan tahmin yöntemlerinden biri de en küçük kareler yöntemidir (EKK).

En küçük kareler yöntemi ile hata kareler toplamı minimize edilmektedir ve bu yöntemin bazı varsayımları bulunmaktadır. Bunlar; hata terimleri sıfır ortalamalı ve sabit varyanslı olmalı, hata terimleri arasında içsel bağlantı sorunu olmamalı ve bağımsız değişkenler arasında doğrusal ilişki olmamalı şeklinde özetlenebilir. Varsayımlarının sağlanması durumunda EKK tahmincileri yansız ve minimum varyanslı tahminci olarak kabul edilmektedir. Bağımsız değişkenler arasında ilişki olması durumuna ise çoklu doğrusal bağlantı sorunu (ÇDB) denilmektedir.

Özellikle büyük veri setlerinde karşılaşılan, çoklu doğrusal bağlantı problemi olması durumunda EKK yöntemi ile güvenilir tahmin sonuçları elde etmek zorlaşabilmektedir. Çünkü ÇDB problemi olduğu durumlarda EKK tahminleri yansız olmakla beraber varyans değerlerindeki artış sonucu EKK tahmini en küçük varyanslı tahmin edici olma özelliğini kaybetmektedir. Elde edilen tahmin değerlerinin güvenilirliğini artırmak için ÇDB problemi giderilmelidir.

ÇDB probleminin giderilmesi için; bir veya birden fazla değişken modelden çıkarılabilir, gözlem sayısı artırılabilir, birbiri ile ilişkili olan iki değişkenin toplamı alınarak modele dâhil edilebilir vb şeklinde uygulamalar yapılabilir. Ancak bu işlemler yapılırken de farklı sorunlar ortaya çıkabilmektedir. Örneğin, değişken çıkarma işlemi yapılmak istendiğinde, hangi değişkenin modelden çıkarıldığı

önemlidir ve çıkarılmaması gereken bir değişken çıkarılması durumunda sonuçlar etkilenebilmektedir. Bu sebeple ÇDB problemi durumunda EKK tahmin doğruluğunu artırmak için alternatif çözüm yöntemlerine başvurulabilir. Bu yöntemlerden biri Shrinkage (daraltıcı, cezalı) yöntemidir (Hoerl ve Kennard, 1970).

Daraltıcı yöntemlerden LASSO (En Küçük Mutlak Daralma ve Seçim Operatörü) tahmincisi ile EKK tahmin denkleminde belirli bir ceza kısıdı getirilerek tahmin doğruluğu artırılmak amaçlanmaktadır. LASSO tahmincisinde ceza kısıdından kaynaklı tahmin denkleminde bir veya birden fazla parametre sıfıra daraltarak özellikle büyük veri setlerinde yorumlanması kolay modeller elde edilmesine olanak sağlamaktadır. Değişken seçimi ve parametre tahminini aynı anda yapılmasına imkân sağlayarak işlem kolaylığı sağlamasından dolayı LASSO yaygın olarak tercih edilen bir tahmincidir.

İlk olarak Tibshirani (1996) tarafından EKK tahmin doğruluğunu artırmak için ortaya atılmıştır ve zaman içerisinde kullanım alanları genişletilmiştir. Günümüzde özellikle sağlık alanı başta olmak üzere birçok farklı alanda kullanılmaktadır. Kullanım alanları incelendiğinde çalışmaların özellikle çok büyük veri setlerinde ve değişken sayısının gözlem sayısından fazla olduğu veri setleri üzerinde yoğunlaştığı görülebilmektedir.

Bu çalışmada çok değişkenli Bernoulli lojistik modellerde LASSO tahmincileri incelendi. Ayrıca model seçiminde kullanılan Akaike Bilgi Kriteri (AIC), Bayesci Bilgi Kriteri (BIC), Genelleştirilmiş Çapraz Doğrulama Yaklaşımı (GACV) ve Bayesci Genelleştirilmiş Çapraz Doğrulama Yaklaşımı (BGACV) kriterleri yapılan farklı simülasyonlar için karşılaştırıldı ve LASSO tahminleri için en güçlü sonuçları veren kriterler belirlendi.

Bu tez 6 bölümden oluşmaktadır. İlk bölümde, LASSO tahminci için yapılan çalışmalar için literatür incelendi ve ikinci bölümde parametre tahmin yöntemlerinden EKK ve EÇOT yöntemleri verildi.

Üçüncü bölümde çoklu doğrusal bağlantı problemi olması durumunda problemin giderilmesinde alternatif olarak kullanılan daraltıcı yöntemlerden bahsedildi. Bu yöntemlerden Ridge regresyonu ve LASSO tahmincisi hakkında bilgi verildi. LASSO tahmincisi için modelde tek tahminci olduğunda ve birden fazla tahminci olması durumunda amaç fonksiyonun nasıl çözüleceği ve LASSO ayar

parametre seçiminde kullanılan yöntemlerden çapraz doğrulama yönteminden bahsedildi.

Dördüncü bölümde Bernoulli dağılımının çok değişkenli durumu incelendi. Çok değişkenli Bernoulli lojistik modele LASSO ceza kısıdı getirilerek elde edilen amaç fonksiyonu incelendi. Ayrıca model ve ayar parametre seçiminde kullanılan model uyum istatistiklerinden AIC, BIC, GACV ve BGACV hakkında bilgi verildi.

Beşinci bölümde çok değişkenli Bernoulli lojistik LASSO model belirlenerek LASSO tahminleri için simülasyonlar yapıldı. Simülasyonlar R programında MVB paketi kullanılarak; gözlem sayısı, bağımlı ve bağımsız değişken sayılarının farklı değerleri belirlenerek yapıldı.

Son olarak ise sonuç ve öneri kısmında, yapılan simülasyon çalışmalarından elde edilen sonuçlar verilmiştir.

2. KURAMSAL TEMELLER VE KAYNAK ÖZETLERİ

Model seçiminde kullanılan istatistiksel yöntemlerden biri altküme seçimidir. Alt küme regresyonu; modelde değişkenlerin belirli bir altkümesinin tutulduğu ve diğer değişkenlerin modelden çıkarıldığı, EKK (En Küçük Kareler) tahmin yöntemine dayalı, yaygın olarak kullanılan bir yöntemdir. Breiman (1995); alt küme regresyon yöntemi ile daha tutarlı sonuçlar elde edilmesi için negatif-olmayan garotte yöntemini geliştirmiştir. Bu yöntem ile EKK tahmin sonuçlarının daha tutarlı olduğu görülmüştür. Fakat bu yöntem değişken sayısının gözlem sayısından daha fazla olduğu ($p > n$) veri setlerinde yetersiz kalmıştır. Tibshirani (1996); negatif olmayan garotte yönteminden yola çıkarak, $p > n$ durumlarında EKK tahmin yöntemi ile daha tutarlı sonuçlar elde edilebilmesi için LASSO tahmin yöntemini ortaya atmıştır (Tibshirani, 2011).

Zaman içerisinde LASSO tahmin yöntemi ile farklı alanlarda çalışmalar yapılmıştır (Chen ve Donoho, 1998; Chen vd, 2001). Tibshirani (1996) genelleştirilmiş lineer modeller üzerinde ve Tibshirani (1997) yaşam verileri için Cox ' un orantısız risk modelleri üzerinde çalışmalar yapılmıştır.

LASSO tahmincisinin tutarlı sonuçlar vermediği durumlar için uyarlamalı LASSO tahmin yöntemi geliştirilmiştir. uyarlamalı LASSO; LASSO tahmincisine göre kısmen aynı algoritmaya sahiptir. Aralarındaki fark; Uyarlamalı LASSO, LASSO tahmincisine oranla daha sade modeller elde edilmesine olanak sağlamaktadır ve uyarlamalı LASSO ile farklı katsayılar için farklı ceza kısıdı belirlenmektedir. Zou (2006), Zhang ve Lou (2007); uyarlamalı LASSO tahmin yöntemlerine üzerine yoğunlaşırken, Park ve Hastie (2007); Van de Geer (2008) LASSO tahmincisi ile genelleştirilmiş lineer modeller üzerinde ve Witten vd (2009) çok değişkenli yöntemler üzerinde çalışmışlardır. Meinhausen ve Bühlmann (2006), Meier vd (2008), Tibshirani (2011), Tibshirani vd (2011), Tibshirani ve Taylor (2012), Tibshirani (2013), Finch ve Hernandez (2016), Dalalyan (2017) LASSO tahmincisi ile yaptıkları farklı alanlardaki çalışmaları ile önemli katkıları bulunmaktadır.

2.1. Model Tahmini

Doğrusal regresyon analizinde, veri seti için çıkarım yapmak amacıyla tahmin modeli oluşturulur ve toplanan veri seti hakkında fikir edinilebilmektedir. Bu şekilde parametreleri tahmin etmek için farklı yöntemler bulunmaktadır. Yaygın olarak kullanılan tahmin yöntemleri; en küçük kareler yöntemi (EKK) ve en çok olabilirlik yöntemidir (EÇOT).

2.1.1. En küçük kareler yöntemi

Doğrusal bir regresyon modeli;

$$y_i = \beta_0 + \sum_{j=1}^p x_{ij}\beta_j + e \quad (2.1)$$

şeklindedir ve burada n gözlem sayısı, y bağımlı değişken, $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$ bağımsız değişkenler, β_0 ve $\beta = (\beta_1, \beta_2, \dots, \beta_p)$ bilinmeyen parametrelerdir ve e ; hata terimidir. En küçük kareler yöntemi (EKK);

$$\min_{\beta_0, \beta} \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p x_{ij}\beta_j)^2 \quad (2.2)$$

şeklinde hata kareler toplamı minimize edilerek tahmin modelinin oluşturulmasına olanak sağlamaktadır (Alma, 2008).

2.2.2. En çok olabilirlik tahmin yöntemi

En çok olabilirlik tahmin yöntemi; belli bir örneklem değerinin gerçekleşme olabilirliğini maksimum yapan kitle parametrelerini bulmaya çalışan bir yaklaşımdır. $x_1, x_2, \dots, x_n$ rastgele örneklem ve kitlenin dağılımı $f(x; \theta_1, \theta_2, \dots, \theta_k)$ olsun. Rastgele örneklemin ortak olasılık veya ortak olasılık yoğunluk fonksiyonu,

$$f(x_1, x_2, \dots, x_n; \theta_1, \theta_2, \dots, \theta_k) = \prod_{i=1}^n f(x_i; \theta_1, \theta_2, \dots, \theta_k) \quad (2.3)$$

eşitliği ile verilebilir yada kısaca

$$f(x; \theta) = \prod_{i=1}^n f(x_i; \theta_1, \theta_2, \dots, \theta_k) \quad (2.4)$$

şeklinde yazılabilir.

$f(x;\theta)$ 'nın fonksiyonunu θ 'nın fonksiyonu olarak düşünülduğünde olabilirlik fonksiyonu tanımlanmaktadır. $x_1, x_2, \dots, x_n$ rastgele örnekleme için olabilirlik fonksiyonu;

$$L(\theta; x) = \prod_{i=1}^n f(x_i; \theta_1, \theta_2, \dots, \theta_k) \quad (2.5)$$

şeklinde tanımlanır ve olabilirlik fonksiyonunu θ 'ya göre en çok yapılarak (kısmi türevleri alınarak) parametreler için tahmin ediciler elde edilebilmektedir (Scholz, 1985; Akdi, 2005).

Eşitlik (2.5) probleminin çözümü için gerekli koşul θ bilinmeyen parametrelerin logolabilirlik fonksiyonunun parametrelere göre birinci türevleri sıfır olmalıdır.

$$\frac{\partial \ln(L(\theta; x))}{\partial \theta} = 0 \quad (2.6)$$

3. ÇOKLU DOĞRUSAL BAĞLANTI GİDERME YÖNTEMLERİ

EKK tahmincileri yansız ve minimum varyanslı olması özelliklerinden dolayı iyi tahminciler olarak kabul edilmektedir. Fakat büyük veri setlerinde bağımsız değişkenler arasında bir ilişki olması durumunda yani çoklu doğrusal bağlantı olması durumunda EKK tahmincilerinin varyans değerlerinde artış gözlemlenir. Böyle bir durumda EKK tahmincileri ile daha güvenilir ve tutarlı sonuçlar elde edebilmek için çoklu doğrusal bağlantı sorunu giderilmeli veya farklı çözüm yöntemlerine başvurulmalıdır. Çoklu doğrusal bağlantı sorununun giderilmesi için kullanılan diğer yaklaşım daraltıcı (Shrinkage) yöntemlerdir. Bilinen daraltıcı yöntemler; Ridge regresyonu ve LASSO tahmincileridir.

3.1. Ridge Regresyonu

ÇDB olması durumunda EKK tahmincisinin varyansında artış gözlemlenebilmektedir. Ridge regresyonu ile yanlış tahminler elde edilir fakat tahminlerin varyansı EKK tahmincisinin varyansından daha küçüktür.

Ridge tahmin edicisi;

$$\hat{\beta}^{ridge} = \min_{\beta} \left\{ \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2 \right\}, \sum_{j=1}^p \beta_j^2 \leq t \text{ iken} \quad (3.1)$$

şeklinde formüle edilir. Burada n gözlem sayısı, y bağımlı değişken, $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$ bağımsız değişkenler, β_0 ve $\beta = (\beta_1, \beta_2, \dots, \beta_p)$ bilinmeyen parametrelerdir. (3.1) algoritmasının Lagrange formu;

$$\hat{\beta}^{ridge} = \min_{\beta} \left\{ \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p \beta_j^2 \right\} \quad (3.2)$$

olmaktadır (Theodoridis, 2015). λ ayar parametresidir ve daralma miktarını kontrol eder.

Ridge regresyonu ile modeldeki katsayılar ceza kısıdı ile sifıra doğru daraltılmaktadır fakat tam olarak sifıra daralabilen katsayılar elde edilememektedir. Diğer taraftan LASSO tahmincisi ile tam olarak sifıra daralabilen katsayılar elde etmek mümkündür. Bu sebeple özellikle büyük veri setlerinde Ridge regresyonuna kıyasla LASSO daha sade modeller elde edilerek, sonuçları değerlendirme ve işlem kolaylığı sağlaması açısından daha avantajlıdır.

3.2. LASSO Tahmincisi

LASSO; En Küçük Mutlak Daralma ve Seçim Operatörü (The Least Absolute Shrinkage and Selection Operator), ilk olarak Tibshirani (1996) tarafından tanıtıldı. İlk zamanlarda sadece EKK tahmin yöntemine alternatif olarak ortaya çıkmasına rağmen zaman içerisinde yapılan çalışmalar sonucunda uygulama alanları genişletildi (Zou ve Hastie, 2005; Candes ve Tao, 2007; Candes ve Tao, 2010; Hans, 2009). LASSO tahmincisi;

$$\min_{\beta_0, \beta} \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2, \quad \|\beta\|_1 < t \text{ için} \quad (3.3)$$

algoritması ile elde edilir. Eşitlik (3.3)' de verilen n gözlem sayısı, p değişken sayısı, y bağımlı değişken, $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$ bağımsız değişkenler, β_0 ve $\beta = (\beta_1, \beta_2, \dots, \beta_p)$ bilinmeyen parametrelerdir. $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$; β ' nın $l-1$ normudur yani mutlak değerlerinin toplamına eşittir ve t değeri daralma miktarını kontrol etmektedir (Tibshirani, 1996).

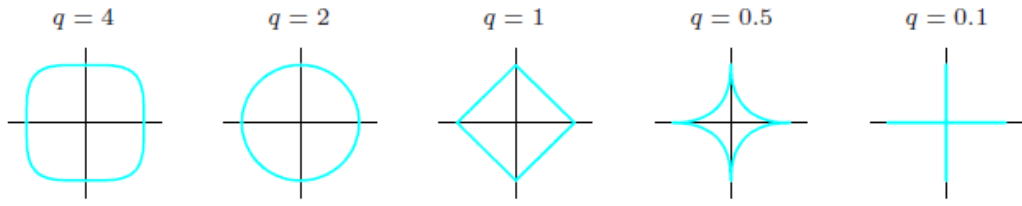
Eşitlik (3.3) ile EKK tahmincisine bir kısıt getirildiği görülebilmektedir. Bu algoritma Lagrange formunda yazıldığında;

$$\min_{\beta} \left\{ \frac{1}{2n} \sum_{i=1}^n \left(y_i - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \lambda \|\beta\|_1 \right\} \quad (3.4)$$

şeklinde elde edilir. Eşitlik (3.4)' de λ ayar parametresidir, lagrange fonksiyonundaki lagrange çarpanı şeklindedir ve daralma miktarını kontrol eder. λ değeri ne kadar artarsa daralma miktarı da o oranda artar. x değişkenleri ölçüm birimlerinden arındırmak için standartlaştırılmakta ve işlem kolaylığı sağlamak amacı ile y değişkenleri merkezleştirilmektedir. Yani, $\frac{1}{n} \sum_{i=1}^n x_{ij} = 0$, $\frac{1}{n} \sum_{i=1}^n x_{ij}^2 = 1$ ve $\frac{1}{n} \sum_{i=1}^n y_i = 0$ alınarak işlemler yapılmaktadır. LASSO tahmincileri bulunurken modeldeki sabit β_0 parametresi başlangıçta ihmal edilmektedir. Diğer parametre tahmin değerleri elde edildikten sonra $\hat{\beta}_0 = \bar{y} - \sum_{j=1}^p \bar{x}_j \hat{\beta}_j$ formülü ile sabit parametre değeri hesaplanmaktadır (Hastie vd, 2015).

LASSO' da $l-1$ normu kullanılmaktadır. $l_q = \|\theta\|_q = \left(\sum_{i=1}^l |\theta_i|^q \right)^{1/q}$, $\theta \in \mathbb{R}^l$ dir ve $q=1$ konveks çözümü verebilen en küçük değerdir. Grafiksel gösterimi Şekil 3.1.' de verilmiştir. q değeri arttıkça değişkenler sıfıra doğru daraltılabilmektedir fakat tam

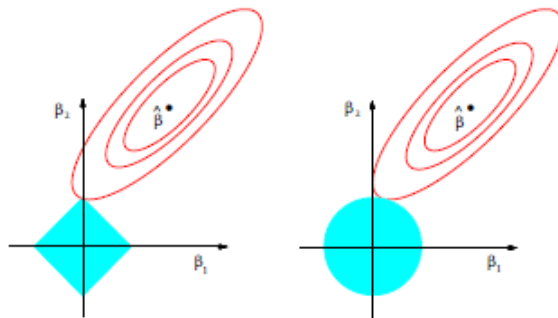
olarak sıfır değeri elde edilememektedir. LASSO ile tam olarak sıfıra daralabilen, dolayısıyla tahmin modeli daha sade olan modeller elde edilebilmektedir. Büyük veri setlerinde, değişken sayısı gözlem sayısından fazla olan modellerde LASSO tahmincisi ile tahmin modelinde önemli olan değişkenlere yer verilerek, diğerleri sıfıra doğru daraltılarak yani modelden çıkarılarak, yorumlanması daha kolay modeller elde edilebilmektedir. Bu sebeple EKK tahmin edicisinin doğruluğunu artırmasına ek olarak, sade (sparse) modeller elde edilmesine imkân sağlamasından dolayı LASSO tercih edilebilmektedir (Zhao ve Yu, 2006; Hastie vd, 2015).



Şekil 3.1. l_q normu için farklı q değerlerinin grafiksel gösterimi (Friedman vd, 2001)

Şekil 3.1.' de düzlem doğrusu üzerinde oluşturulan alanlar model için belirlenen kısıt alanını göstermektedir. EKK teğet doğrusu köşelere temas ettiği noktalarda katsayılar tam olarak sıfır olmaktadır dolayısıyla sadece $q=1$ için yani LASSO'da tam olarak sıfır olan katsayılar elde edilebildiği görülebilmektedir (Friedman vd, 2001).

Aşağıda Şekil 3.2. ile LASSO ve Ridge regresyonlarının tahminleri grafiksel gösterim ile verilmiştir.



Şekil 3.2. LASSO ve Ridge regresyonu tahminleri (Hastie vd, 2015)

Şekil 3.2.' de elipsoid ile gösterilen mavi alan LASSO tahmincisine, daire ile gösterilen mavi alan Ridge tahmincisine aittir. Üst taraftaki kırmızı daireler EKK tahmincinin yayılımını göstermektedir. Bir vektör konveks fonksiyonunu belirli bir

noktada kesiyor ise bu noktada fonksiyon minimum olmaktadır. Dolayısıyla Şekil 3.2. ile LASSO tahmincisi ile tam olarak sıfır olabilen katsayılar elde edilebildiği, Ridge regresyonu ile ise EKK için belirli bir kısıda kadar daralabildiği fakat tam olarak sıfır olabilen değerler elde edilemediği sonucuna ulaşılabilmektedir.

3.2.1. LASSO algoritması çözümü

LASSO algoritmasının çözümü için kullanılan birçok farklı yöntem vardır. Öncelikle LASSO modelinin tek tahmincisi olması durumunda model (3.5)' de verildiği gibidir.

$$\min_{\beta} \left\{ \frac{1}{2n} \sum_{i=1}^n (y_i - x_i \beta_i)^2 + \lambda |\beta| \right\} \quad (3.5)$$

Algoritmanın çözümü için kullanılan en yaygın yöntemlerden birisi, β' ya göre birinci türev alınarak sıfıra eşitlemektir. Bu şekilde parametre tahmini elde edilerek algoritma sonuçlanmaktadır. Ancak burada ceza kısıdında $l-1$ normundan kaynaklı mutlak değerli terim olması sebebi ile türev alma işleminde problem yaşanabilmektedir. Bu problem hafif eşikleme (Soft Thresholding) operatörü ile giderilebilmektedir (Donoho ve Johnstone, 1994). Hafif eşikleme operatörü, EKK tahmincisinin işaretine göre LASSO algoritmasının çözümünde kullanılabilmektedir.

$\hat{\beta}; \beta'$ nın tahmini olmak üzere;

$$\hat{\beta} = \begin{cases} \frac{1}{N} \langle x|y \rangle - \lambda & , \frac{1}{N} \langle x|y \rangle > \lambda \text{ iken} \\ 0 & , \frac{1}{N} \langle x|y \rangle \leq \lambda \text{ iken} \\ \frac{1}{N} \langle x|y \rangle + \lambda & , \frac{1}{N} \langle x|y \rangle < -\lambda \text{ iken} \end{cases} \quad (3.6)$$

Yani $\hat{\beta} = S_{\lambda} \left(\frac{1}{N} \langle x|y \rangle \right)$ dir ve burada hafif eşikleme operatörü; $S_{\lambda}(x) = \text{sign}(x)(|x| - \lambda)_+$ dir. Bu şekilde mutlak değerli terimden dolayı türev alma işleminde yaşanan problem ortadan kaldırılarak LASSO tahmincisi elde edilebilmektedir (Friedman vd, 2010).

LASSO modelinde birden fazla tahminci olduğu durumlarda ise çözüm için Koordinat Azalma yöntemi kullanılabilmektedir. Wu ve Lange, (2008) koordinat azalma yöntemi ile modelde birden fazla tahminci olduğu durumlarda öncelikle tahminciler için herhangi bir başlangıç değeri belirlenmekte, daha sonra bir tahminci seçerek ve diğerlerini geçerli değerlerinde sabit tutarak seçilen tahminci için tek değişkenli LASSO algoritması çözümü ile LASSO tahmini elde edilmektedir. Bu

şekilde tüm koordinatlar (tahminciler) için döngüsel bir şekilde tahmin değerleri elde edilmektedir. Bu yöntem işlem kolaylığı ve hızlı olmasından dolayı yaygın olarak kullanılmaktadır. Örneğin j. adımda β_j ' nin tahmin değeri bulunmak istendiğinde, modelde geriye kalan koordinatlar ($\tilde{\beta}_k, k \neq j$) sabit kabul edilmektedir ve LASSO amaç fonksiyonu;

$$\frac{1}{2n} \sum_{i=1}^n (y_i - \sum_{k \neq j} x_{ik} \tilde{\beta}_k)^2 + \lambda \sum_{k \neq j} |\tilde{\beta}_k| + \lambda |\beta_j| \quad (3.7)$$

olmaktadır. Eşitlik (3.7)' de $r_i^j = y_i - \sum_{k \neq j} x_{ik} \tilde{\beta}_k$; j. koordinat için kısmi rezidüyü vermektedir. Yani fonksiyon (x_{ij}, r_i^j) şekline dönüşmektedir (Friedman vd, 2001).

Ayrıca LARS (En Küçük Açık Regresyonu ve Daraltıcısı) yöntemi (Efron vd, 2004) ve Sıralı (Pathwise) koordinat azalma yöntemi (Friedman vd, 2007) LASSO tahmincilerini elde etmek için kullanılan alternatif yöntemlerdendir. LARS yöntemi farklı alanlarda kullanılarak geliştirilmiştir (Yuan ve Lin, 2006; Yau ve Hui, 2017; Behmanesh ve Saghaei, 2017).

3.2.2. LASSO ayar parametre seçimi

λ ayar parametre seçimi için birçok yöntem bulunmaktadır. Bunlardan model seçiminde ve ayar parametre seçiminde yaygın olarak kullanılan yöntemlerden birisi de Çapraz Doğrulama yöntemidir. Kendi içerisinde üçe ayrılmaktadır.

3.2.2.1. K katlı çapraz doğrulama

Öncelikle veri seti K tane eşit parçaya ayrılmaktadır. Genellikle K, 5 ile 10 arasında seçilmektedir. Ayrılan gruplardan her seferinde bir grup test grubu olarak dışarıda tutulmaktadır ve λ ayar parametresinin farklı değerleri ile geriye kalan K-1 tane çalışma grubu için LASSO tahmin modelleri elde edilmektedir. Elde edilen K-1 tane model ile dışarıda bırakılan test grubu için beklenen tahmin hatası hesaplanmaktadır. Elde edilen değerlerden en küçük tahmin hatalı model seçilerek en doğru model tahminini veren λ değerine ulaşılabilmektedir. Çapraz doğrulama (3.8) nolu eşitlikte verildiği gibi formülize edilmektedir.

$$CV(\lambda) = \frac{1}{n} \sum_{k=1}^K \sum_{i \in F_k} \left(y_i - \hat{f}_{\lambda}^{-k}(x_i) \right)^2 \quad (3.8)$$

Burada $\hat{f}_\lambda^k$; k. grup dışarıda bırakıldığında elde edilen tahmin modelidir ve $\hat{\lambda} = \min_{\lambda} CV(\lambda)$ ile en küçük tahmin hatalı model seçimi ile en doğru ayar parametre değerine ulaşılabilmektedir.

3.2.2.2. Birini dışarıda bırakarak çapraz doğrulama

Bu yöntem K katlı çapraz doğrulamanın farklı bir türüdür denilebilir. Burada her bir gözlem bir grup olarak kabul edilmektedir ve her seferinde bir gözlem dışarıda tutularak K katlı çapraz doğrulama yönteminde olduğu gibi kalan gözlem değerleri için tahmin modelleri elde edilmektedir. Elde edilen modeller için dışarıda bırakılan gözlem değerleri için tahmin hatası hesaplanmaktadır. Bu şekilde tüm gözlemler bir kez dışarıda bırakılarak işlemler devam edilmektedir. Son olarak en küçük tahmin hatalı model seçilerek ayar parametre değerine ulaşılmaktadır. Bu yöntem K katlı çapraz doğrulama yöntemine göre daha maliyetli ve işlem süresi daha uzun olabilmektedir.

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^n \left(y_i - \hat{f}_\lambda^i(x_i) \right)^2 = \frac{1}{n} \sum_{i=1}^n \left[\frac{y_i - \hat{f}(x_i)}{1 - S_{ii}} \right]^2, \hat{y} = Sy \text{ iken} \quad (3.9)$$

$\hat{y} = (\hat{f}(x_1), \dots, \hat{f}(x_n))$ ve S herhangi bir matris olmak üzere ve S_{ii} ; S matrisinin i. köşegen elemanıdır. $\hat{f}_\lambda^i$; i.nci değişken dışarıda bırakıldığında elde edilen tahmin modelidir.

3.2.2.3. Genelleştirilmiş çapraz doğrulama

Genelleştirilmiş çapraz doğrulama, Birini dışarıda bırakarak çapraz doğrulamanın genelleştirilmiş halidir ve daha kullanışlı bir yaklaşım oluşturmaktadır. Aşağıdaki şekilde gösterilmektedir;

$$\frac{1}{n} \sum_{i=1}^n \left[\frac{y_i - \hat{f}(x_i)}{1 - iz(S)/n} \right]^2 \quad (3.10)$$

Eşitlik (3.10)' deki $iz(S)$; S matrisinin izidir yani köşegen elemanlarının toplamıdır.

4. ÇOK DEĞİŞKENLİ BERNOULLI LOJİSTİK MODEL

İstatistiksel analizlerde değişkenler nicel-nitel, sürekli-sürekli ve bağımlı-bağımsız şeklinde ayrılmaktadır. Nicel değişkenler sayılarla-ölçümlerle ifade edilebilen, nitel değişkenler ise ölçümlerle veya sayılarla ifade edilemeyen, grup veya kategori olarak ifade edilebilen değişkenlerdir. Nitel değişkenler (kategorik değişkenler) için farklı analiz yöntemleri mevcuttur. Bunlar veri setine bağlı olarak; eğer bağımlı ve bağımsız değişken olarak ayrımı yapılmamış ise logaritmik doğrusal modeller ile, eğer bir veya daha fazla değişkenin bağımlı (nitel olmalı) olarak ayrımı yapılmak isteniyorsa lojistik regresyon analizi ile veri seti analiz edilebilmektedir.

4.1. Logaritmik Doğrusal ve Lojistik Modeller

Logaritmik doğrusal modeller kategorik veri analizinde kullanılan çok değişkenli istatistiksel yöntemlerden biridir. Satır ve sütun sayısı fazla olan çapraz tablolarda analiz yaparken kolaylık sağladığı için yaygın olarak tercih edilmektedir. Bu modellerde değişkenler arasında bağımlı-bağımsız ayrımı yapılmamaktadır. Dolayısıyla amaç; değişkenler arasındaki ilişkilerin incelenmesidir. Logaritmik doğrusal modeller; en çok olabilirlik tahminlerine (EÇOT) ve olabilirlik oran (G^2) testlerine dayanmaktadır (Bülbül, 2006).

Diğer taraftan lojistik regresyon analizi; değişkenler arasında bir veya birden fazlasının bağımlı-bağımsız ayrımı yapılmak istendiğinde ve bağımlı değişkenlerin nitel olduğu, bağımsız değişkenlerin nitel veya nicel olabildiği durumlarda kullanılabilir. Ayrıca lojistik regresyonda doğrusal regresyon analizindeki varsayımlar aranmadığı için yaygın olarak tercih edilmektedir. Burada amaç; bağımlı değişkeni tahmin edebilecek en sade modeli bulmaktır. Özellikle tıp alanı başta olmak üzere sosyo-ekonomik, askeri, biyoloji, ziraat ve eğitim alanlarında yaygın olarak kullanılmaktadır. Lojistik model kendi içerisinde bağımlı değişkendeki kategori sayısına bağlı olarak üç gruba ayrılmaktadır; İkili (binary), Sıralı (ordinal), İsimli (multinomial) lojistik modeller şeklindedir (Çokluk, 2010).

4.2. Bernoulli Dağılımı

Bernoulli dağılımı, kesikli bir denemenin sonuçları başarılı-başarısız veya olumlu-olumsuz birbirleriyle bağdaşmayan (0-1) iki sonuçlu olaylarda kullanılmaktadır. Bernoulli dağılımının olasılık yoğunluk fonksiyonu;

$$f_Y(y) = p^y (1-p)^{1-y} \quad (4.1)$$

şeklinde. Y değişkeni 0-1 değerlerini almaktadır ve burada p yapılan denemede başarılı gelme olasılığını göstermektedir. (4.1)' de verilen durum tek değişken olduğu durum içindir.

İki değişkenli olduğu durumlarda Bernoulli dağılımının olasılık yoğunluk fonksiyonu aşağıdaki şekilde olmaktadır;

$$P(Y=y) = p(y_1, y_2) = p_{11}^{y_1 y_2} p_{10}^{y_1 (1-y_2)} p_{01}^{(1-y_1) y_2} p_{00}^{(1-y_1)(1-y_2)} \quad (4.2)$$

Bir olasılık fonksiyonun üstel dağılım ailesi şeklinde yazılabilmesi, üstel dağılım ailesi fonksiyonunda bulunan b fonksiyonu kaynaklı birinci türev ve ikinci türev alınmasında beklenen değer ve varyans değerlerinin elde edilmesinde işlem kolaylığı sağlamasından dolayı avantaj sağlamaktadır. (4.2) fonksiyonu üstel dağılım ailesi şeklinde yazıldığında;

$$p(y_1, y_2) = \exp \left\{ \log(p_{00}) + y_1 \log \left(\frac{p_{10}}{p_{00}} \right) + y_2 \log \left(\frac{p_{01}}{p_{00}} \right) + y_1 y_2 \log \left(\frac{p_{11} p_{00}}{p_{10} p_{01}} \right) \right\} \quad (4.3)$$

olmaktadır. (4.3) gösterimi sadeleştirildiğinde,

$$p(y_1, y_2) = \exp \{ \log(p_{00}) + y_1 f^1 + y_2 f^2 + y_1 y_2 f^{12} \} \quad (4.4)$$

elde edilmektedir. Burada f^1 ; birinci değişkenin modele etkisi ve f^{12} ; birinci ve ikinci değişkenin arasındaki etkileşimin modele etkisi şeklinde ifade edilebilir. Modelde k bağımlı değişken sayısı ve p bağımsız değişken sayısı iken $2^k - 1$ tane tahmin edilen katsayı vektörü bulunacaktır ve toplam parametre sayısı sabit katsayı (β_0) dâhil $(p+1) \times (2^k - 1)$ dir (Dai, 2013).

(4.4)' te iki değişkenli Bernoulli dağılımı üstel dağılım ailesi üyesi olarak gösterildiğinde bir log-lineer model ile benzer formda olduğu görülebilmektedir.

(4.4)' teki fonksiyonun çözümü için logaritmik işlemlerde kolaylık sağlamasından dolayı EÇOT yöntemi daha yaygın olarak tercih edilebilmektedir.

(4.4)' deki iki değişkenli Bernoulli dağılımı için negatif log-olabilirlik fonksiyonu;

$$l(y,f) = -\frac{1}{n} \sum_{i=1}^n [y_1(i)f^1(x_i) + y_2(i)f^2(x_i) + y_1(i)y_2(i)f^{12}(x_i) - b(f(x_i))] \quad (4.5)$$

ve burada $b(f(x_i)) = \log[1 + \sum_{K=1,2,12} \exp(f^K)]$ dir. - $b(f)$ fonksiyonu modeldeki sabit terimin ihmal edildiğinin bir göstergesidir. Modeldeki f fonksiyonları; genelleştirilmiş lineer modellerdeki, bağımsız değişkenler ile bilinmeyen beta katsayılarının bir kombinasyonu olan ve bağımsız değişkenler hakkındaki bilgiyi modele dâhil eden lineer öngörücüler şeklindedirler.

Burada log-olabilirlik fonksiyonu iki değişkenli Bernoulli dağılımı üzerinden gösterilmiştir fakat değişken sayısı ikiden daha fazla da olabilmektedir yani Bernoulli değişken sayısı artırılabilir.

Bernoulli dağılımı çok değişkenli Bernoulli dağılımına genişlettiğinde elde edilen sonuçlar lojistik regresyon modeli ile benzer olduğu sonucuna varılmıştır. Bu sebeple çok değişkenli lojistik modelde bağımlı değişken 0-1 şeklinde ikili (binary) değerler aldığında bu model için çok değişkenli Bernoulli lojistik regresyon modeli ifadesi kullanılmıştır (Dai, 2012).

4.3. İki Değişkenli Bernoulli Lojistik LASSO Model

Eşitlik (3.2)' de verilen fonksiyonda LASSO ceza kısıdı EKK tahmincisine eklenmiş hali verilmişti. Burada ise LASSO ceza kısıdını iki değişkenli Bernoulli lojistik modelin negatif olabilirlik fonksiyonuna ekleyerek iki değişkenli Bernoulli lojistik LASSO modelinin amaç fonksiyonu elde edilebilmektedir. Hedef fonksiyon;

$$T_\lambda(y,f) = l(y,f) + J_\lambda(f) \quad (4.6)$$

şeklindedir ve burada $l(y,f)$; negatif log-olabilirlik fonksiyonu, $J_\lambda(f)$; LASSO ceza kısıdıdır. İki değişkenli durum için LASSO kriteri;

$$J_\lambda(f) = \lambda_1 \sum_{j=1}^p |c_j^1| + \lambda_2 \sum_{j=1}^p |c_j^2| + \lambda_{12} \sum_{j=1}^p |c_j^{12}| \quad (4.7)$$

olmaktadır. Burada bağımlı değişken sayısı 2 olduğu için 3 adet λ vardır. Örneğin bağımlı değişken sayısı 3 iken ceza fonksiyonunda 2^3-1 yani 7 adet λ değeri bulunacaktır $(\lambda_1, \lambda_2, \lambda_3, \lambda_{12}, \lambda_{13}, \lambda_{23}, \lambda_{123})$.

Bu amaç fonksiyon minimize edilerek iki deęişkenli Bernoulli lojistik LASSO model için LASSO tahminleri elde edilebilmektedir (Dai, 2013). Benzer şekilde Bernoulli deęişken sayısı artırılarak işlemler yapılabilmektedir.

4.4. Model Uyum İstatistikleri

Tahmin sonuçlarının elde edilmesinde kullanılan farklı uyum iyilięi ölçütleri bulunmaktadır. Çalışmada bunlardan; Akaike bilgi kriteri (AIC), Bayesci bilgi kriteri (BIC), Genelleştirilmiş çapraz doğrulama yaklaşım (GACV) kriteri ve Bayesci genelleştirilmiş çapraz doğrulama yaklaşım (BGACV) kriteri üzerinde durulmaktadır.

4.4.1. Akaike bilgi kriteri

Akaike Bilgi Kriteri; ilk olarak Akaike (1973) tarafından tanıtılmıştır. Birden fazla iyi model olduğunda aralarından en uygun modelin seçiminde kullanılan bir ölçüttür. Genellikle AIC ile gösterilmektedir. Parametre sayısına ve hata kareler toplamına dayalıdır.

$$AIC = n \ln \left(\frac{HKT}{n} \right) + 2k \quad (4.8)$$

şeklinde formülize edilebilir. Farklı gösterimleri de bulunmaktadır. Burada;

k = Parametre sayısı

n = Modeldeki gözlem sayısı

HKT = Hata kareler toplamı

dır. Modeller arasından AIC değeri en küçük olan model en uygun model olarak seçilmektedir (Akaike, 1974).

Modele yeni bağımsız deęişken ilave edildikten sonra HKT daki azalma, AIC kriterinde düşüşe sebep olmadığı sürece modele deęişken ilave edilmesi gerekli deęildir.

4.4.2. Bayesci bilgi kriteri

Bayesci Bilgi Kriteri; ilk olarak Schwarz (1978) tarafından ortaya atılmıştır. Bu sebeple Schwarz kriteri olarak da bilinmektedir. Genellikle BIC ile gösterilmektedir.

$$BIC = -2 \ln(L) + k \log(n) \quad (4.9)$$

olarak elde edilmektedir. Burada;

L = Olabilirlik fonksiyonu

k = Parametre sayısı

n = Gözlem sayısı

dır. AIC gibi en uygun modelin seçiminde kullanılmaktadır. Modeller arasından BIC değeri en düşük olan model tercih edilmektedir (Wasserman, 2000).

Akaike bilgi kriteri ile ilişkilidir. Aralarındaki fark; ceza kısıtları farklıdır. BIC daha fazla kısıt getirir. Bu sebeple AIC'dan daha sade modeller seçme eğilimindedir.

4.4.3. Genelleştirilmiş çapraz doğrulama yaklaşım kriteri

Çapraz doğrulama yöntemi, model seçimi ve ayar parametre seçimi için kullanılan yaygın yöntemlerden biridir (Wahba ve Wold, 1975). Çapraz doğrulama yaklaşımı kendi içine üç alt gruba ayrılmaktadır; K katlı çapraz doğrulama, Biri dışarıda bırakılarak çapraz doğrulama ve Genelleştirilmiş çapraz doğrulama (Wahba ve Craven, 1978; Golub vd, 1979).

GACV; Biri dışarıda bırakılarak çapraz doğrulamanın farklı bir yaklaşımıdır ve ikili veri setlerinde parametre seçiminde kullanılmaktadır (Xiang ve Wahba, 1996). Kullback-Leibler (KL) uzaklığını minimize etmek için kullanılabilmektedir. Zaman içerisinde uygulama alanları genişletilmiştir (Wahba, 1994; Xiang ve Wahba, 1997; Wahba vd, 1999; Lin vd, 2000; Gu ve Xiang, 2001).

KL uzaklığı; iki olasılık fonksiyon arasındaki uzaklığı ölçmek için kullanılmaktadır. Uzaklık; gerçek bir uzaklık değildir, doğru model ile tahmin edilen model arasındaki bilgi farkını belirtmektedir. Genel gösterimi; P ve Q iki olasılık fonksiyonu olmak üzere; $\sum P(i) \log \frac{P(i)}{Q(i)}$ şeklindedir. Bernoulli dağılımı için KL uzaklığının λ ayar parametresi ile ilgili kısmı;

$$CKL(\lambda) = \frac{1}{n} \sum_{i=1}^n \left[-E \left(y_i f_{\lambda}(x_i) + b \left(f_{\lambda}(x_i) \right) \right) \right] \quad (4.10)$$

olmaktadır. Burada $b(f_{\lambda}) = \log(1 + e^f)$ olarak ifade edilir ve KL uzaklığı için Biri dışarıda bırakılarak çapraz doğrulama kriteri;

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^n \left[-y_i f_{\lambda}^i(x_i) + b(f_{\lambda}(x_i)) \right] \quad (4.11)$$

şeklindedir. Eşitlik (4.11)' de $f_{\lambda}^i(x_i)$; i.nci gözlem dışarıda bırakıldığında elde edilen tahmin modelidir. Bu yöntem ile her bir gözlem bir grup olarak ele alındığı için büyük veri setlerinde zaman kısıdı ve fazla maliyetli olması sebebi ile bu yöntem için birinci ve ikinci dereceden Taylor seri yaklaşımı oluşturulmaktadır (Shi vd, 2006).

Taylor serisi; fonksiyonların verilen bir noktadaki sayısal değerlerini bulmak için kullanılabilir. Ek olarak, türev ya da integral işlemlerinde daha kolay işlem yapılabilmesine imkân vermektedir.

Biri dışarıda bırakılarak çapraz doğrulama için birinci ve ikinci dereceden Taylor serisi yaklaşımı ile;

$$ACV(\lambda) = \frac{1}{n} \sum_{i=1}^n \left[-y_i f_{\lambda}(x_i) + b(f_{\lambda}(x_i)) \right] + \frac{1}{n} \sum_{i=1}^n h_{ii} \frac{y_i (y_i - \mu_{\lambda}(x_i))}{1 - \sigma_{\lambda i}^2 h_{ii}} \quad (4.12)$$

elde edilmektedir. Burada h_{ii} ; Taylor serisi yaklaşımı sonucu elde edilen H matrisinin i.nci köşegen elemanıdır. Eşitlikte $\sigma_{\lambda i}^2 = p_{\lambda}(x_i) (1 - p_{\lambda}(x_i))$ üstel dağılım ailesi fonksiyonundaki $b(f)$ fonksiyonun ikinci türevine karşılık gelmektedir ve $\mu_{\lambda}(x_i)$ ise $b(f)$ fonksiyonunun birinci türevine karşılık gelmektedir.

Biri dışarıda bırakılarak çapraz doğrulamaya Taylor serisi ile yakınsama sonucu elde edilen model geliştirildiğinde;

$$GACV(\lambda) = \frac{1}{n} \sum_{i=1}^n \left[-y_i f_{\lambda} + b(f_{\lambda}(x_i)) \right] + \frac{tr(H)}{n} \frac{y_i (y_i - \mu_{\lambda}(x_i))}{tr(I - W^{1/2} H W^{1/2})} \quad (4.13)$$

modeli elde edilmektedir. $iz(H)$; H matrisinin izidir yani köşegen elemanlarının toplamıdır ve W; ii.nci elemanı $\sigma_{\lambda i}^2$ olan köşegen matristir.

(4.12) ve (4.13)' de verilen modellerde gösterildiği şekilde dönüşümler yapılarak GACV kriteri elde edilmektedir (Zhang vd, 2004).

4.4.4. Bayesci geliştirilmiş çapraz doğrulama yaklaşım kriteri

BGACV kriteri Eşitlik (4.13)' de verilen GACV kriterinin Bayes ceza kısıdı eklenmiş versiyonudur. Ceza kısıdı GACV ye oranla daha fazla olduğu için modele alınacak değişken sayısında daha fazla seçici olmaktadır.

$$GACV(\lambda) = \frac{1}{n} \sum_{i=1}^n \left[-y_i f_{\lambda} + b \left(f_{\lambda}(x_i) \right) \right] + \frac{\text{tr}(H)}{n} \frac{\log(n)}{2} \frac{y_i (y_i - \mu_{\lambda}(x_i))}{\text{tr}(I - W^{1/2} H W^{1/2})} \quad (4.14)$$

şeklindedir (Dai, 2012).

5. BULGULAR VE TARTIŞMA

Bu çalışmada, R programında MVB paketi kullanılarak farklı simülasyon çalışmaları yapılmıştır. Gözlem sayısının 50, 100, 200, 250 ve 300 olduğu durumlarda, bağımlı değişken sayısı 2 ile 3 ve bağımsız değişken sayısı 5 olarak belirlenerek simülasyonlar yapılmıştır.

Ayrıca bağımlı değişken sayısı 2 değeri için sabit tutularak gözlem sayısı 50, 40 ve 30 için farklı bağımsız değişken sayıları ile AIC, BIC, GACV ve BGACV kriterleri ile LASSO tahminleri elde etmek için tekrarlı simülasyonlar yapılmıştır.

n = Gözlem sayısı, p = Bağımsız değişken sayısı, k = Bağımlı değişken sayısıdır ve başlangıç beta değerleri Çizelge 5.1. ' e göre belirlenmiştir.

Çizelge 5.1. Bağımlı değişken sayısı 2 iken başlangıç beta değerleri

	T=1	T=2	T=12
C1	1,5	0	0
C2	0	-1,5	0
C3	0	0	0
C4	0	0	0
C5	0	0	1

Çizelge 5.1. ' de $c^T = (c_0^T, c_1^T, \dots, c_p^T)$; tahmin edilen katsayı vektörüdür ve beta değerleri; $\beta_0 = (c_0^1, c_0^2, c_0^{12})$, $\beta_1 = (c_1^1, c_1^2, c_1^{12})$, ..., $\beta_5 = (c_5^1, c_5^2, c_5^{12})$ şeklindedir. Model tahmini genelleştirilmiş lineer modellerde lineer öngörücüler şeklinde bağımsız değişkenler ile bilinmeyen parametrelerin lineer kombinasyonu şeklinde elde edilmektedir. Örnek olarak, $n=50$, $p=5$ ve $k=2$ için belirlenen başlangıç beta değerleri için AIC, BIC, GACV ve BGACV uyum iyiliği ölçütleri ile elde edilen LASSO tahmin değerleri (sabit beta değeri; β_0 hariç) Çizelge 5.2. ' de verilmiştir.

Çizelge 5.2. $n=50$, $p=5$ ve $k=2$ için uyum iyiliği ölçütleri ile elde edilen beta tahminleri

	AIC VE BIC			GACV VE BGACV		
	T=1	T=2	T=12	T=1	T=2	T=12
C1	0,0968	0,0000	0,0000	0,1028	0,0000	0,0000
C2	0,0000	-0,3250	0,0000	0,0000	-0,3987	0,0000
C3	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
C4	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
C5	0,0000	0,0000	0,1388	0,0000	0,0000	0,0542

LASSO' da amaç; başlangıç beta değerlerinde bulunan sıfır değerlerini yine sıfır olarak, sıfırdan farklı değerleri ise sıfıra doğru daraltılarak en doğru tahmin modelini elde etmektir. Çizelge 5.2. ' den görüldüğü üzere, AIC ve BIC tahmini beta değerleri aynı ve başlangıç beta değerlerinde sıfır olan hücre değerleri yine sıfır olarak elde edilmiştir, diğer taraftan GACV ve BGACV ile elde edilen tahmini beta değerlerinin aynı sonuçları verdiği ve burada da başlangıç beta değerlerinde sıfır olan hücre değerlerinin yine sıfır olarak elde edildiği görülebilmektedir. Elde edilen LASSO tahmin sonuçlarına bakıldığında, tahmin değerlerinde farklılıklar görülmektedir. Birinci satırın ilk hücresine bakıldığında GACV ve BGACV ile elde edilen tahmin değerinin daha büyük olduğu gözlenmektedir. İkinci satırın ikinci hücresine bakıldığında ise negatif belirlenen beta değeri için (-1 için) AIC ve BIC uyum iyiliği ölçütü ile daha büyük LASSO tahmin değeri elde edildiği ve başlangıç beta değerine daha yakın tahmin sonucu elde edildiği görülebilmektedir. Son olarak, beşinci satırın üçüncü hücresine bakıldığında; AIC ve BIC uyum iyiliği ölçütleri ile elde edilen LASSO tahmin değerinin daha büyük olduğu ve başlangıç beta değerine daha yakın bir değer aldığı görülebilmektedir.

Bir diğer örnek olarak, $n=300$, $p=5$, $k=2$ için ve Çizelge 5.1. ' deki başlangıç beta değerleri kullanılarak 4 farklı uyum iyiliği ölçütü için elde edilen tahmini beta değerleri aşağıda Çizelge 5.3. ' de verilmiştir.

Çizelge 5.3. $n=300$, $p=5$ ve $k=2$ için uyum iyiliği ölçütleri ile elde edilen beta tahminleri

	AIC			BIC		
	T=1	T=2	T=12	T=1	T=2	T=12
C1	1,2527	0,0000	0,0000	0,8973	0,0000	0,0000
C2	-0,0101	-1,2344	0,0000	0,0000	-1,0370	0,0000
C3	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
C4	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
C5	0,0000	0,0000	0,5502	0,0000	0,0000	0,2732
	GACV			BGACV		
	T=1	T=2	T=12	T=1	T=2	T=12
C1	0,9980	0,0000	0,0000	1,2501	0,0000	0,0000
C2	0,0000	-1,0472	0,0000	0,0000	-1,3762	0,0000
C3	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
C4	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
C5	0,0000	0,0000	0,4460	0,0000	0,0000	0,6634

Bu örnekte ise elde edilen tahminler arasında farklılıklar görülmektedir. Örneğin, AIC kriteri sonuçlarına bakıldığında ikinci satırın birinci hücresinde başlangıç değeri sıfır olmasına rağmen tahmin değeri sıfırdan farklı olarak elde edilmiştir. Diğer kriterler için bu hücre sıfır olarak yani başlangıç değeri ile aynı olacak şekilde elde edilmiştir. Diğer elde edilen tahmin değerleri incelendiğinde; belirli bir kararlılık görülmemesiyle birlikte başlangıç beta değerlerine en yakın tahminlerin GACV uyum iyiliği ölçütü ile elde edildiği görülebilmektedir. Çizelge 5.2. ve Çizelge 5.3. beraber incelendiğinde; gözlem sayısında değişiklik (artış) olduğunda elde edilen LASSO tahmin sonuçlarında farklılık olabileceği sonucuna ulaşılabilmektedir.

Farklı gözlem sayısı ve bağımlı değişken sayıları için yapılan simülasyonlar için sonuçlar Çizelge 5.4. ' de verilmiştir. Elde edilen sonuçlar 100 tekrarlı denemelerde elde edilen tahmin değerlerinin her bir kriter için yüzdelik değerini göstermektedir. Yani LASSO tahmincisi ile elde edilen tahmini beta değerlerinin başlangıç beta değerlerine yüzdelik olarak yakınlığını belirtmektedir. Fark değerleri ise başlangıç modeli ile tahmin modeli arasındaki farkları göstermektedir. Fark değerlerinin düşük olması beklenmektedir.

Çizelge 5.4. ' de belirtilen FARK1; tarafımızdan belirlenen başlangıç beta değerleri ile AIC kriteri için elde edilen LASSO tahminleri arasındaki fark değerini göstermektedir. Aynı şekilde olmak üzere FARK2; BIC kriteri için, FARK3; GACV kriteri için ve FARK4; BGACV kriteri için elde edilen farkları belirtmektedir.

Çizelge 5.4. $p=5$, $k=2, 3$, $n=50, 100, 200, 250$ ve 300 iken elde edilen sonuçlar

n	p	k	AIC	BIC	GACV	BGACV	FARK1	FARK2	FARK3	FARK4
50	5	2	84,14634	93,9024	87,60163	88,61789	0,1710	0,1747	0,1578	0,1627
		3	93,87097	96,58065	92,58065	94,87097	0,1071	0,1054	0,1275	0,1144
100	5	2	85,25000	95,75000	89,16667	92,58333	0,1551	0,1555	0,1315	0,1387
		3	94,61290	97,64516	92,87097	94,51613	0,0857	0,0898	0,0948	0,0891
200	5	2	86,83333	95,58333	90,33333	90,75000	0,1264	0,1216	0,1169	0,1183
		3	95,37634	98,60215	96,98925	97,95699	0,0910	0,0799	0,0718	0,0754
250	5	2	84,25000	95,58333	94,25000	96,75000	0,1173	0,1143	0,1129	0,1089
		3	92,41935	97,74194	97,41935	99,67742	0,0901	0,0724	0,0633	0,0789
300	5	2	80,75000	95,83333	95,16667	96,33333	0,1073	0,1165	0,0962	0,1030
		3	93,11828	97,63441	99,24731	99,46237	0,0708	0,0763	0,0775	0,0738

Sonuçlardan, bağımsız değişken sayısı 5 ve gözlem sayısı 50, 100 ve 200 için bağımlı değişken sayısı 2 ile 3 alındığında yüzdelik olarak en büyük değeri BIC

kriteri vermiştir. Yani bu gözlem sayıları için en iyi modeli veren en güçlü uyum iyiliği ölçütünün BIC kriteri olduğu sonucuna ulaşılabilmektedir.

Diğer taraftan bağımsız değişken sayısı 5 ve gözlem sayısı 250 ile 300 için bağımlı değişken sayısı 2 ile 3 alındığında ise yüzdelik olarak en büyük değeri BGACV kriteri vermiştir.

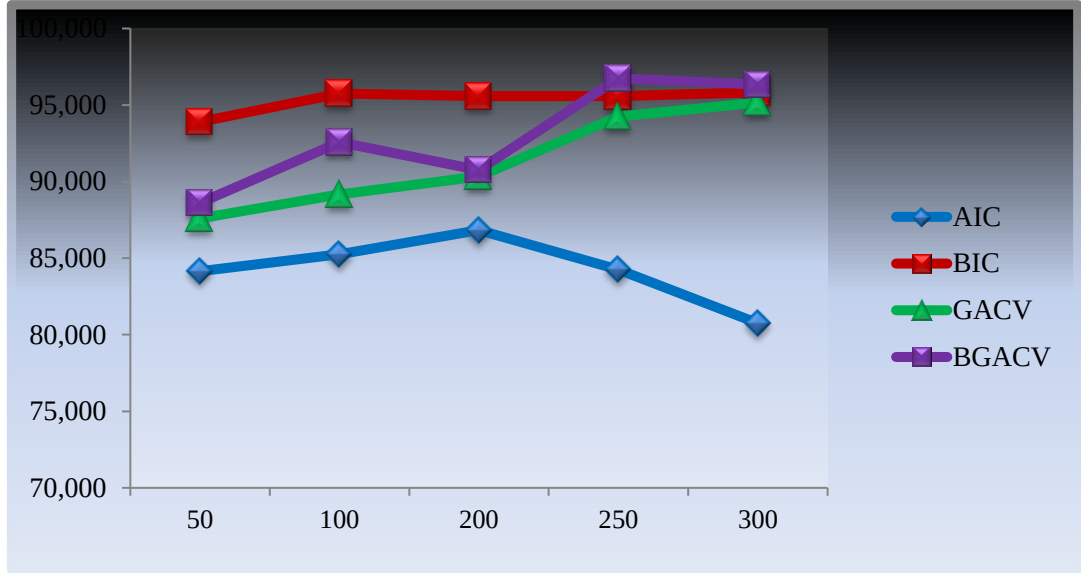
Gözlem sayısının artması ile birlikte en iyi tahmin modelini veren uyum ölçütlerinin farklılık gösterdiği Çizelge 5.4.'den elde edilen sonuçlardan görülebilmektedir.

Belirtildiği gibi fark değerleri; belirlenen beta değerleri ile AIC, BIC, GACV ve BGACV uyum iyiliği ölçütleri için elde edilen LASSO tahmin değerleri arasındaki değerlerini göstermektedir. Çizelge 5.4.'de ki fark değerleri incelendiğinde farklılık görülmektedir fakat genel duruma bakıldığında GACV kriteri ile en düşük farklar elde edilmektedir yani başlangıç modeline en yakın tahmin değerleri GACV kriteri ile elde edilebilmektedir.

Çalışmada LASSO tahmincilerinin özellikle büyük veri setlerinde tercih edilmelerinden dolayı gözlem sayısı artırılarak uyum iyiliği ölçütleri incelenmek amaçlanmıştır. Fakat zaman kısıdı sebebi ile gözlem sayıları yapılan uygulamadaki şekli ile incelenebilmiştir. LASSO tahmincilerinin büyük veri setlerinde ve özellikle bağımsız değişken sayısının gözlem sayısından fazla olduğu veri setlerinde analizlerde işlem kolaylığı sağlaması ve büyük veri setlerinde görülen bazı problemlerin giderilmesinde katkısı olmasından dolayı tercih edilmektedir. Bu sebeple aşağıda ilerleyen bölümlerde değişken sayısının gözlem sayısından büyük olduğu veri setleri hazırlanarak en güçlü tahmin modellerinin elde edilmesinde AIC, BIC, GACV ve BGACV uyum iyiliği ölçütlerinin incelenmesi için uygulamalar ve sonuçları verilmiştir.

Fakat öncelikle Çizelge 5.4. ile verilen simülasyon sonuçları grafiksel olarak incelenmek için aşağıda bağımlı değişken sayısındaki farklılıklara göre iki farklı grafik ve yorumları verilmiştir.

Çizelge 5.4. değerlerinden bağımsız değişken sayısı 5, bağımlı değişken sayısı 2 değeri sabit olarak alındığında gözlem sayısı 50, 100, 200, 250 ve 300 için elde edilen sonuçların grafiksel gösterimi Şekil 5.1.' de verilmiştir.



Şekil 5.1. $p=5$ ve $k=2$ değerleri sabit iken $n=50, 100, 200, 250$ ve 300 için kriterler

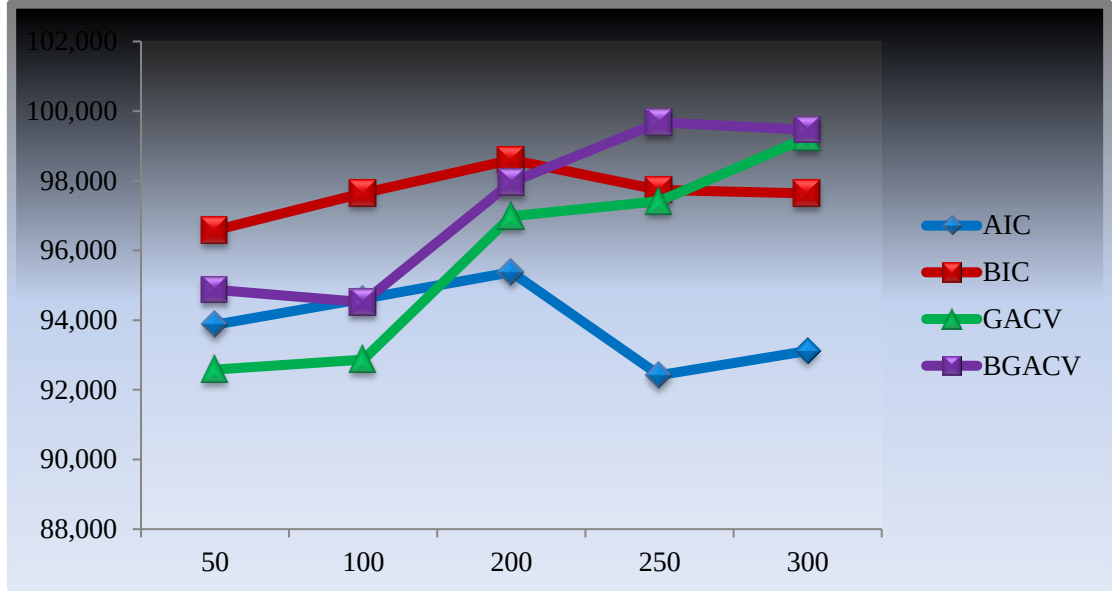
Grafikte, AIC kriteri için; gözlem sayısı 50 ile 200 arasında artış gösterirken daha sonra gözlem sayısı 200 ile 300 arasında artmasına rağmen AIC eğrisi azalış göstermektedir.

BIC kriteri için; gözlem sayısı 50 ile 200 arasında çok az bir artış göstermektedir ve gözlem sayısı 200 ile 300 arasında belirli bir miktar azalış göstermektedir. Farklı gözlem değerleri arasında farklı grafik çizmesine rağmen genel olarak BIC kriteri belirli bir artış göstermektedir. Genel olarak bakıldığında gözlem sayısındaki artışa paralel olarak BIC kriteri eğrisi artış göstermektedir.

GACV kriteri için; gözlem sayısındaki artış ile birlikte genel olarak sürekli bir artış gözlenmektedir. Gözlem sayısı 50 ile 200 arasında artış göstermektedir ve gözlem sayısı 200 ile 300 arasında artış eğrisinde yükseliş görülmektedir. Yani gözlem sayısının artış miktarı ile doğru orantılı olarak grafik çizmektedir.

Son olarak BGACV kriterine bakıldığında; GACV kriterine oranla daha fazla düzensizlik göstermesine rağmen, BGACV kriteri incelendiğinde gözlem sayısındaki artış ile birlikte dalgalanmalar göstermesine rağmen genel olarak artış görülmektedir.

Diğer bir grafiksel gösterim için, simülasyon çalışmaları özet tablo değerlerinden bağımsız değişken sayısı 5, bağımlı değişken sayısı 3 değeri sabit olarak alındığında gözlem sayısı 50, 100, 200, 250 ve 300 için grafiksel gösterimi Şekil 5.2.' de verilmiştir.



Şekil 5.2. $p=5$ ve $k=3$ değerleri sabit iken $n=50, 100, 200, 250$ ve 300 için kriterler

Grafikte, AIC kriteri için; gözlem sayısı 200 ' e kadar Şekil 5.1. ile benzer olarak artış göstermektedir fakat Şekil 5.1.' den farklı olarak gözlem sayısı 250 ile 300 arasında azalış değil artış görülmektedir. Yani bağımlı değişken sayısının artırılması ile AIC kriterinde gözlem sayısı arttığı durumlarda değişimler olabildiği görülmektedir.

BIC kriteri için; gözlem sayısı 200 ' den sonra Şekil 5.1. ' de göstermiş olduğu düşüşten farklı olarak azalma eğrisinin arttığı görülmektedir. Burada bağımlı değişken sayısının artırılması BIC kriterinde değişime neden olduğu sonucuna ulaşılabilmektedir.

GACV uyum iyiliği ölçütü için; bağımlı değişken sayısı 2 olduğu duruma kıyasla düzensizliğinde artış görülmesine rağmen genel olarak artış gözlemlenmektedir.

Ve BGACV uyum iyiliği ölçütü incelendiğinde; Şekil 5.1. ile kıyaslandığında yani bağımlı değişken sayısında farklılık (artış) olması düzensizlik göstermesinde bir katkısı olmasına rağmen genel olarak artış gözlemlenmektedir.

LASSO tahmincisi, özellikle deęişken sayısının gözlem sayısından fazla olduęu durumlarda birçok parametreyi sıfıra daraltarak yorumlanması kolay modeller elde edilmesinde katkısı olmasından dolayı tercih edilmektedir. Benzer veri seti için AIC, BIC, GACV ve BGACV uyum iyilięi ölçütleri ile elde edilen LASSO tahminlerini incelemek amacıyla R programında MVB paketinden yararlanılarak tekrarlı simülasyonlar yapılmıştır.

İlk olarak, bağımlı deęişken sayısı 2 için sabit tutularak, gözlem sayısı 50 iken bağımsız deęişken sayıları 5, 15, 20, 25, 50 ve 55 olarak belirlenerek kriterler için tahmin deęerleri elde edilmiştir. Elde edilen deęerler Çizelge 5.5.' de verilmiştir.

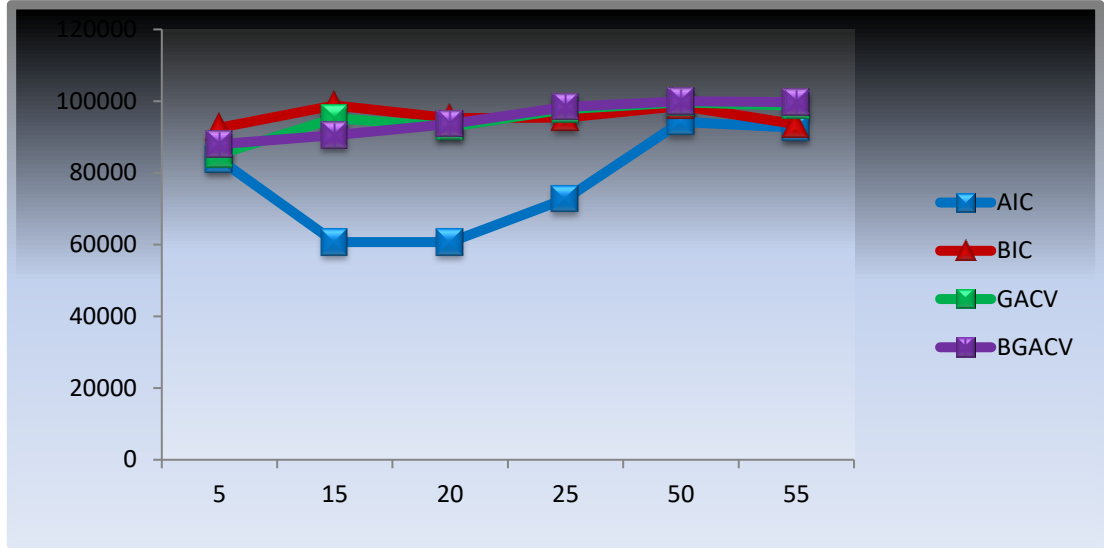
Çizelge 5.5. $k=2$, $n=50$ ve $p=5, 15, 20, 25, 50$ ve 55 iken kriterler

Bağımlı deę.say.(k)	Gözlem say. (n)	Deęişken say. (p)	AIC	BIC	GACV	BGACV
2	50	5	83,7500	92,5833	85,1666	88,0000
2	50	15	60,71429	98,80952	95,23810	90,47619
2	50	20	60,70175	95,26316	92,80702	93,68421
2	50	25	72,77778	95,41667	97,77778	98,33333
2	50	50	94,21769	98,63946	99,65986	100,0000
2	50	55	92,59259	93,51852	98,76543	99,69136

Çizelge 5.5. incelendięinde elde edilen sonuçlardan, gözlem sayısı 50 iken bağımsız deęişken sayıları 5, 15 ve 20 için yüzdelik olarak en yüksek sonuçları BIC uyum iyilięi ölçütünün verdięi görülürken dięer taraftan bağımsız deęişken sayısı 25, 50 ve 55 için BGACV uyum iyilięi ölçütü ile en güçlü tahmin sonuçlarının elde edildięi gözlenmektedir.

Elde edilen bilgiler ışığında gözlem sayısı 50 için deęişken sayısı 25 ve üzeri deęerleri ile oluşturulan veri setlerinde BGACV kriteri ile daha kararlı ve güvenilir tahmin sonuçları elde edilebileceęi sonucuna ulaşılabilmektedir. Daha önce de belirtildięi gibi LASSO tahmincileri çoğunlukla deęişken sayısının gözlem sayısından fazla olduęu ($p > n$) veri setleri üzerinde tercih edildięi göz önünde bulundurulursa $p=55$ ($p > 50$) için BGACV uyum iyilięi ölçütü ile daha güçlü tahminler elde edildięi görülmektedir.

Çizelge 5.5. ile verilen bağımlı değişken sayısı 2, gözlem sayısı 50, bağımsız değişken sayıları 5, 15, 20, 25, 50 ve 55 iken elde edilen analiz sonuçları için aşağıda Şekil 5.3. ile grafiksel gösterimi verilmiştir.



Şekil 5.3. $k=2$ ve $n=50$ iken $p=5, 15, 20, 25, 50$ ve 55 için kriterler

Grafikte, AIC kriteri için; değişken sayısı 5 ile 15 arasında iken AIC eğrisinde azalış olduğu, daha sonra 15 ile 50 arasında değişken sayısı artarken AIC kriterinde de artış olduğu ve daha sonra değişken sayısı 50 ile 55 arasında iken AIC kriterinin sabite yakın bir grafik çizdiği görülmektedir. Yani değişken sayısındaki düzenli artış olmasına rağmen AIC kriterinde dalgalanmalar görülmekle birlikte genel olarak bakıldığında değişken sayısının artması ile birlikte artış gösterdiği sonucuna ulaşılabilir.

BIC kriteri; değişken sayısındaki artış ile orantılı olarak düzenli bir artış göstermemesine rağmen AIC kriteri ile kıyaslandığında daha tutarlı olduğu görülebilmektedir ve genel olarak sabit bir grafik çizmek ile birlikte az da olsa dalgalanmalar görülebilmektedir.

GACV kriteri için; değişken sayısı 5 ile 20 arasında belirli bir miktar dalgalanma görülmesine rağmen değişken sayısı 20 ile 55 arasında artan ve sabite yakın bir eğim görülmektedir. AIC ve BIC kriterlerine oranla daha tutarlı olduğu sonucuna ulaşılabilir.

BGACV kriteri; değişken sayısı ile doğru orantılı olarak sürekli artan ve sabite yakın bir grafik çizdiği ve GACV kriterine göre daha kararlı olduğu görülmektedir.

Analiz çalışmalarından ikinci olarak; bağımlı değişken sayısı 2 için sabit tutularak, gözlem sayısı 40 iken bağımsız değişken sayıları 5, 10, 20, 40, 45, 50 ve 60 olarak belirlenerek kriterler için tahmin değerleri elde edilmiştir. Elde edilen değerler Çizelge 5.6.' de verilmiştir.

Çizelge 5.6. k=2, n=40 ve p=5, 10, 20, 40, 45, 50 ve 60 iken kriterler

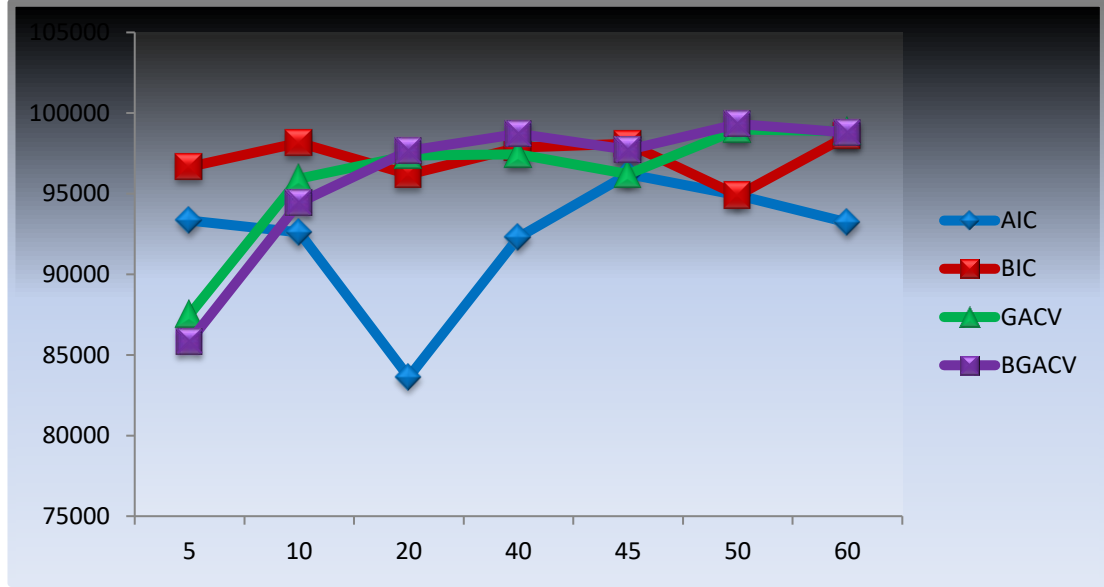
Bağımlı değ.say.(k)	Gözlem say. (n)	Değişken say. (p)	AIC	BIC	GACV	BGACV
2	40	5	93,33333	96,66667	87,50000	85,83333
2	40	10	92,59259	98,14815	95,92593	94,44444
2	40	20	83,62573	96,19883	97,36842	97,66082
2	40	40	92,30769	97,86325	97,43590	98,71795
2	40	45	96,21212	98,10606	96,21212	97,72727
2	40	50	94,89796	94,89796	98,97959	99,31973
2	40	60	93,22034	98,58757	98,87006	98,8800

Çizelge 5.6. ' de elde edilen sonuçlara bakıldığında; değişken sayısı 5 ile 10 değerinin aldığı en güçlü tahmin sonuçlarının BIC uyum iyiliği ölçütü ile elde edildiği görülmektedir. Değişken sayısı 20 ile 40 değerlerini aldığı ise BGACV uyum iyiliği ölçütü ile en güçlü LASSO tahmin sonuçlarının elde edildiği sonucuna ulaşılmaktadır. Değişken sayısı 45 değerini aldığı BIC kriteri ile ve değişken sayısı 50 ve 60 olduğunda BGACV kriteri ile en güçlü tahmin sonuçları elde edildiği görülebilmektedir.

Elde edilen sonuçlardan değişken sayısı 45 olduğunda değişiklik göstermesine rağmen genel olarak bakıldığında değişken sayısı 20 ve üzeri değerler aldığı çoğunlukla BGACV uyum iyiliği ölçütü ile en güçlü sonuçlar elde edilebildiği görülmektedir.

LASSO tahmincisi için değişken sayısının gözlem sayısından büyük olduğu durumlar özellikle incelediğinde ($p > n$); değişken sayısı 40 ve üzeri değerler aldığı bazı farklılıklar olmasına rağmen ($p=45$ için) genel olarak BGACV uyum iyiliği ölçütü ile daha güçlü ve tutarlı sonuçlar elde edilebildiği sonucuna ulaşılmaktadır.

Çizelge 5.6. ile verilen bağımlı değişken sayısı 2, gözlem sayısı 40, bağımsız değişken sayıları 5, 10, 20, 40, 45, 50 ve 60 iken elde edilen analiz sonuçları için aşağıda Şekil 5.4. ile grafiksel gösterimi verilmiştir.



Şekil 5.4. $k=2$ ve $n=40$ iken $p=5, 10, 20, 40, 45, 50$ ve 60 için kriterler

Şekil 5.4. incelendiğinde; AIC kriteri için; değişken sayısı 5 ile 10 arasında artmasına rağmen azalış grafiği çizmektedir ve yine değişken sayısı 10 ile 20 arasında artmasına rağmen azalış grafiğinin artırarak devam ettiği görülmektedir. Daha sonra değişken sayısı 20 ile 45 arasındaki artış ile birlikte AIC grafiği de artış göstermektedir ve de değişken sayısı 45 ile 60 arasında ilerleyişini sürdürmesine rağmen AIC kriteri tekrar azalış grafiği çizdiği görülebilmektedir. Yani AIC uyum iyiliği ölçütünün değişken sayısındaki sürekli artış ile farklılık gösterdiği ve değişken sayısının farklı değerler almasından etkilendiği sonucuna ulaşılabilmektedir.

BIC kriteri; değişken sayısının artması ile birlikte farklı değişken sayıları için aralıklarda farklı grafik çizdiği ve dalgalanmalar olduğu görülmektedir. Genel olarak bakıldığında sabite yakın bir eğim çizdiği görülmektedir.

GACV kriteri; bazı dalgalanmalar göstermesine rağmen değişken sayısındaki artış ile birlikte genel olarak artan bir grafik çizmektedir.

BGACV kriteri; değişken sayısındaki artış ile paralel olarak küçük dalgalanmalar göstermesine rağmen GACV ile kıyaslandığında daha tutarlı artan bir grafik çizdiği görülebilmektedir.

Analiz çalışmalarından üçüncü olarak; bağımlı değişken sayısı 2 için sabit tutularak, gözlem sayısı 30 iken bağımsız değişken sayıları 5, 10, 15, 30, 35, 40 ve 45 olarak belirlenerek kriterler için tahmin değerleri elde edilmiştir. Elde edilen değerler Çizelge 5.7.' de verilmiştir.

Çizelge 5.7. k=2, n=30 ve p=5, 10, 15, 30, 35, 40 ve 45 iken kriterler

Bağımlı değ.say.(k)	Gözlem say. (n)	Değişken say. (p)	AIC	BIC	GACV	BGACV
2	30	5	90,83333	93,33333	81,66667	77,50000
2	30	10	81,11111	95,18519	91,11111	92,59259
2	30	15	73,80952	80,95238	94,04762	95,23810
2	30	30	86,78161	96,55172	98,27586	97,70115
2	30	35	95,58824	93,62745	98,03922	98,0394
2	30	40	97,00855	97,43590	98,71795	99,57265
2	30	45	96,21212	98,10606	98,48485	98,86364

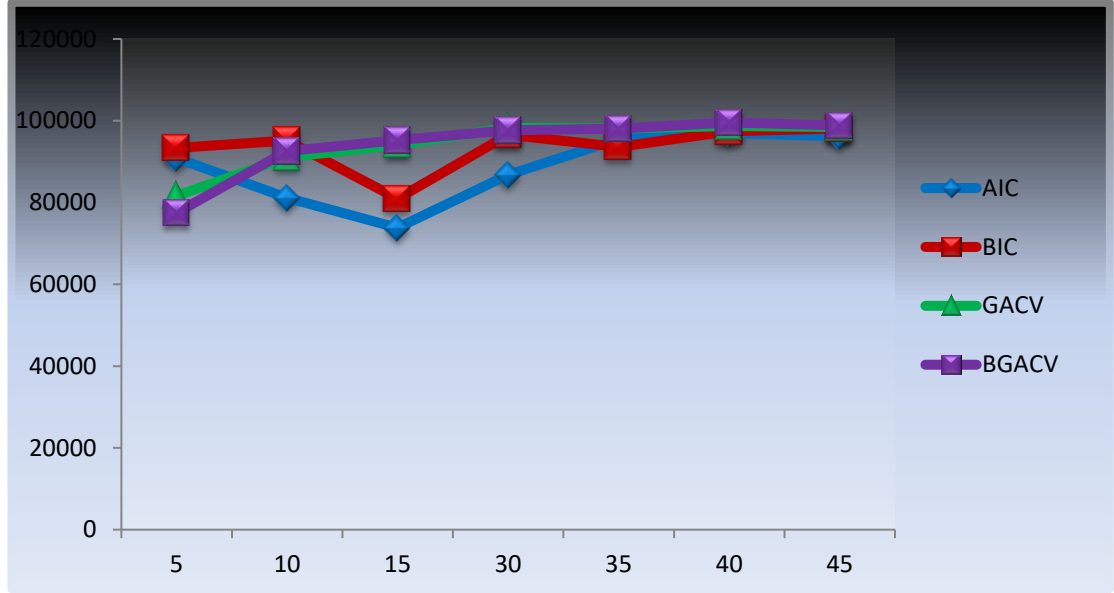
Çizelge 5.7. incelendiğinde; değişken sayısı 5 ile 10 değerlerinin aldığı BIC uyum iyiliği ölçütü ile daha güçlü LASSO tahminlerinin elde edildiği görülmektedir. Değişken sayısı 15 değerinin aldığı BGACV uyum iyiliği ölçütünün diğer kriterlere oranla daha güçlü sonuçlar verdiği sonucuna ulaşılabilmektedir.

Değişken sayısı 30 olduğunda ise GACV uyum iyiliği ölçütü en güçlü sonuçları vermektedir.

Daha sonra değişken sayısı 35, 40 ve 45 değerlerini aldığı ise en yüksek yüzdelik oranlar BGACV uyum iyiliği ölçütü ile elde edilmektedir.

Değişken sayısının gözlem sayısından fazla olduğu durumlar ($p > n$) incelendiğinde ise yani $p=35, 40$ ve 45 iken; yukarıda da değinildiği gibi değişken sayısı 35 değerinin aldığı GACV uyum iyiliği ölçütü ile ve değişken sayısı 40 ve 45 değerlerini aldığı BGACV uyum iyiliği ölçütü ile en güçlü tahmin sonuçlarının elde edildiği sonucuna ulaşılabilmektedir. Yani değişken sayısının gözlem sayısından fazla olduğu veri setlerinde genel bir istikrar elde edilmemesine rağmen genel olarak bakıldığında BGACV uyum iyiliği ölçütü ile $p > n$ olduğu durumlarda daha yüksek oranlar elde edildiğine Çizelge 5.7. ' den ulaşılabilmektedir.

Çizelge 5.7. ile verilen bağımlı değişken sayısı 2, gözlem sayısı 30, bağımsız değişken sayıları 5, 10, 15, 30, 35, 40 ve 45 iken elde edilen analiz sonuçları için aşağıda Şekil 5.5. ile grafiksel gösterimi verilmiştir.



Şekil 5.5. $k=2$ ve $n=30$ iken $p=5, 10, 15, 30, 35, 40$ ve 45 için kriterler

Şekil 5.5. ' de AIC kriteri; değişken sayısı 5 ile 15 arasındaki artış ile ters yönde bir grafik çizerek azalış göstermektedir, daha sonra değişken sayısı 15 ve üzeri değerler aldığında artış göstermiştir. Yani değişken sayısındaki sürekli artışa karşılık farklılıklar göstermiş ve dalgalanmalar bulunmaktadır fakat genel olarak bakıldığında değişken sayısının artması ile birlikte artan bir grafik çizdiği görülebilmektedir.

BIC kriterine bakıldığında; değişken sayısı 5 ile 10 arasında iken artış, 10 ile 15 arasında azalış daha sonra 15 ile 30 değerleri arasında tekrar yükseliş eğrisi çizmiştir ve çok fazla olmasa da dalgalanmalar devam etmiştir. Yani değişken sayısının farklı değerlerinden etkilendiği sonucu elde edilebilmektedir. Geneline bakıldığında artış grafiği çizdiği görülebilmektedir.

GACV kriteri; değişken sayısının artması ile birlikte atan bir grafik çizdiği ve genel olarak tutarlı bir eğime sahip olduğu anlaşılmaktadır.

BGACV kriteri ise; GACV uyum iyiliği ölçütü ile benzer bir yayılım göstermektedir fakat GACV kriterine oranla artış miktarı biraz daha fazladır ve değişken sayısı ile birlikte artan bir grafik çizmektedir.

Bağımlı değişken sayısı 2 ile sabit tutularak gözlem sayısı 50 için değişken sayıları sırasıyla 5, 15, 20, 25, 50 ve 55 olarak belirlenerek AIC, BIC, GACV ve BGACV kriterleri ile R programında LASSO tahminleri elde edilmiştir. Başlangıçta tarafımızdan belirlenen beta değerleri ile analiz sonucunda elde edilen beta tahmin değerleri arasındaki farklar her bir kritere göre hesaplanarak sonuçlar aşağıda Çizelge 5.8. ile verilmiştir. Fark değerleri ile elde edilen tahminlerin başlangıç beta değerine yakınlığını incelemek amaçlanmıştır. Yani hangi kriter veya kriterler ile gerçekçi tahmin değerleri elde edilebileceği araştırılmak istenmiştir.

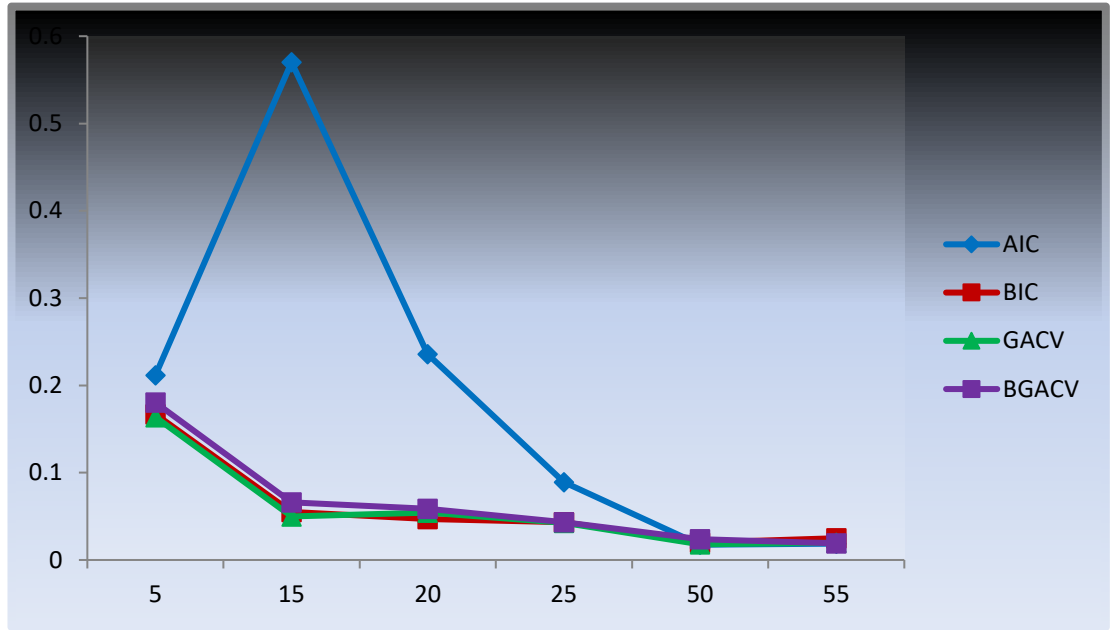
Çizelge 5.8. $k=2$, $n=50$ ve $p=5, 15, 20, 25, 50$ ve 55 iken fark değerleri

Bağımlı değ.say.(k)	Gözlem say. (n)	Değişken say. (p)	AIC	BIC	GACV	BGACV
2	50	5	0,2116894	0,1670205	0,1637165	0,1803190
2	50	15	0,57020431	0,05510832	0,04987814	0,06601267
2	50	20	0,23596264	0,04711491	0,05431673	0,05880109
2	50	25	0,08930131	0,04317916	0,04236971	0,04324139
2	50	50	0,01748834	0,01997945	0,01795729	0,02380956
2	50	55	0,01871052	0,02497719	0,01989538	0,01906744

Çizelge 5.8. incelendiğinde; değişken sayısı 5 ve 15 iken fark değerleri arasında en küçük değer GACV uyum iyiliği ölçütü ile elde edildiği görülmektedir. Değişken sayısı 20 iken BIC ve değişken sayısı 25 iken tekrar GACV kriteri ve son olarak değişken sayısı 50 ve 55 iken AIC kriteri ile en küçük fark değerleri elde edilmiştir.

Fark değerleri arasında belirli bir kararlılık görülememektedir. Fakat genel olarak bakıldığında değişken sayısının küçük değerleri için (5, 15) GACV ve artan değişken sayısı değerleri için (50 ve 55) AIC uyum iyiliği ölçütü ile en küçük fark değerleri yani başlangıç beta değerlerine en yakın tahmin değerleri elde edilmiştir.

Fark değerleri arasındaki ilişkiyi incelemek için Çizelge 5.8. ' in grafiksel gösterimi aşağıda Şekil 5.6. ' da verilmiştir.



Şekil 5.6. $k=2$ ve $n=50$ iken $p=5, 15, 20, 25, 50$ ve 55 için farklar

Şekil 5.6. ' da AIC kriteri; değişken sayısı 5 ile 15 arasındaki artış ile paralel olarak artış göstermiştir ve değişken sayısı 15 ile 50 arasında artarken AIC kriteri azalış grafiği çizmiştir ve de ani sapmalar görülmektedir. Değişken sayısındaki artışın belirli bir noktadan sonra etkisi olduğu sonucuna ulaşılabilir ($p=15$ den sonra).

BIC kriteri; değişken sayısındaki artış ile azalış göstermiştir fakat dalgalanmalar söz konusudur. AIC uyum iyiliği ölçütü ile karşılaştırıldığında daha tutarlı olduğu görülebilmektedir.

GACV kriteri; değişken sayısının 5 ile 15 arasındaki artışı doğrultusunda düzenli olara azalış grafiği sergilemiştir ve daha sonra değişken sayısının 15 ile 55 arasındaki artı doğrultusunda belirli bir miktar dalgalanma göstermesine rağmen sabite yakın bir azalış grafiği çizdiği görülebilmektedir.

BGACV kriteri ise; değişken sayısındaki artış ile paralel olarak azalış grafiği göstermiştir. GACV uyum iyiliği ölçütü ile karşılaştırıldığında başlangıç değişken sayılarında daha yüksek değerler aldığı, değişken sayısı arttıkça genel olarak aldığı değerler GACV kriterinin aldığı değerlere yaklaştığı görülebilmektedir.

Bağımlı değişken sayısı 2 alınarak, gözlem sayısı 40 iken bağımsız değişken sayıları 5, 10, 20, 40, 45, 50 ve 60 şeklinde belirlenerek kriterlere göre fark değerleri elde edilmiştir. Gözlem sayısı 40 için fark değerleri sonuçları Çizelge 5.9. ' de verilmiştir.

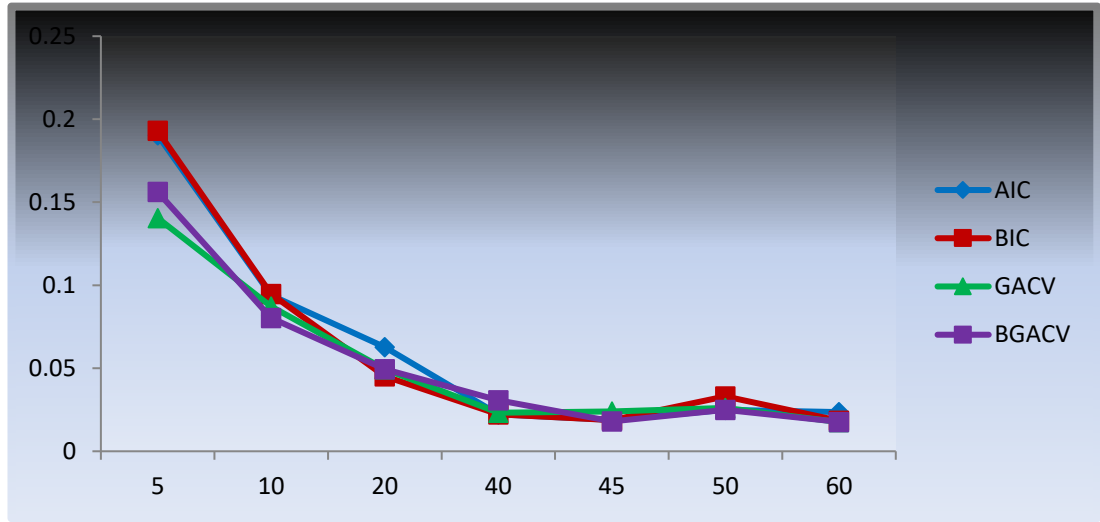
Çizelge 5.9. $k=2$, $n=40$ ve $p=5, 10, 20, 40, 45, 50$ ve 60 iken fark değerleri

Bağımlı değ.say.(k)	Gözlem say. (n)	Değişken say. (p)	AIC	BIC	GACV	BGACV
2	40	5	0,1904546	0,1929956	0,1404920	0,15622119
2	40	10	0,09402993	0,09484096	0,08718514	0,08037726
2	40	20	0,06274219	0,04516084	0,04948127	0,04941111
2	40	40	0,02224184	0,02219564	0,02315778	0,030736
2	40	45	0,01890874	0,01871058	0,02391775	0,0179889
2	40	50	0,02490205	0,03310353	0,02613119	0,02496436
2	40	60	0,02358912	0,01841892	0,01763292	0,01772599

Çizelge 5.9. ' da; değişken sayısı 5 değerini aldığı anda fark değerleri arasında en küçük değeri veren yani başlangıç beta değerlerine en yakın tahmin değerlerinin elde edilmesini sağlayan uyum iyiliği ölçütünün GACV uyum iyiliği ölçütü olduğu görülmektedir. Değişken sayısı 10 iken gerçek değerlere en yakın tahmin sonuçlarının BGACV uyum iyiliği ölçütü ile elde edildiği sonucuna ulaşılmaktadır. Devam edildiğinde, değişken sayısı 20 ve 40 değerlerini aldığı anda ise BIC uyum kriteri ile en küçük fark değerinin elde edildiği, değişken sayısı 45 iken BGACV uyum iyiliği ölçütü ile, değişken sayısı 50 iken AIC kriteri ile ve değişken sayısı 60 değerini aldığı anda GACV uyum iyiliği ölçütünün en küçük fark değerlerine ulaşılmasına katkı sağladığı görülebilmektedir.

Genel olarak bakıldığında, Çizelge 5.9. ' da verilen yapılan simülasyonlar sonucunda elde edilen sonuç değerleri incelendiğinde değişken sayısındaki artış ile orantılı olarak kararlı bir uyum iyiliği ölçütü elde edilemediği, sonuçların yani tarafımızdan belirlenen beta değerleri ile tahmin değerleri arasındaki fark değerlerinin değişken sayısının farklı değerler alması ile birlikte farklılık gösterdiği görülebilmektedir.

Çizelge 5.9. sonucu elde edilen fark değerleri için grafiksel gösterim aşağıda Şekil 5.7. ' de verilmiştir.



Şekil 5.7. $k=2$ ve $n=40$ iken $p=5, 10, 20, 40, 45, 50$ ve 60 için farklar

Şekil 5.7. ' de; AIC kriteri; değişken sayısı 5 ile 40 arasında artış gösterirken AIC grafiği beklenildiği şekilde azalış göstermektedir, değişken sayısı 40 ve üzeri değerler aldığı anda ise sabite yakın bir grafik çizmektedir. Fark değerlerinin küçük olması beklenildiği için kriter eğrilerinin azalış göstermesi beklenilmektedir fakat AIC eğrisi genel olarak azalan bir eğim çizmesi ile birlikte dalgalanmalar da bulunmaktadır.

BIC kriteri; AIC uyum iyiliği ölçütü ile benzer bir grafik çizdiği görülmektedir. Değişken sayısı 10 ile 40 değerleri arasında AIC kriteri ile eğrisinde farklılıklar gözlenmektedir. Değişken sayısı 40 ve üzeri değerler aldığı durumlarda dalgalanmalar olması ile birlikte genel olarak sabite yakın bir grafik çizmektedir.

GACV uyum iyiliği ölçütü; AIC ve BIC uyum iyiliği ölçütleri ile karşılaştırıldığında daha kararlı olduğu görülebilmektedir. Değişken sayısının artması ile birlikte GACV uyum iyiliği ölçütünün fark eğrisi genel olarak azalış grafiği çizdiği görülebilmektedir.

BGACV kriteri incelendiğinde, GACV uyum iyiliği ölçütü ile benzer bir grafik çizdiği sonucuna ulaşılabilir. Ayrıca genel olarak azalan bir grafik çizmesine rağmen GACV kriteri ile benzer olarak farklı değişken sayıları için dalgalanmalar yani kararsızlık göstermektedir.

Bağımlı değişken sayısı 2 alınarak, gözlem sayısı 30 iken bağımsız değişken sayıları 5, 10, 15, 30, 35, 40 ve 45 şeklinde belirlenerek kriterlere göre fark değerleri elde edilmiştir. Gözlem sayısı 30 için fark değerleri sonuçları aşağıda Çizelge 5.10. ' de verilmiştir.

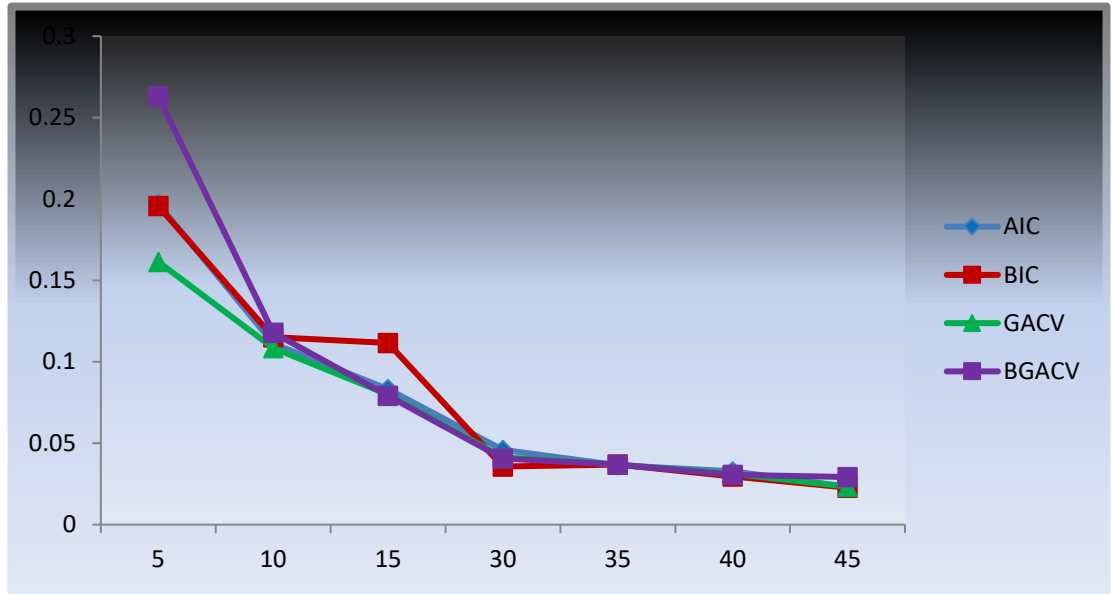
Çizelge 5.10. $k=2$, $n=30$ ve $p=5, 10, 15, 30, 35, 40$ ve 45 iken fark değerleri

Bağımlı değ.say.(k)	Gözlem say. (n)	Değişken say. (p)	AIC	BIC	GACV	BGACV
2	30	5	0,1970067	0,1956799	0,1614125	0,2629182
2	30	10	0,1111654	0,1151862	0,1084233	0,1178033
2	30	15	0,08326589	0,11156934	0,07999956	0,07913841
2	30	30	0,04588137	0,03579153	0,04141068	0,04059910
2	30	35	0,03658719	0,03679618	0,0368093	0,03682093
2	30	40	0,03277299	0,02968867	0,03123577	0,03057201
2	30	45	0,02328015	0,02272489	0,02336273	0,02925417

Çizelge 5.10. ' da; değişken sayısı 5 ve 10 için en küçük fark değerini veren uyum iyiliği ölçütü GACV kriteridir. Değişken sayısı 15 olduğunda BGACV uyum iyiliği ölçütü ile en küçük fark değerinin elde edildiği görülmektedir. Değişken sayısı 30 ve üzeri değerleri aldığında BIC uyum iyiliği ölçütü ile en küçük fark değeri elde edilmiştir, sadece değişken sayısı 35 iken AIC uyum iyiliği ölçütü ile en küçük fark değeri elde edilmiştir ve değişken sayısının 30 ve üzeri değerleri arasında farklılık göstermiştir.

Gözlem sayısı 30 için yapılan çalışma sonucunda elde edilen en küçük fark değerlerinin elde edilmesinde uyum iyiliği ölçütleri arasında farklılık ve belirli bir kararlılık göze çarpmamaktadır. Fakat Çizelge 5.10. genel olarak incelendiğinde; değişken sayısının küçük değerleri için başlangıç beta değerlerine en yakın tahmin değerlerini GACV uyum iyiliği ölçütü ile, diğer taraftan değişken sayısının artan değerleri için bazı farklılıklar ve sapmalar olmasına rağmen BIC uyum iyiliği ölçütü ile en küçük fark değerlerinin elde edildiği sonucuna ulaşılabilmektedir.

Çizelge 5.10. sonucu elde edilen fark değerleri için grafiksel gösterim aşağıda Şekil 5.8. ' de verilmiştir.



Şekil 5.8. $k=2$ ve $n=30$ iken $p=5, 10, 15, 30, 35, 40$ ve 45 için farklar

Şekil 5.8. incelendiğinde AIC kriteri; değişken sayısı artarken genel olarak azalan bir grafik çizmektedir. Diğer gözlem sayıları ile kıyaslandığında ($n=50, 40$ için) $n=30$ için fark değerleri AIC kriteri ile daha az dalgalanma göstermesine rağmen hafif dalgalanmalar ile birlikte azalan bir grafik çizdiği görülmektedir.

BIC kriteri; genel olarak azalan bir grafik çizdiği görülmektedir fakat değişken sayısındaki sürekli artış ile farklı değişken sayıları için değişkenlik göstermektedir. AIC kriteri ile kıyaslandığında daha fazla dalgalanma mevcuttur.

GACV uyum iyiliği ölçütü incelendiğinde; BIC ve AIC uyum iyiliği ölçütü ile kıyaslandığında daha karalı olarak azalan bir grafik çizmektedir. Azalma miktarları arasında ani hareketler görülmemektedir. Değişken sayısı 5 iken GACV uyum iyiliği ölçütünün başlangıç noktası diğer kriterlerden daha düşüktür. Yani diğer kriterlerdeki azalma oranlarının daha fazla olduğu Şekil 5.8. ' den görülebilmektedir.

BGACV uyum iyiliği ölçütünün grafiğine bakıldığında; değişken sayısı 5 iken grafiğin başlangıç noktası diğer kriterlere göre daha büyüktür ve dolayısıyla azalama miktarı daha fazladır. Değişken sayısı 5 ile 10 arasındaki azalma miktarı diğer değişken sayısı değerlerine göre fazla olmasına rağmen genel olarak bakıldığında değişken sayısının artışı ile azalan bir grafik çizdiği görülebilmektedir.

6. SONUÇ VE ÖNERİLER

İstatistiksel yöntemler arasında yaygın olarak kullanılan regresyon analizi ile toplanan veri setleri hakkında bilgi sahibi olmak ve ileriye yönelik fikir edinilmesine katkı sağlayacak şekilde tahmin modelleri oluşturulması amaçlanmaktadır. Modeller oluşturulurken gerçek gözlemler hakkında en doğru bilgiyi verebilecek en sade tahmin modeli elde edilmelidir. En küçük kareler tahmin yöntemi (EKK) ve en çok olabilirlik tahmin yöntemi (EÇOT) bu tahmin yöntemlerine birer örnektir. Matematiksel hesaplama kolaylığı sağlamasından dolayı yaygın olarak tercih edilen EKK yöntemi ile hata kareler toplamını minimize etmek amaçlanmaktadır. Fakat EKK tahmincisi ile bazı durumlarda kararlı tahmin sonuçları elde etmek zorlaşmaktadır. Özellikle büyük veri setlerinde iki veya daha fazla değişken arasında ilişki olması durumunda EKK tahmini kararsızlık gösterebilir ve elde edilen tahmin sonuçlarının gerçek sonuçları yansıtmadığı durumlar elde edilebilmektedir. Bu gibi problemlerin çözülmesi için bazı yanlı tahmin yöntemleri alternatif olarak kullanılabilir. Yanlı tahmin yöntemleri ile daha yanlı fakat daha etkili modeller elde edilebilmektedir. Bu yanlı yöntemlerden birisi LASSO tahmin yöntemidir ve LASSO ile EKK tahmin denkleminde belirli bir ceza kısıdı getirilerek daha tutarlı tahmin değerleri elde etmek amaçlanmaktadır.

LASSO ile değişken seçimi ve parametre tahmini aynı anda yapılabilir ve özellikle büyük veri setlerinde daha tutarlı sonuçlar elde etmek için kullanılmaktadır. Ayrıca değişken sayısının gözlem sayısından fazla olduğu büyük veri setlerinde birçok parametreyi sıfıra daraltarak yorumlanması daha kolay modeller elde edilmesine olanak sağlamaktadır.

Bu çalışmada; LASSO tahmincisi çok değişkenli Bernoulli lojistik modeller üzerinde incelenmiş ve model seçiminde kullanılan uyum iyiliği ölçütlerinden AIC, BIC, GACV ve BGACV uyum iyiliği ölçütleri ile tahmin modelleri elde edilerek daha kararlı ve güçlü kriterleri belirlemek için karşılaştırmalar yapılmıştır.

Öncelikle, çok değişkenli Bernoulli lojistik modellerde LASSO tahmincilerinin incelenmesi amacı ile R programında MVB paketi kullanılarak hazır veri seti ile belirlenen başlangıç beta değerleri için gözlem sayısı 50, 100, 200, 250 ve 300 olduğu durumlarda, bağımlı değişken sayısı 2,3 ve bağımsız değişken sayısı 5 alınarak simülasyonlar yapılmıştır. Her bir durum için 100 tekrar yapılarak ve ayar

parametre seçiminde AIC, BIC, GACV ve BGACV uyum iyiliği ölçütleri kullanılarak tahmin sonuçları elde edilmiştir. Simülasyon sonuçlarından, gözlem sayısı 50,100 ve 200 için bağımlı değişken sayısı 2 ve 3 olduğu durumlarda BIC kriteri ile diğer taraftan gözlem sayısı 250 ve 300 için bağımlı değişken 2 ve 3 olduğu durumlarda ise BGACV kriteri ile en güçlü LASSO tahmin modelleri elde edildiği sonucuna ulaşılmıştır.

Grafiksel gösterimlerden Şekil 5.1. ve Şekil 5.2. incelendiğinde ise sadece bağımlı değişken sayısı değiştirildiğinde en güçlü tahmin modelinin elde edilmesinde uyum iyiliği ölçütlerinde farklılıklar gözlemlendiği sonucuna varılmıştır. Bağımlı değişken sayısı artırıldığında, AIC ve BIC kriterleri farklı sonuçlar vermekte ve sonuçlarında dalgalanmalar görülmektedir. GACV ve BGACV kriterlerinde ise, bağımlı değişken sayısındaki artış ile orantılı olarak artma gösterdikleri ve dalgalanmalarının daha az olduğu yani daha kararlı oldukları sonuçlarına ulaşılmıştır.

Özellikle bağımsız değişken sayısının gözlem sayısından fazla olduğu durumlarda LASSO tahmincilerinin kullanımı daha yansız modelleri ortaya çıkaracaktır. Bu sebeple küçük veri setleri için farklı gözlem sayıları ile $p > n$ olduğu durumlar oluşturularak ve bağımlı değişken sayıları sabit tutularak LASSO tahminleri uyum iyiliği ölçütlerine göre karşılaştırılmıştır. Sonuçlar; gözlem sayısı 50 iken değişken sayısı 25 ve üzeri değerleri için BGACV kriteri ile daha güçlü LASSO tahminleri elde edildiği, gözlem sayısı 40 iken değişken sayısı 20 ve üzeri değerleri için ve de gözlem sayısı 30 iken değişken sayısı 15 ve üzeri değerleri için genel olarak BGACV kriteri ile en güçlü LASSO tahmin sonuçları elde edildiği şeklindedir.

Ayrıca gözlem sayısı 50, 40 ve 30 için $p > n$ olacak şekilde oluşturulan veri setleri için sonuçta elde edilen tahmin değerleri ile başlangıçta tarafımızdan belirlenen beta değerleri arasındaki fark değerleri tablo ve grafik olarak verilerek incelenmiştir. Fark ne kadar az olursa tahmin değerleri gerçek değerlerine o kadar yakın olmaktadır. Bu sebeple fark değerleri arasında küçük değerler uyum iyiliği ölçütlerine göre karşılaştırılmıştır. Analiz sonuçlarında; değişken sayısındaki artışa bağlı olarak genel olarak bir kararlılık elde edilememesine rağmen genel olarak bakıldığında AIC ve BIC kriterlerinin tahmin değerlerinde farklı değerler ile dalgalanmaların daha fazla olduğu, GACV ve BGACV kriterlerinin diğer iki kritere göre daha kararlı olduğu sonucuna ulaşılabilmektedir. Süre kısıdı olması ve

sürelerinin uzun olmasından dolayı tez çalışması için simülasyon sayıları kısıtlı tutulmuştur. Bu sebeple, gözlem sayılarına göre değişken sayıları artırılarak analizler yapıldığında fark değerleri arasında daha net bir kararlılık elde edilebileceği öngörülmektedir.

Sonuç olarak; LASSO tahmincisi özellikle büyük veri setlerinde ve $p > n$ durumlarında daha sade modeller elde edilmesinde ve daha güvenilir sonuçlar vermesi sebebi ile yaygın olarak kullanılmaktadır. Bu çalışmada ise LASSO tahmincisi, daha küçük veri setleri üzerinde ve yine $p > n$ durumlarında incelenmiştir. Çalışma sonucunda LASSO tahmin modellerinin elde edilmesinde AIC ve BIC kriterlerine bir alternatif olarak GACV ve BGACV uyum iyiliği ölçütleri ile daha kararlı ve güçlü sonuçlar elde edildiği sonucuna ulaşılmıştır.

KAYNAKLAR

- Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. Proc. 2nd Inter. Symposium on Information Theory, 267- 281, Budapest.
- Akaike, H. 1974. A new look at the statistical model identification. IEEE transactions on automatic control, 19 (6), 716- 723.
- Akdi, Y. 2005. Matematiksel istatistiğe giriş. Bıçaklar Kitabevi.
- Alma, Ö. G. 2008. Regresyon analizinde kullanılan en küçük kareler ve en küçük medyan kareler yöntemlerinin karşılaştırılması. *Süleyman Demirel Üniversitesi Fen Dergisi*, 3 (2).
- Behmanesh, R. and Saghaei, A. 2017. Robust least angle regression and lasso using bootstrap aggregation (real case study: Iranian poultry farms). International journal of industrial and systems engineering, 26 (2), 201- 227.
- Breiman, L. 1995. Better subset regression using the nonnegative garrote. *Technometrics*, 37 (4), 373-384.
- Bülbül, S. 2006. Üç boyutlu çapraz tablolarda logaritmik doğrusal analiz: Çocuk işgücü değişkenleri arasındaki etkileşimler.
- Candes, E. J. and Tao, T. 2010. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56 (5), 2053-2080.
- Candes, E. and Tao, T. 2007. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 2313-2351.
- Chen, S. S. and Donoho, D. L. 1998. Application of basis pursuit in spectrum estimation. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on* (Vol. 3, pp. 1865-1868). IEEE.
- Chen, S. S., Donoho, D. L. and Saunders, M. A. 2001. Atomic decomposition by basis pursuit. *SIAM review*, 43 (1), 129-159.
- Çokluk, Ö. 2010. Lojistik regresyon analizi: Kavram ve uygulama. *Kuram ve uygulamada eğitim bilimleri*, 10 (3), 1357-1407.
- Dai, B. (2012). Multivariate Bernoulli distribution models. Technical Report, Department of Statistics, University of Wisconsin, Madison, WI 53706.
- Dai, B. 2013. MVB: Multivariate Bernoulli log-linear model. R package version, 1.
- Dai, B., Ding, S. and Wahba, G. 2013. Multivariate Bernoulli distribution. *Bernoulli*, 19 (4), 1465-1483.
- Dalalyan, A. S., Hebiri, M. and Lederer, J. 2017. On the prediction performance of the lasso. *Bernoulli*, 23 (1), 552-581.
- Donoho, D. L. and Johnstone, J. M. 1994. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81 (3), 425-455.
- Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. 2004. Least angle regression. *The Annals of statistics*, 32 (2), 407-499.

- Finch, W. H. and Hernandez, M. E. 2016. Regularization methods for fitting linear models with small sample sizes: Fitting the lasso estimator using R. *Practical Assessment, Research and Evaluation*, 21(7).
- Friedman, J., Hastie, T. and Tibshirani, R. 2001. *The elements of statistical learning* (1). Springer, Berlin: Springer series in statistics.
- Friedman, J., Hastie, T. and Tibshirani, R. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33 (1), 1.
- Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. 2007. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1 (2), 302-332.
- Golub, G. H., Heath, M. and Wahba, G. 1979. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21 (2), 215-223.
- Gu, C. and Xiang, D. 2001. Cross-validating non-gaussian data: Generalized approximate cross-validation revisited. *Journal of Computational and Graphical Statistics*, 10 (3), 581-591.
- Hans, C. 2009. Bayesian lasso regression. *Biometrika*, 835-845.
- Hastie, T., Tibshirani, R. and Wainwright, M. 2015. *Statistical learning with sparsity: the lasso and generalizations*. CRC Press.
- Hoerl, A. E. and Kennard, R. W. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12 (1), 55- 67.
- Lin, X., Wahba, G., Xiang, D., Gao, F., Klein, R. and Klein, B. 2000. Smoothing spline ANOVA models for large data sets with Bernoulli observations and the randomized GACV. *Annals of Statistics*, 1570-1600.
- Meier, L., Van De Geer, S. and Bühlmann, P. 2008. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70 (1), 53-71.
- Meinshausen, N. and Bühlmann, P. 2006. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, 1436-1462.
- Park, M. Y. and Hastie, T. 2007. L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69 (4), 659-677.
- Scholz, F. W. 1985. Maximum likelihood estimation. *Encyclopedia of statistical sciences*.
- Schwarz, G. 1978. Estimating the dimension of a model. *The Annals of Statistics*, 6 (2), 461- 464.
- Shi, W., Wahba, G., Wright, S., Lee, K., Klein, R. and Klein, B. 2006. Lasso-patternsearch algorithm with application to ophthalmology data. *arXiv preprint math/0610916*.
- Theodoridis, S. 2015. *Machine learning: a Bayesian and optimization perspective*. Academic Press.
- Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

- Tibshirani, R. 1997. The lasso method for variable selection in the Cox model. *Statistics in medicine*, 16 (4), 385-395.
- Tibshirani, R. 2011. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73 (3), 273-282.
- Tibshirani, R. J. 2011. The solution path of the generalized lasso. Stanford University.
- Tibshirani, R. J. 2013. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7, 1456-1490.
- Tibshirani, R. J. and Taylor, J. 2012. Degrees of freedom in lasso problems. *The Annals of Statistics*, 1198-1232.
- Tibshirani, R. J., Hoefling, H. and Tibshirani, R. 2011. Nearly-isotonic regression. *Technometrics*, 53 (1), 54-61.
- Van de Geer, S. A. 2008. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 614-645.
- Wahba, D. X. G. 1994. A generalized approximate cross validation for smoothing splines.
- Wahba, G. and Craven, P. 1978. Smoothing noisy data with spline functions. estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31, 377-404.
- Wahba, G. and Wold, S. 1975. A completely automatic French curve: Fitting spline functions by cross validation. *Communications in Statistics-Theory and Methods*, 4 (1), 1-17.
- Wahba, G., Lin, Y. and Zhang, H. 1999. Generalized approximate cross validation for support vector machines, or, another way to look at margin-like quantities.
- Wasserman, L. 2000. Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44, 92-107.
- Witten, D., Tibshirani, R. and Hastie, T. 2009. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis.
- Wu, T. T. and Lange, K. 2008. Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, 224-244.
- Xiang, D. and Wahba, G. 1996. A generalized approximate cross validation for smoothing splines with non-Gaussian data. *Statistica Sinica*, 675-692.
- Xiang, D. and Wahba, G. 1997. Approximate smoothing spline methods for large data sets in the binary case. In *Proceeding-American Statistical Association Biometrics Section* (pp. 94-99).
- Yau, C. Y. and Hui, T. S. 2017. Lars- type algorithm for group lasso. *Statistics and Computing*, 27 (4), 1041-1048.
- Yuan, M. and Lin, Y. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68 (1), 49-67.

- Zhang, H. H. and Lu, W. 2007. Adaptive lasso for Cox's proportional hazards model. *Biometrika*, 94(3), 691-703.
- Zhang, H. H., Wahba, G., Lin, Y., Voelker, M., Ferris, M., Klein, R. and Klein, B. 2004. Variable selection and model building via likelihood basis pursuit. *Journal of the American Statistical Association*, 99 (467), 659-672.
- Zhao, P. and Yu, B. 2006. On model selection consistency of lasso. *Journal of Machine learning research*, 7 (Nov), 2541-2563.
- Zou, H. 2006. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101 (476), 1418-1429.
- Zou, H. and Hastie, T. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67 (2), 301-320.

ÖZGEÇMİŞ

Adı ve Soyadı : Aslı YAMAN
Doğum Yeri : Merzifon/ AMASYA
Doğum Tarihi : 29.05.1989
Yabancı Dili : İngilizce

Eğitim Durumu

Lise : Merzifon Anadolu Lisesi (2007)
Lisans : Çukurova Üniversitesi- Fen Edebiyat Fakültesi- İstatistik Bölümü (2012)
Yüksek Lisans : Ondokuz Mayıs Üniversitesi- Fen Bilimleri Enstitüsü- İstatistik Anabilim Dalı (02. 2014- Halen)

Çalıştığı Kurum/ Kuruluşlar ve Yıl

Eskişehir Erişim Müşteri Hizmetleri A.Ş.- 01.2016- 10.2016
Merzifon GM Teknik Cam A.Ş. - 02.2014- 01.2015