Süleyman Demirel Üniversitesi
Sağlık Bilimleri Dergisi
Cilt 13, Sayı 3, 485 - 496, 2022

Suleyman Demirel University
Journal of Health Sciences
Volume 13, Issue 3, 485- 496, 2022

Genomik Veri Setlerinin LASSO ve Elastik Net Regresyon Yöntemleri ile Analizi

Analysis of Genomic Data Sets by LASSO and Elastic Net Regression Methods

Hikmet ORHAN [1], Merve VERGİLİ [1*]

[1] Süleyman Demirel Üniversitesi, Tıp Fakültesi, Biyoistatistik ve Tıbbi Bilişim AD, Isparta, Türkiye

ÖZET

Amaç: Bu çalışmanın amacı büyük boyutlu genomik veri setlerinin değişken seçim yöntemleri kullanılarak daha küçük boyutlara indirgenip daha az maliyet ve zaman ile analizlerin gerçekleştirilebileceğini göstermektir.

Gereç ve Yöntem: Bu çalışmada NCBI veri tabanından Bioconductor yardımı ile R programına aktarılan GDS4906 numaralı veri seti kullanılmıştır. Veri seti 10-katlı çapraz doğrulama ile LASSO ve Elastik Net regresyon yöntemleri kullanılarak analiz edilmiştir.

Bulgular: Veri seti LASSO regresyon yöntemi ile analiz edildiğinde veri setinden 5 adet gen seçilmiş olup, sonrasında farklı iterasyonlarda seçilen değişkenler ve değişken sayılarında farklılık gözlendiğinden kararlılık seçimi yöntemi uygulanarak 2 adet gen seçilmiş ve modelin R^2 değeri 0,85 olarak bulunmuştur. Aralıklı arama yöntemi kullanılarak uygulanan Elastik Net regresyon yönteminde 19 adet gen seçilmiş ve R^2 değeri 0,92 olarak bulunmuştur.

Sonuç: Elde edilen sonuçlara göre LASSO ve Elastik Net regresyon yöntemlerinin genomik veri setlerinde iyi bir performans gösterdiği anlaşılmıştır.

Anahtar Kelimeler: Çoklu bağlantı, Elastik Net, Genomik Veri, LASSO

Alınış / Received: 10.11.2022 Kabul / Accepted: 29.11.2022 Online Yayınlanma / Published Online: 20.12.2022

DOI: 10.22312/sdusbed
ARAŞTIRMA MAKALESİ

*Corresponding author: mvergili@hotmail.com

ABSTRACT

Objective: The purpose of this study is to show that large-sized genomic datasets can be reduced to smaller sizes using variable selection methods, and that analysis can be performed with less cost and time.

Materials and methods: This study uses dataset number GDS4906, which is transferred from the NCBI database to the R program using Bioconductor. The dataset was analyzed using LASSO and Elastic Net regression methods with 10-fold cross-validation.

Results: When the dataset is analyzed using the LASSO regression method, 5 genes were selected from the dataset and 2 genes were selected and the R^2 values of the model were found as 0.85 by applying the determination selection method, as the variables and variable numbers selected in different iterations were then different. In the Elastic Net regression method applied using the interval search method, 19 genes were selected and R^2 were found as 0.92.

Conclusion: According to the results obtained, LASSO and Elastic Net regression methods have shown a good performance in the genomic datasets.

Keywords: Elastic Net, Genomic Data, Multicollinearity, LASSO

1. Giriş

Regresyon analizi, değişkenler arasındaki ilişkiyi modellemek ve keşfetmek amacıyla kullanılan istatistiksel bir tekniktir. Regresyonun mühendislik, fizik ve kimya bilimleri, iktisat, yaşam ve biyoloji bilimleri ve sosyal bilimler gibi birçok kullanım alanı olması sebebiyle en yaygın kullanılan istatistiksel teknik sayılabilmektedir. Regresyon yöntemlerinden çoklu doğrusal regresyonun sağlaması gereken varsayımlar vardır. Bunlardan biri regresyon modelindeki değişkenlerin birbirleri arasında ilişkinin olmamasıdır. Bu varsayımın sağlanmadığı durumlarda çoklu doğrusal bağlantı sorunu meydana gelir ve gerçektekinden önemli ölçüde farklı kestirim ile sonuçlanabilmektedir [1].

Çoklu doğrusal bağlantı sorunu sağlık, kimya ve biyoloji verilerinde yaygındır. Çoklu doğrusal bağlantı sorununa çözüm olarak kararlı tahminler yapabilmek amacıyla yanlı tahmin ediciler kullanılmaktadır. Genomik veri setleri içinde benzer işleve sahip genler arasındaki yüksek korelasyon nedeniyle çoklu doğrusal bağlantı sorunu olmasından dolayı bu çalışmada yanlı tahmin edicilerden Tibshirani tarafından (1996) önerilmiş olan LASSO, Zou ve Hastie (2005) tarafından önerilmiş olan Elastik Net regresyon yöntemleri kullanılmıştır.

Pripp ve Stanis (2017) tarafından yapılan bir çalışmada, lezyon bölgesindeki 30 inflamasyon ve anjiyogenez biyobelirteçleri ile 93 hastanın seçilmiş klinik ve radyolojik özellikleri arasındaki ilişkiyi değerlendirmek için LASSO regresyonu kullanılmıştır [2].

Kohannim ve arkadaşları tarafından (2012) Alzheimer Hastalığı Nörogörüntüleme Girişimi'nin (ADNI) bir parçası olarak taranan 729 denekten MRI'dan türetilen bir temporal lob hacmi ölçümleri kullanılarak, beyin görüntülemenin genom çapında ilişkilendirme çalışmalarında (GWAS) gen etkilerini LASSO regresyonu kullanılarak değerlendirmişlerdir [3].

Çiftsüren ve Akkol (2018), düzenleme yöntemlerini kullanarak iç yumurta kalitesi özelliklerinin tahmini ve değişken seçimini Ridge, Lasso ve Elastik Net regresyon yöntemlerini kullanarak yapmıştır. Çalışmada 117 Japon bıldırcını kullanılarak yumurtaların iç kalite özellikleri yumurta sarısı ağırlığı ve yumurta akı ağırlığı; dış kalite özellikleri yumurta genişliği, yumurta uzunluğu, yumurta ağırlığı, şekil indeksi ve kabuk ağırlığı ölçümleri yapılmıştır. Veri setindeki çoklu doğrusallık olması sebebiyle Ridge, LASSO ve Elastik Net yöntemleri uygulanmıştır. Hem yumurta sarısı ağırlığı hem de yumurta akı ağırlığı için iki tahmin edici içeren LASSO regresyon yönteminin modelin tahmin doğruluğu açısından en iyi sonuçları verdiği bulunmuştur [4].

Cho ve ark (2009), romatoid artritin GWAS'da bütün bir genom boyunca hastalığa neden olan genleri tespit etmek için Elastik Net çoklu lojistik regresyon modelini kullanan basit bir aşamalı yaklaşım önermişlerdir. Elastik Net regresyon yöntemi, GWAS'da hastalığa neden olan SNP'leri birlikte tanımlamada bazı avantajlara sahip olduğunu belirtmişlerdir. Bu avantajlardan ilki, otomatik değişken seçimi ve sürekli daraltma aynı anda gerçekleştirilebilmekte; ikincisi, klasik çoklu doğrusal regresyonlarda çoklu bağlantı problemine neden olabilen yüksek korelasyona sahip SNP'lerden oluşan grupları seçebilmekte; üçüncüsü, Elastik Net regresyonun daraltma özelliği sayesinde, SNP'ler ve genotipik olmayan faktörler arasındaki tüm etkileşim terimlerinin yanı sıra SNP ana etkilerinin de modele dahil edilmesini sağlamakta olduğunu belirtmişlerdir. Ek olarak, doğrudan kromozomlar arasında potansiyel SNP'leri aramak yerine, bu yaklaşımın GWAS'ta çok sayıda potansiyel SNP modelini işlemek için çok adımlı bir prosedür kullanarak verimli arama sağladığını bulmuş ve rapor etmişlerdir [5].

Bu çalışmada, NCBI veri tabanından alınan GDS4906 numaralı KOAH isimli büyük boyutlu genomik veri setinin R yazılımı kullanılarak LASSO ve Elastik Net regresyon yöntemleri ile daha basit ve başarılı modeller oluşturulabileceğinin gösterilmesi amaçlanmıştır.

2. Materyal ve Metot

Çalışmada kullanılan veri seti NCBI (National Center for Biotechnology Information) Gene Expression Omnibus (GEO) veri tabanından alınmıştır (6). “Egzersiz eğitiminin kronik obstrüktif akciğer hastalığı hastalarına etkisi: vastus lateralis kası” başlıklı veri seti GSE27536 referans serisi altında bulunan GDS4906 numaralı, GPL570: Affymetrix Human Genome U133 Plus 2.0 Array (HG-U133_Plus_2) mikroarrayleri ile ölçülmüş 54 hastanın gen ekspresyon verilerini içermektedir. Veri seti kronik obstrüktif akciğer hastalığı (KOAH) olan hastalara 8 haftalık egzersiz öncesi ve sonrası alınan kas analizinden oluşmaktadır. İskelet kasının işlev bozukluğu, kasların zayıflaması, küçülmesi ve kaybı sonucu güç ve hareket kabiliyeti düşüklüğü KOAH’ın ayırt edilebilen sistemik etkilerindendir.

LASSO Regresyon Yöntemi

Yanlı tahmin yöntemlerinden LASSO (Least Absolute Shrinkage and Selection Operator) regresyon yöntemi Tibshirani tarafından 1996 yılından önerilmiştir. Ridge regresyon yöntemine benzeyen LASSO regresyon yöntemi katsayılar üstüne ceza terimi uygulanması ile bazı katsayıları sıfıra indirgeyerek çalışmaktadır. LASSO regresyon yönteminin tahmin edicisi eşitlik (1)’de verilmiştir.

$$\hat{\beta}_{lasso} = \underset{\beta}{arg\,min}\left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right\} \quad (1)$$

Bu eşitlikte, n gözlem sayısı, y bağımlı değişken, p değişken sayısı, β_0 ve $\beta = (\beta_1, \beta_2, \dots, \beta_p)$ bilinmeyen parametreler, $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$ bağımsız değişkenleri ve λ ceza terimini (ayar parametresi) ifade etmektedir. Ceza terimi (λ), daralma (shrinkage) miktarını kontrol eden parametredir ve λ’nın aldığı değer ne kadar artarsa daralma miktarı aynı oranda artmaktadır [7]. Ceza terimi sıfırdan büyük bir değer ($\lambda > 0$) olmalıdır.

$l_1 = \sum_{j=1}^{p}|\beta_j|$ ifadesi ise ceza fonksiyonu olarak adlandırılmaktadır. Ceza fonksiyonun alacağı değer regresyon modeline girecek olan değişken sayısını etkilemektedir ve aldığı değer büyüdükçe modele giren değişken sayısı artmaktadır.

LASSO regresyon yöntemi katsayıları sıfıra indirgeyebilmesi sayesinde modelde daha az değişken bulunmasına imkân sağlayarak yorumlanması kolay ve net regresyon modelleri elde edilmesini

sağlamaktadır. Bu sayede çok fazla sayıda gözlem ve değişken barındıran büyük veri setlerinde (big data) ya da değişken sayısı gözlem sayısından büyük olan (p>n) verilerde fayda sağlamaktadır.

Elastik Net Regresyon Yöntemi

Zou ve Hastie (2005), LASSO regresyonun bazı eksikliklerine çözüm getirebilmek için Ridge ve LASSO regresyon yöntemlerinin birlikte kullanılması ile Elastik Net regresyon yöntemini önermişlerdir. Kısaca bu eksikliğe değinilecek olunursa; değişken sayısının gözlem sayısından büyük olduğu durumlarda (p>n) LASSO regresyon modele en fazla n değişken seçebilmekte ve bu durum kısıtlayıcı olabilmektedir. Ayrıca veri seti içerisinde aralarında yüksek korelasyona sahip değişken grupları bulunduğu durumlarda LASSO regresyon değişken grubu içerisinden yalnızca birini modele dahil eder ve diğer değişkenleri modelden dışarı atmaktadır.

Genomik veri setlerinde (gen ekpresyonu) benzer işlevlere sahip genler arasında yüksek korelasyon olmasından kaynaklı grup halinde modellenmesi gerekmektedir. Bu durumda LASSO regresyonun tahmin performansı Ridge regresyona göre daha düşük olduğu söylenmektedir. Bu sebeplerden ötürü Elastik Net regresyon yöntemi değişken seçimi ve katsayıları daraltma yaparken ilişkili değişken gruplarını da seçebilmektedir [8]. Elastik Net tahmin edicisi eşitlik (2)'de verilmiştir.

$$\hat{\beta}_{EN} = \underset{\beta}{arg\,min}\left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda_1 \sum_{j=1}^{p}|\beta_j| + \lambda_2 \sum_{j=1}^{p}\beta_j^2\right\} \quad (2)$$

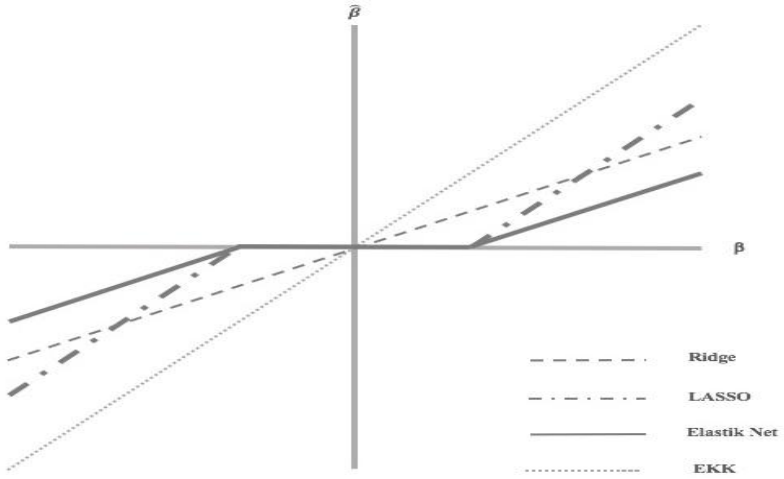
Bu eşitlikte yer alan $\lambda_1 \sum_{j=1}^{p}|\beta_j| + \lambda_2 \sum_{j=1}^{p}\beta_j^2$ Elastik Net tahmin edicisinin ceza terimini ifade etmektedir. Buradan anlaşıldığı üzere Ridge (l_2) ve LASSO (l_1) tahmin edicilerin birlikte kullanılmasından oluştuğu anlaşılmaktadır.

$\alpha = {\lambda_2}/{(\lambda_1 + \lambda_2)}$, $y = (y_1, y_2, \ldots, y_n)^T$ yanıt değişkeni, $X = \left(x_1 | \ldots | x_p\right)$ model matrisi, $|\beta|^2 = \sum_{j=1}^{p}\beta_j^2$ ve $|\beta|_1 = \sum_{j=1}^{p}|\beta_j|$ olsun. Böylece t kısıtı altındaki $\hat{\beta}$ tahmin edicisi Eşitlik (3)'te verilmiştir.

$$\hat{\beta} = \underset{\beta}{arg\,min}|y - X\beta|^2,\ (1-\alpha)|\beta|_1 + \alpha|\beta|^2 \le t \quad iken \quad (3)$$

Bu eşitlikteki $(1-\alpha)|\beta|_1 + \alpha|\beta|^2$ ifade LASSO ve Ridge ceza terimlerinin konveks birleşimi olan Elastik Net ceza terimini (ayar parametresi) ifade etmektedir. Ceza terimindeki α ifadesi $(0 \le \alpha \le 1)$, $\alpha = 1$ iken Ridge regresyon yöntemine; $\alpha = 0$ iken LASSO regresyon yöntemine denk gelmektedir [8].

Ridge, LASSO ve Elastik Net tahmin edicilerin katsayılara etkisi



Şekil 1: Ridge, LASSO ve Elastik Net tahmin edicilerin katsayılara etkisi (Zou ve Hastie (2005))

Ridge, LASSO ve Elastik Net tahmin edicilerinin katsayılara etkisi Şekil 1'de verilmektedir. 45◦'lik noktalı gri çizgi, referans çizgisi olarak kısıtlamasız EKK tahminini göstermektedir. Grafiğe

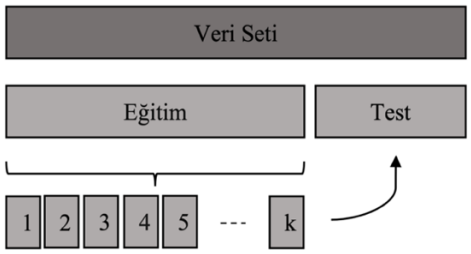
bakıldığında Elastik Net tahmin edicisinin LASSO gibi katsayıları sıfıra indirgeyebildiği ve Ridge tahmin edicisine paralel bir çizgide olduğu görülmektedir. Böylelikle Elastik Net regresyonun Ridge ve LASSO arasında bir ayarlama yapmakta olduğu anlaşılmaktadır.

Gruplama Etkisi (The Grouping Effect)

Bir veri setindeki gözlem sayısının değişken sayısından küçük olması durumuna literatürde gruplandırılmış değişkenler (grouped variables) denmektedir. Genomik veri setleri de bu tip bir veri olmasından dolayı Segal ve Conklin (2003) gruplandırılmış genleri bulmak amacıyla düzenlileştirme kullanılmasını önermişlerdir [8,9]. Gruplama etkisi (grouping effect), herhangi bir regresyon yönteminde aralarında yüksek korelasyon olan bağımsız değişkenlerin oluşturduğu değişken gruplarındaki regresyon katsayılarının eşit olması durumuna verilen isimdir. Elastik Net regresyon yönteminde gruplama etkisini ortadan kaldırmak amaçlanarak eş katsayılara aynı katsayı ataması yapılır [8].

Model parametrelerinin seçimi

LASSO regresyonda yalnızca λ parametresi belirlenirken, Elastik Net regresyonda α parametresinin de ayarlanması gerekmektedir. LASSO ve Elastik Net regresyon yöntemlerinin ayar parametresi (ceza terimi) seçimi çapraz doğrulama ile yapılabilmektedir.



Şekil 2: k-katlı çapraz doğrulama gösterim şeması

k-katlı çapraz doğrulama (n-folds cross validation) yönteminde veri seti ilk olarak eğitim ve test olarak ikiye ayrılmakta sonrasında eğitim seti k eşit parçaya bölünmektedir. Veri setinin büyüklüğüne göre k değeri belirlenmekte ve genellikle 5 veya 10 değerini almaktadır. Doğrulama yönteminde k adet bölünmüş olan olan gruplardan sırasıyla bir grup doğrulama grubu (validation group) olarak ayrı tutularak geriye kalan k-1 adet grup ile model oluşturulur ve daha sonra ayrı tutulan grup ile model test edilir. Oluşturulan k-1 modelin beklenen tahmin hataları karşılaştırılarak minimum hataya sahip olan model seçilir. Böylece doğru tahmin sonucunu veren ayar parametreleri belirlemiş olmaktadır.

Kararlılık Seçimi

Cezalı regresyon modelleri daha az değişken (öznitelik) ile yüksek tahmin performanslı regresyon modelleri kurmaya yardımcı olmaktadır. Fakat iterasyon veya n-folds değeri değiştirildiğinde her seferinde farklı değişkenler seçiliyorsa, seçilen değişkenlerin kararlı olmadığı durumda modele olan güven azalmaktadır. Bu sebeple Meinshausen ve Bühlmann (2010), kararlılık seçimi (stability selection) yaklaşımını önermişlerdir [10].

Kararlılık seçimi, LASSO regresyon gibi değişken seçme yöntemlerinin yeniden örnekleme ile birleştirilerek uygulandığı bir yöntemdir. Yerine koyulmadan çekilmiş alt örneklemlere karşılık gelen değişken seçim yöntemi uygulanarak, her değişken için değişkenin uygun modele dahil edildiği alt örneklemlerin oranı olarak seçim olasılıkları tahmin edilebilmektedir. Tahmin edilen seçim olasılıkları yardımıyla kararlı değişkenler belirlenebilmektedir. Kararlılık seçimi, tahmini kararlı değişken setine yanlış şekilde değişken atayan I. Tip Hata oranlarını kontrol etmek için teorik bir çerçeve sağlamaktadır [10].

Her bir değişkenin düzenlileştirme yolu (regularization path) boyunca seçim olasılığı, kararlılık yolu (stability path) olarak adlandırılmaktadır. Koordinat iniş algoritması yeniden örnekleme yöntemleri ve

düzenlileştirme yolu için hesaplama verimliliği açısından kullanılabilmektedir. Sill ve ark. koordinat iniş algoritmasını kullanılabilmek için öncelikle alt örnekleme ile alt kümeler oluşturulur, ardından koordinat iniş algoritması yardımıyla her alt örneklem için düzenlileştirme yolları hesaplanır ve ortalaması alınarak bir kararlılık yolu hesaplanır [11].

Analiz için kullanılacak araç ve paketler

Sill ve ark. (2014), genomik veriler gibi yüksek boyutlu tahmin modelleri için R programının işlevselliğini geliştirmeyi amaçlayarak geliştirdikleri "c060" isimli R paketi kullanılmıştır [11]. Ek olarak R programında bulunan "glmnet", "epsgo" ve "penalizedSVM" paketleri kullanılarak analize ayarlamalar yapılmıştır [12,13]. Ayrıca Bioconductor isimli açık kaynaklı yazılımlar geliştiren oluşum yardımıyla genomik verilerin analizi yapılabilmektedir [14]. NCBI veri tabanından alınan genomik veri setini analize uygun hale getirilebilmek için Bioconductor'de bulunan "GEOquery", "Biobase" ve "hgu133plus2.db" isimli paketler kullanılmıştır.

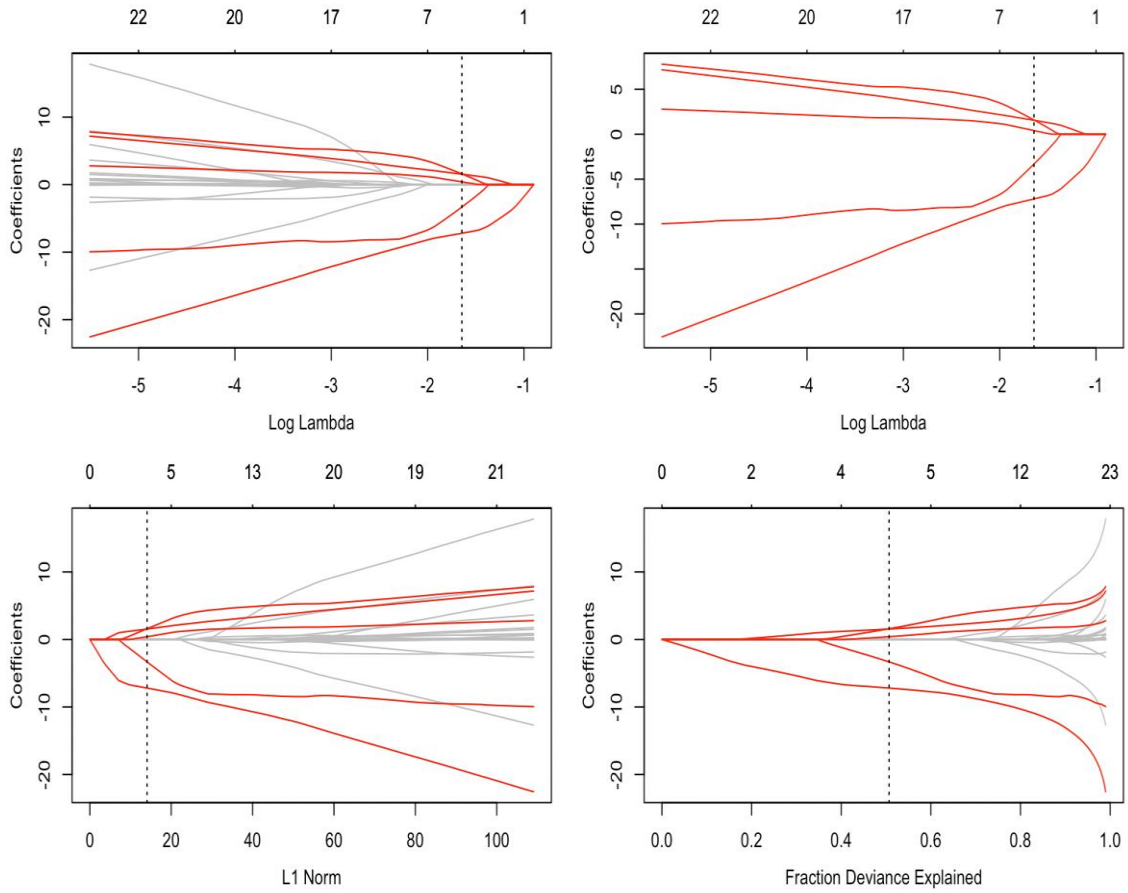
3. Bulgular

Bionconductor yardımıyla NCBI'dan R programına çekilen genomik veri setinin, ekspresyon seti ve fenotip bilgilerini içeren veri matrisi birleştirilerek modelde kullanılacak olan veri oluşturulmuştur. Veri setinde, 30'u hasta ve 24'ü sağlıklı birey olmak üzere toplamda 54 hasta bulunmaktadır. Analize başlamadan önce oluşturulan veri setinin %80'i eğitim, %20'si test olarak ikiye ayrılmıştır. Bu çalışmanın amacı, gen ekspresyonunu içeren veri setleri üzerinde oluşturulan regresyon modellerine uygun genlerin bulunmasıdır. Veri setindeki hastalık durumu (disease state) regresyon modelinin bağımlı değişkeni olarak seçilmiştir. Hastalık durumu iki düzeyli bir değişken (health: 1, chronic obstructive pulmonary disease: 2) olmasından dolayı "glmnet" ve "cv.glmnet" fonksiyonlarında "family" argümanı "binomial" olarak seçilmiştir. Regresyon modeline en uygun LASSO ceza parametresi değerinin belirlenmesi için "cv.glmnet" fonksiyonu kullanılarak 10-katlı çapraz doğrulama (10-folds cross validation) yapılmıştır. Çapraz doğrulama sonucunda minimum lamda değeri λ=0,193 (log λ=−0,714) olarak bulunmuştur. Optimum lamda değeri ile kurulan model sonuçları ve seçilmiş olan özelliklerin katsayı tahminleri Tablo 1'de verilmiştir.

Tablo 1: KOAH veri seti için LASSO regresyon modelinin sonuçları

Gen İfadeleri	**Gen Sembolü**	**Gen Adı**	**Katsayı Tahmini**
Regresyon sabiti	-	-	26,499
200009_at	GDI2	GDP dissociation inhibitör 2	-3,328
203984_s_at	CASP9	Caspase 9	1,586
204491_at	PDE4D	Phosphodiesterase 4D	-7,197
222315_at	LOC100996756	Uncharacterized LOC100996756	1,549
232810_at	AIG1	Androgen induced 1	0,406
Lamda: 0,02 **HKO:** 3,026 **HKOK**: 1,739 **HMO:** 1,416 **R^2**:0,85			

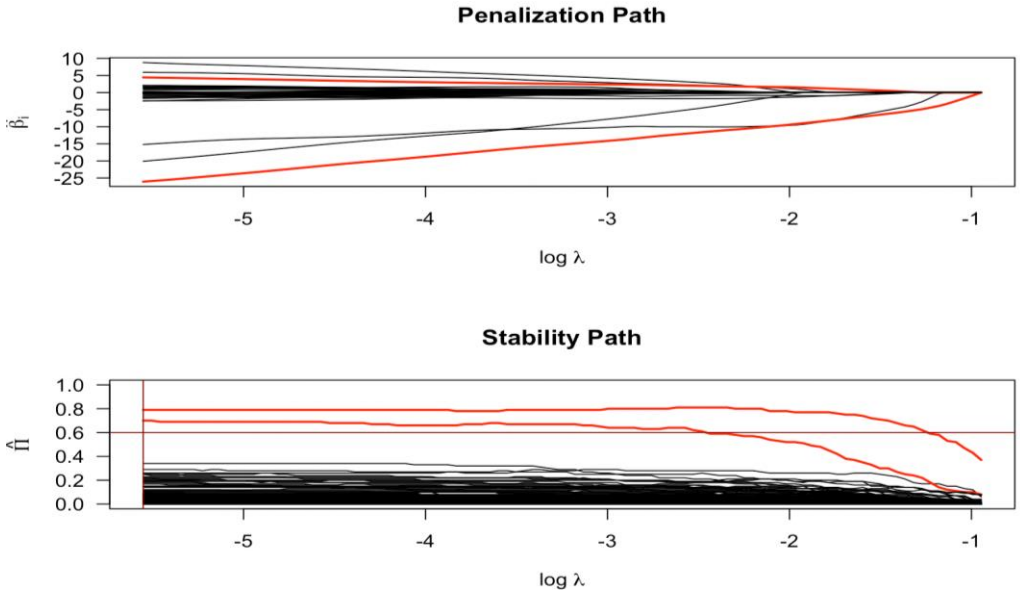
LASSO regresyon modeline göre veri setindeki 54676 adet özellik içerisinden 5'inin modele seçilmiş olduğu görülmektedir. GDI2, CASP9, PDE4D, LOC100996756 ve AIG1 gen sembollerine sahip özelliklerin katsayı tahminleri sırasıyla -3,328, 1,586, -7,197, 1,549 ve 0,406 olduğu görülmektedir. Kurulan LASSO regresyon modelinin HKO 3,026, R^2 değeri ise %85 (0,85) olarak bulunmuştur.



Şekil 3: KOAH veri setine uygulanan LASSO regresyonun katsayılara etkisi

LASSO regresyon modeline göre seçilen özelliklerin regresyon katsayılarının ceza parametresine karşı göstermiş olduğu değişim Şekil 3'te verilmiştir. Grafikteki kırmızı çizgi, 10-katlı çapraz doğrulama ile belirlenen en düşük lamda değeriyle oluşturulan model tarafından seçilen sıfır olmayan katsayıları temsil etmektedir. Sol üstteki grafik bütün katsayıları içerirken sağ üstteki ise yalnızca seçilmiş olan katsayıları içermektedir. Sol alttaki grafik katsayıların L1 normu için katsayı yollarını gösterirken, sağ alttaki grafik ise açıklanan sıfır kısmi en çok olabilirlik sapmasının (the null partial log-likelihood deviance explained) kesrine göre katsayı yollarını göstermektedir [11]. Grafiklerde bulunan dikey çizgiler λ değerini temsil etmektedir. Her grafikte bir dikey çizgi olmasının sebebi minimum sapma ve minimum sapmanın bir standart sapması içindeki en büyük λ değerlerinin aynı çıkmasından kaynaklanmaktadır. λ'nın değeri arttıkça yani log λ'nın değeri düştükçe, modele girecek olan en fazla etki büyüklüğüne sahip özelliklerin sayısı azalmaktadır.

KOAH veri setinde bireylerin hasta veya sağlıklı olmalarının üstünde etkili olan prognostik özellikleri belirlemek amacıyla kararlılık seçimi yöntemi kullanılmıştır. Kararlılık seçiminin performansını arttırmak amacıyla LASSO regresyon modelinin kararlılık yolunu hesaplamak için R'da bulunan "c060" paketindeki "stabpath" fonksiyonu ve bu fonksiyonda yer alan "weakness argument" yardımıyla her özelliğe uygulanan cezalandırmanın üzerine ek olarak yeniden ağırlıklandırma ile "rastgele LASSO" olarak da geçen "ek rastgeleleştirme" yapılmaktadır [10,11]. Stabpath fonksiyonu, ilk olarak alt kümeleri oluşturur takiben "parallel" paketi yardımıyla paralel olarak kararlılık yolunu hesaplar ve stabsel fonksiyonu ile kararlı özellikler tahmin edilir. Tüm işlemler sonucunda PDE4D ve LOC100996756 genlerinin kararlı olduğu anlaşılmıştır.



Şekil 4: KOAH veri setine uygulanan LASSO regresyonun katsayı ve kararlılık yolları

KOAH veri setine uygulanan LASSO regresyonun katsayı ve kararlılık yolları Şekil 4'te verilmiştir. Kırmızı çizgiler ile vurgulanmış olan iki özellik PDE4D ve LOC100996756 isimli kararlı gen ifadelerini belirtmektedir. Çapraz doğrulama (10-folds) ile kurulan LASSO regresyon modeli sonucunda 5 özellik seçilmiştir. Fakat kararlılık seçimi ile bu değişkenlerin tümünün çok kararlı olmadığı ve düzenlileştirme (regularization) miktarı azaldığında (log λ azaldığında) modele giren özellik sayısının arttığı gözlemlenmektedir.

Veri setine Elastik Net regresyon uygulandığında LASSO regresyondan farklı olarak α ve λ parametre değerlerinin birlikte seçilmesi gerekmektedir. Bunun için aralıklı arama algoritması (the interval search algorithm) kullanılmıştır. KOAH veri setine 10-katlı çapraz doğrulama ile uygulanan Elastik Net regresyonun aralıklı arama çıktısının ilk 5 satırı Tablo 2.'de verilmiş ve tablonun altında optimal modelin sonuçları belirtilmiştir. Optimal modelde α=0,99 ve λ=0,07 olarak tespit edilmiştir. Belirlenmiş olan ayar parametrelerine göre model kurulduğunda modelin R^2 değeri 0,92 olarak bulunmuştur.

Tablo 2: KOAH veri setine uygulanan Elastik Net regresyonun aralıklı arama çıktısı

Model	**Alfa**	**Lamda**	**Sapma**	**Değişken Sayısı**
1	0,734	0,086	0,163	40
2	0,853	0,081	0,161	25
3	0,532	0,078	0,158	77
4	0,131	0,530	0,183	234
5	0,608	0,086	0,161	57
Optimal	0,992	0,07	0,150	19
Optimal modelin; SH:0,033 **R^2:** 0,92				

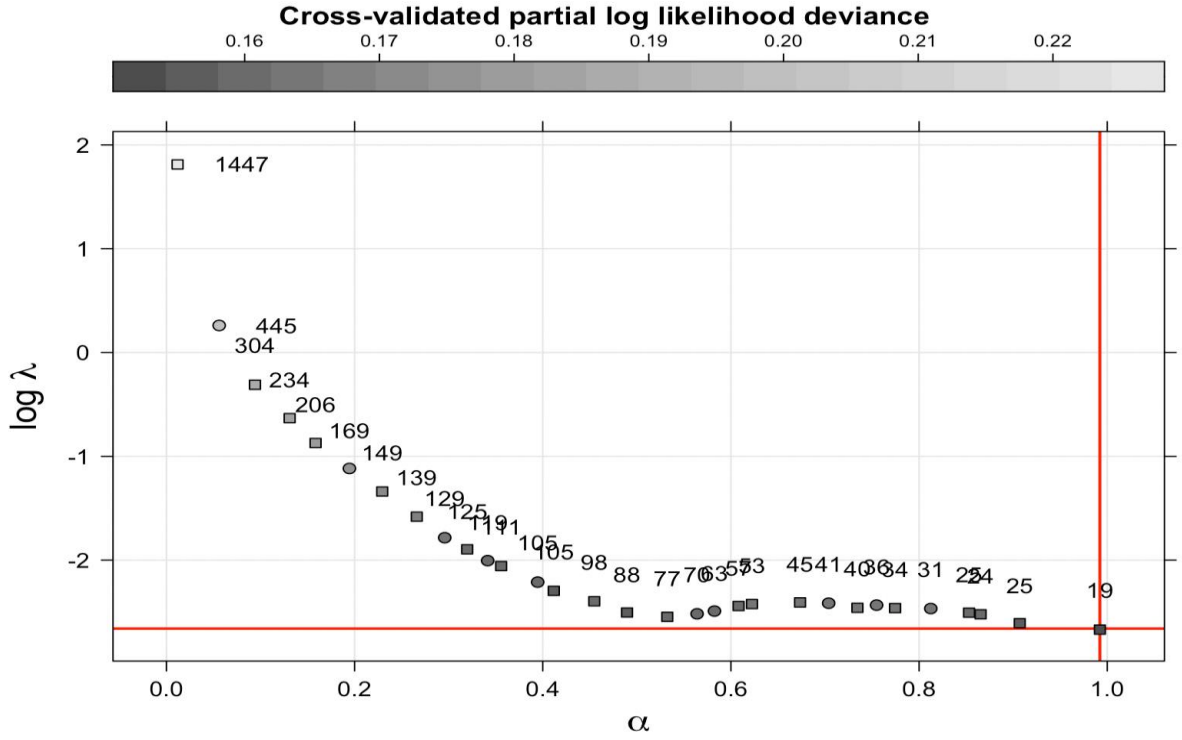
Tablo 3: KOAH veri seti için α=0,99 ve λ=0,07 iken Elastik Net regresyonun katsayı tahmini

Gen İfadeleri	Gen Sembolü	Gen Adı	Katsayı Tahmini
(Intercept)	-	-	77,494
1561445_at	-	-	0,714
1563318_s_at	MAGIX	MAGI family member, X-linked	3,373
200009_at	GDI2	GDP dissociation inhibitor 2	-9,840
200704_at	LITAF	lipopolysaccharide induced TNF factor	-0,290
200862_at	DHCR24	24 dehydrocholesterol reductase	-1,751
204491_at	PDE4D	phosphodiesterase 4D	-12,297
205757_at	ENTPD5	ectonucleoside triphosphate diphosphohydrolase 5	0,170
208112_x_at	EHD1	EH domain containing 1	-5,784
209077_at	TXN2	thioredoxin 2	0,847
217879_at	CDC27	cell division cycle 27	-0,242
218158_s_at	APPL1	adaptor protein, phosphotyrosine interacting with PH domain and leucine zipper 1	2,341
220786_s_at	SLC38A4	solute carrier family 38 member 4	-0,108
222315_at	LOC100996756	uncharacterized LOC100996756	2,127
222629_at	REV1	REV1, DNA directed polymerase	0,694
225420_at	GPAM	glycerol-3-phosphate acyltransferase, mitochondrial	-0,377
227340_s_at	RGMB	repulsive guidance molecule family member b	0,281
229679_at	LOC101060443 ///C12orf76	uncharacterized LOC101060443/// chromosome 12 open reading frame 76	-0,254
231935_at	ARPP21	cAMP regulated phosphoprotein 21	-1,073
242842_at	-	-	0,209

Optimal model sonucunda 19 gen ifadesi seçilmiş ve katsayı tahminleri Tablo 3'te verilmiştir. Gen ifadelerine bakıldığında LASSO regresyon modeli yardımıyla GDI2, PDE4D ve LOC100996756 özelliklerinin Elastik Net regresyon modelinde de seçilmiş olduğu görülmektedir. PDE4D ve

LOC100996756 özelliklerinin, kararlılık seçimiyle de seçilmiş olması bu özelliklerin tutarlı ve kararlı olmasının bir göstergesi olarak yorumlanabilmektedir.

Elastik Net regresyonun α ve logλ parametrelerini ayarlamasının bir fonksiyonu olan çapraz doğrulanmış kısmi log-olabilirlik sapmasının grafiği Şekil 5'te yer almaktadır. Grafikteki yer alan noktalar α değerlerine karşılık gelmekte ve kısmi log-olabilirlik sapması azaldıkça α değerleri açık griden siyaha doğru gitmektedir. Kare sembol başlangıç noktalarını ve daire sembol ise iterasyon noktalarını temsil etmektedir. Her noktanın yanında yer alan sayılar ise o noktaya denk gelen α ve log λ parametreleri ile oluşturulan model sonucunda seçilen gen sayısını belirtmektedir. Kırmızı çizgilerin kesişimindeki nokta ise kayıp fonksiyonun minimum standart hatadaki nihai çözüm noktasını yani optimum modeli temsil eder. Bu nokta α=0,99 ve λ=0,07 olduğu noktaya denk gelerek log-olabilirlik sapmasını minimum yapmaktadır. Bu noktadan 19 genin modele seçildiği tespit edilmektedir.



Şekil 5: KOAH veri seti Elastik Net regresyon modelinin α ve log λ parametrelerini ayarlamasının bir fonksiyonu olan çapraz doğrulanmış kısmi log-olabilirlik sapmasının görselleştirilmesi

4. Tartışma ve Sonuç

Bu çalışmada büyük boyutlu genomik veri setlerinin LASSO ve Elastik Net regresyon yöntemleri kullanılarak analiz edilebileceği, gen setlerinin daha küçük boyutlara indirgenebileceği ve hastalık durumunun daha az değişken ile tahmin edilebileceği gösterilmek istenmiştir. Bunun için NCBI veri tabanından Bioconductor yardımıyla alınan veri seti R programında bulunan "c060", "glmnet", "penalizedSVM" ve "epsgo" paketleri ile analiz edilmiştir.

Model parametrelerini ayarlamak ve optimum modele ulaşabilmek amacıyla 10-katlı çapraz doğrulama yöntemi kullanılmıştır. LASSO regresyon yönteminde ek olarak kararlı ve tutarlı değişkenlerin seçilebilmesi için kararlılık seçimi yöntemi uygulanmıştır. Elastik Net regresyon yönteminde iki parametre birden ayarlanacağından aralıklı arama algoritması kullanılarak α ve λ parametreleri belirlenmiştir.

Elde edilen sonuçlar ile Elastik Net regresyon yönteminin katsayıları daraltma yaparken aynı zamanda ilişkili değişken gruplarını da seçebilmesinden dolayı genomik veri setlerinde iyi bir performans gösterdiği anlaşılmıştır.

Yaptığımız çalışma ile benzer bir çalışmada Kohannim ve arkadaşları, gen merkezli bir LASSO regresyon yaklaşımı kurarak beyin yapısı üzerindeki gen etkilerini keşfetmek istemişler ve büyük miktardaki genomik veriyi eleyerek verimli bir varyant seti oluşturmayı amaçlamışlardır. LASSO regresyon yönteminden faydalanarak her bir gen içindeki ilişkili SNP'ler (Single-nucleotide polymorphism) arasından seyrek SNP alt kümelerini seçerek çalışmalarını desteklemişlerdir. Genom çapında önemli 22 gen keşfetmişler ve LASSO regresyon ile bulunan SNP'lerin p değerlerine göre tek değişkenli GWAS ile bulunanlara göre önemli genler oldukları araştırıcılar tarafından ortaya konmuştur [3]. Cho ve arkadaşları, genom çapında ilişkilendirme (GWAS) çalışmalarında birçok SNP arasından hastalığa neden olan genleri bulmanın çoklu bağlantı sorunu açısından zorluğu olmasından dolayı, çoklu bağlantıyı ele almaya izin veren değişken seçim yöntemi olan Elastik Net regresyonu kullanarak hastalığa neden olan SNP'leri aynı anda tanımlayan bir prosedür önermişlerdir. Birinci adımda, SNP'leri taramak amacıyla tek işaretli ilişkilendirme analizi (the single -marker association analysis) yapılmış ikinci adımda Elastik Net düzenlemesine dayalı çoklu ilişkilendirme analizi (the multiple-marker association) ile taranmıştır. Tarama adımında seçilen SNP'ler genellikle 6. Kromozom üzerinde yer alırken, Elastik Net yaklaşımı artan bir oranda diğer kromozomlar üzerindeki hastalık ile ilişkili olduğu düşünülen SNP'leri tanımlamıştır. Elastik Net regresyon yönteminin SNP belirlemede çeşitli avantajları olduğu belirtilmiştir. Otomatik değişken seçimi ve sürekli daraltma aynı anda gerçekleştirilmesi, çoklu doğrusal regresyonda çoklu bağlantı sorunu oluşturabilecek yüksek derecede ilişkili SNP'den oluşan grupları seçebilmesi, son olarak Elastik Net'in daraltma özelliği sayesinde SNP'ler ve genotipik olmayan faktörler arasındaki etkileşim terimlerini olduğu gibi SNP ana etkilerini de modele dahil etmeyi sağladığını belirtmişlerdir [5]. Yaptığımız çalışma ve yapılan çalışmalarda LASSO ve Elastik Net regresyon yönteminin değişken seçimi ve daraltma özelliği sayesinde genomik veri setleri üzerinde etkili olması konusunda benzerlik göstermiştir.

LASSO ve Elastik Net regresyon yöntemleri sağlık alanında çoklu bağlantı sorununa çözüm olarak farklı çalışmalarda da kullanılabilir.

Etik Beyanı

Bu çalışmada, "Yükseköğretim Kurumları Bilimsel Araştırma ve Yayın Etiği Yönergesi" kapsamında uyulması gerekli tüm kurallara uyulduğunu, bahsi geçen yönergenin "Bilimsel Araştırma ve Yayın Etiğine Aykırı Eylemler" başlığı altında belirtilen eylemlerden hiçbirinin gerçekleştirilmediğini taahhüt ederiz. Bu çalışma "DOĞRUSAL REGRESYONDA RIDGE, LASSO VE ELASTİK NET YÖNTEMLERİNİN SAĞLIK ALANINDA UYGULANMASI" isimli yüksek lisans tezinden uyarlanmıştır.

Kaynakça

[1] Khuri, A. I. 2013. Introduction to Linear Regression Analysis. 4th edition by Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining. International Statistical Review.

[2] Pripp, A. H., Stanišić, M. 2017. Association between biomarkers and clinical characteristics in chronic subdural hematoma patients assessed with lasso regression. PLoS ONE 12(11).

[3] Kohannim, O., et al. 2012. Discovery and Replication of Gene Influences on Brain Structure Using LASSO Regression. Front Neurosci, 6(115).

[4] Çiftsüren, N. M., Akkol, S. 2018. Prediction of internal egg quality characteristics and variable selection using regularization methods: ridge, LASSO and elastic net. Archives Animal Breeding, 61(3), 279-284.

[5] Cho, S., Kim, H., Oh, S., Kim, K., Park, T. 2009. Elastic-net regularization approaches for genome-wide association studies of rheumatoid arthritis. BMC Proc., 3(7), 25.

[6] KOAH Veri seti. 2013. NCBI, National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS4906 (Erişim Tarihi: 10.01.2022).

[7] Tibshirani, R. 1996. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology),58(1), 267-288.

[8] Zou, H., 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.

[9] Segal, M., Dahlquist, K., Conklin, B. 2003. Regression approach for microarray data analysis. J Computnl Biol., 10(6), 961–980.

[10] Meinhausen, N., Bühlmann, P. 2010. Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology),72(4), 417-473.

[11] Sill, M., Hielscher, T., Becker, N., Zucknick, M. 2014. c060: Extended Inference with Lasso and Elastic-Net Regularized Cox and Generalized Linear Models. Journal of Statistical Software, 62(5), 1-22.

[12] Becker, N., Werft, W., Benner, A. 2012. Benner A. penalizedSVM: Feature Selection SVM Using Penalty Functions. R package version 1.1. http://CRAN.R-project.org/package= penalizedSVM. (Erişim Tarihi: 11.10.2022).

[13] Froehlich, H., Zell, A. 2005. Efficient Parameter Selection for Support Vector Machines in Classification and Regression via Model-Based Global Optimization. In Proceedings of the International Joint Conference of Neural Networks., 31 Temmuz-4 Ağustos, Canada.

[14] Bioconductor. https://bioconductor.org (Erişim Tarihi: 07.11.2022).