ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH

İlişkili Bileşen Regresyonu: DNA Hasarını Belirleme Modeli Üzerinde Uygulanması

Correlated Component Regression: Application on Model to Determination of DNA Damage

Sadi ELASAN,[a]
Sıddık KESKİN,[a]
Elif ARI[b]

[a]Biyoistatistik AD,
Yüzüncü Yıl Üniversitesi Tıp Fakültesi
Van
[b]Nefroloji Kliniği,
Kartal Dr. Lütfi Kırdar Eğitim ve
Araştırma Hastanesi,
İstanbul

Geliş Tarihi/*Received:* 13.10.2015
Kabul Tarihi/*Accepted:* 01.02.2016

Yazışma Adresi/*Correspondence:*
Sadi ELASAN
Yüzüncü Yıl Üniversitesi Tıp Fakültesi,
Biyoistatistik AD, Van,
TÜRKİYE/TURKEY
sadielasan@gmail.com

ÖZET Amaç: Açıklayıcı değişken sayısının, örneklem büyüklüğü yaklaştığı veya örnek genişliğini geçtiği durumlarda, diğer bir ifade ile yüksek boyutlu veri setlerinde, regresyon modelleri ile tahminde, güvenilirliğin nasıl artırılacağı önemli sorunlardan birisidir. Bu amaçla kullanılabilecek yeni yöntemlerden birisi, İlişkili Bileşen Regresyonudur. Bu çalışmada, İlişkili Bileşen Regresyonu hakkında bilgi verilerek, bir uygulama ile birlikte tanıtılması amaçlanmıştır. **Gereç ve Yöntemler:** Regresyon analizi yapılması gereken bilimsel çalışmalarda, açıklayıcı değişken sayısı; örnek genişliğine yaklaştığında veya örnek genişliğini geçtiğinde (yüksek boyutlu veri setlerinde), standart regresyon analizi yöntemiyle yapılacak tahminlerde, tahmin edilen katsayılar, çoklu bağlantı (kovaryans matrisinin tekil olması) nedeniyle değişkenlik göstermektedir. Bunun için alternatif bir yöntem olarak, "İlişkili Bileşen Regresyonu" (İBR) problemin çözümüne yardımcı olabilir. Cevap değişkeninin sürekli olması durumunda, İBR-Doğrusal Regresyon ve ikili (binary) olması durumunda, İBR Logistik Regresyon kullanılabilirken, sağkalım verilerinde İBR-Cox Regresyonu kullanılabilir. Yöntem, K adet ilişkili bileşenleri kullanır. Bu ilişkili bileşenler, program tarafından belirlenebileceği gibi araştırıcı tarafından da belirlenebilir. **Bulgular:** Çalışmada örnek genişliğinin küçük, korelasyon katsayılarının orta düzeyde ve değişken sayısının fazla olması nedeniyle, doymuş regresyon modelinde aşırı uyum olduğundan m-kat çapraz geçerlilik testi yapılarak İBR modeliyle bu aşırı uyum elemine edilebilmiştir. **Sonuç:** Regresyon analizlerinde karşılaşılabilecek; çoklu bağlantı, uyum eksikliği veya aşırı uyum gibi sorunların çözümü için İBR'nin kullanılabileceği ve daha yüksek gücü yakalayabileceği söylenebilir.

Anahtar Kelimeler: İlişkili bileşen regresyonu, regresyon çözümlemesi; lojistik modeller; doğrusal modeller; DNA hasarı

ABSTRACT Objective: The number of explanatory variables, where the sample size is approaching or passing the sample size, the high-dimensional data sets in other words, the estimated regression model is one of the important questions of how to increase the reliability. One of the new methods that can be used, Correlated Component Regression Centerproduct. In this study, providing information about the associated Component Regression is intended to be introduced along with an application. **Material and Methods:** Regression analysis in the scientific work to be done, the number of explanatory variables; sample width when approached or when the sample width late (in high-dimensional data sets), an estimate will be made with standard regression analysis method, the estimated coefficients, multiple connections (to be singular of covariance matrix) varies due. As an alternative to this, "Associated Component Regression" (CCR) can help solving the problem. If there are continuous response variables, Iberlinear regression and binary if there is, CCR is used in logistic regression, CCR is available on-Cox regression in survival data. The method uses K number associated components. These related components, as determined by the researchers can be determined by the program. **Results:** In this study sample size is small, the correlation coefficients are moderate and numbers of variables are high. Therefore, we performed m-fold cross-validation test due to extreme overfit of the saturated regression model. **Conclusion:** Encountered in regression analysis; multiple connections, CCR can be used for solving problems such as lack of or excessive integration and said can capture higher power.

Key Words: Correlated component regression, regression analysis; logistic models; linear models; DNA damage

doi: 10.5336/biostatic.2015-48311

Turkiye Klinikleri J Biostat 2016;8(1):45-52

Bilimsel çalışmalarda, ilgilenilen değişken ile açıklayıcı değişkenler arasındaki doğrusal veya doğrusal olmayan ilişkiyi belirlemeye yönelik geliştirilen istatistik analiz yöntemleri, genel olarak regresyon analizi (çözümlemesi) yöntemleri olarak bilinir. Cevap değişkeni ile açıklayıcı değişkenlerin sürekli olması durumunda kullanılan standart çoklu regresyon analizi, varsayımların veya önşartların sağlanması durumunda, güçlü bir analiz yöntemidir. Ancak uygulamalarda, neden-sonuç ilişkisi ve ya tahmin denklemi bulmaya yönelik çalışmalarda, bu varsayımlarla ilişkili olarak; açıklayıcı değişken sayısı, örnek genişliğine yaklaşmakta veya örnek genişliğini geçmektedir. Diğer bir ifade ile yüksek boyutlu veri setleri ile karşılaşılmaktadır. Bu durumda standart regresyon analizi yöntemiyle yapılacak tahminlerde, tahmin edilen katsayılar çoklu bağlantı (kovaryans matrisinin tekil olması) nedeniyle yüksek değişkenlik göstermektedir. Bu problemin çözümü için; Cezalı regresyon yöntemleri (Penalized Regression methods) Lasso ve Elastik ağ (net) içeren "Sınırlı model yaklaşımları" ile Temel bileşenler regresyonu ya da Kısmi En Küçük Kareler Regresyonunu içeren "Boyut İndirgeme Yaklaşımları" olmak üzere iki yaklaşım yaygın olarak kullanılmaktadır. Cezalı regresyon yöntemleri, düzenlileştirme (regularization) için ceza terimini (penalty term) modele dahil etmektedir. Kullanılan bu terime göre de yöntemler arasında farklılıklar bulunmaktadır. Lasso regresyon, ceza terimi olarak "Manhattan Uzaklığı"nı kullanırken, Ridge regresyon "Öklid Uzaklığı"nı kullanır. Elastik ağ ise her iki yöntemin ceza terimini kullanır ve hibrid yaklaşım olarak adlandırılır. Temel Bileşenler regresyonu ile Kısmi En Küçük Kareler regresyonu ise çoklu bağlantı problemini gidermek üzere, öncelikle orijinal değişkenleri temel bileşenlere çevirerek, bu temel bileşenleri cevap değişkenini tahmin etmek üzere modele açıklayıcı değişken olarak alır. Boyut indirgeme yaklaşımlarına alternatif olarak önerilen bir diğer yöntem, İlişkili Bileşen Regresyonudur (Correlated Component Regression).[1]

GEREÇ ve YÖNTEMLER

Bu çalışmada uygulama materyali olarak Van Yüksek İhtisas Hastanesi Nefroloji Polikliniğine başvuran 25 hemodiyaliz hastasına ait veriler kullanılmıştır.[2]

Çalışmada Cevap Değişkeni olarak; DNA Hasarı (8-OHdG/dG, oran) alınmış; Açıklayıcı Değişkenler olarak; Yaş, Vücut Kitle İndeksi (BMI) Sistolik Kan Basıncı (SKB), Diastolik Kan Basıncı (DKB), Total Kolesterol (TKOL), Düşük Yoğunluklu Kolesterol (LDL), Yüksek Yoğunluklu Kolesterol (HDL), Trigliserit (TRIG), C-Reaktif Protein (CRP), Paratroid Hormon (PTH), Albümin (ALB), Katalaz (CAT), Malondialdehid (MDA), Glutasyon Peroksidas (GPX), Superoksidaz Mutaz (SOD) olmak üzere 15 değişken alınmıştır. Bu değişkenlerin DNA hasarı ile doğrusal ilişkili olabilecekleri varsayılmıştır. Bu değişkenlere ait tanımlayıcı istatistikler Tablo 1'de sunulmuştur.

İlişkili Bileşen Regresyonu ilk olarak Magidson (2010) tarafından geliştirilmiş olup, değişken sayısının gözlem sayısına yaklaşması veya gözlem sayısını geçmesi durumunda, diğer bir ifade ile yüksek boyutlu verilerde kullanılan bir yöntemdir. Yüksek boyutlu verilerde, bilinen klasik yöntemlerin kullanılması durumunda, karşılaşılabilen önemli sorunlardan birisi de aşırı uyum veya uyum eksikliğidir. Bunun yanı sıra,

TABLO 1: Değişkenlere ait tanımlayıcı istatistikler.

	N	Ort.	St. Hata	Min.	Mak.
DNA	25	2,396	0,852	1,09	3,90
YAŞ	25	43,680	17,911	20,00	73,00
BMI	25	21,520	4,202	14,60	34,00
SKB	25	123,600	21,531	90,00	165,00
DKB	25	70,200	14,612	45,00	100,00
TKOL	25	159,952	57,689	87,70	298,00
LDL	25	93,188	25,587	58,00	149,00
HDL	25	34,284	4,515	24,00	41,00
TRIG	25	145,996	88,448	48,00	477,00
CRP	25	4,048	2,544	0,30	7,80
PTH	25	1237,136	182,158	1008,00	1763,00
ALB	25	4,0160	0,443	2,50	4,70
CAT	25	35,680	4,488	27,00	44,00
MDA	25	5,7141	2,050	3,45	10,45
GPX	25	7,623	4,259	2,04	17,86
SOD	25	5,179	1,823	1,41	7,82

çoklu bağlantı probleminin olması durumunda, İBR diğer yöntemlere göre daha etkin bir yöntem olarak kullanılabilmektedir.[1]

İBR düzenlileştirme (regularization) algoritması olarak bilinir ve cevap değişkeninin tipine göre bu algoritmalar farklı şekilde adlandırılır. Cevap değişkeninin sürekli olması durumunda; İBR-Doğrusal Regresyon, ikili (binary) olması durumunda İBR-Logistik Regresyon ve cevap değişkeninin sağkalım (survival) verileri olması durumunda İBR-Cox Regresyon olarak adlandırılır.[3]

DÜZENLİLEŞTİRME ALGORİTMASI

İlişkili Bileşen Regresyonu algoritması, K adet ilişkili bileşeni kullanır. Bu ilişkili bileşenler, program tarafından belirlenebileceği gibi araştırıcı tarafından da belirlenebilir. Temel Bileşenler Analizinde olduğu gibi her bileşen orijinal değişkenlerin doğrusal kombinasyonudur. Ancak Temel Bileşenler Analizinden farklı olarak, İBR'de ağırlıklar, cevap değişkenini (Y) tahmin etmede en yüksek tahmin gücüne sahip olacak bileşenler şeklinde belirlenir. İBR, cevap değişkenini tahmin etmek için P adet açıklayıcı değişken yerine, K adet ilişkili bileşeni kullanır. Bu K adet bileşen vektör olarak; $S = (S_1, S_2, \ldots, S_k)$ şeklinde yazılabilir ve genelde K, açıklayıcı değişken sayısından (P) daha az olur yani K<P dir. Her bileşen (S_k), açıklayıcı değişkenlerin (orijinal değişkenlerin) doğrusal kombinasyonudur [$X= (X_1, X_2, \ldots, X_p)$] ve tahminleme modelinde, orijinal değişkenlerin yerine düzenlileştirme modeli olarak kullanılır. Birinci bileşen S_1 cevap değişkeni üzerine en çok etkili olan orijinal değişkenleri bulur. İBR-Doğrusal Regresyon algoritması aşağıdaki gibi çalışır.[1,3,4]

g = 1, 2, … , P olmak üzere S_1 üzerinde her bir açıklayıcı değişkenin yükü olan $\lambda_g^{(1)}$ tahmin edilir. Bu katsayılar Y için regresyon denklemindeki basitçe;

$$\lambda_g^{(1)} = \frac{\mathrm{cov}(Y, X_g)}{\mathrm{var}(X_g)}$$

şeklindeki regresyon katsayıları olarak düşünülebilir. Daha sonra S_1, birinci bileşendeki açıklayıcı değişkenlerin ağırlıklandırılmış ortalama etkisi;

$$S_1 = \frac{1}{P}\sum_{g=1}^{P} \lambda_g^{(1)} X_g$$

şeklinde tanımlanır. Birinci bileşen İBR modelindeki Y için tahmin edicilere ait katsayılar, S_1 ile Y arasındaki basit standart doğrusal regresyon modelinden elde edilir. Benzer şekilde ikinci bileşen İBR modeli için tahmin edicilere ait katsayılar da Y ile S_1 ve S_2 arasındaki basit standart regresyon modeli ile tahmin edilir. Buradaki ikinci bileşen olan S_2, baskılayıcı (suppressor) değişkenlerin etkisini bulmaya çalışır ki, bu değişkenler doğrudan etkiye sahip bir veya daha fazla açıklayıcı değişkenden ilgisiz varyasyonun uzaklaştırılması ile tahmin edilir.[1,3,4]

1. Adım: Birinci Bileşenin Tahmini

Birinci bileşen, P adet açıklayıcı değişkenin modele dahil edilmesi ile Y için bütün açıklayıcı değişkenlerin "1-tahmin edici model" de ağırlıklandırılmış ortalamasıdır.

Buna göre model:

$$y = \alpha_g^{(1)} + \lambda_g^{(1)}\chi_g + \varepsilon_g^{(1)} \quad g = 1,\ 2, ...,\ P$$

olarak yazılır. Buradan birinci bileşen olan S_1, "1-tahmin edici model" den bulunan λ_g parametrelerinin doğrusal kombinasyonu olarak ve $\alpha_g^{(1)}$ ihmal edilerek;

$$S_1 = \sum_{g=1}^{P} \lambda_g^{(1)}\chi_g$$

olarak yazılır. Bu bileşen, "1-tahmin edici model" de Y nin X'lere olan regresyonunu, için X'lere ait doğrudan etkileri belirleyen bileşendir.

Buradan "1-Bileşen İBR" modeli;

$$\hat{Y} = \alpha^{(1)} + b_1^{(1)} S_1$$

olarak yazılır.

Burada S_1 açıklayıcı değişkenlere (X'lere) ait vektörden oluşur. Görüldüğü üzere burada α ve b katsayıları Y için birinci bileşen olan S_1 ile oluşturulan regresyon eşitliğinden yeniden tahmin edilir. Buradan eşitlik, X'lere ait düzenlileştirilmiş etkileri belirlemek üzere;

$$\hat{Y} = \alpha^{(1)} + b_1^{(1)} \sum_{g=1}^{P} \lambda_g^{(1)}\chi_g$$

şeklinde yazılabilir. Bu eşitlik basitleştirilerek;

$$\hat{Y} = \alpha^{(1)} + \sum_{g=1}^{P} \lambda^{*(1)}_{g} \chi_g$$

olarak yazılır. Burada; $\lambda^{*(1)}_{g} = b^{(1)}_{1} \lambda^{(1)}_{g}$ olup, bunlar Xg'ler için "1-Bileşen İBR" modelinde düzenlileştirilmiş katsayılardır.

2. Adım: İkinci Bileşenin Tahmini

İkinci bileşen; $\lambda^{(2)}_{g} \chi_g$ (g = 1, 2, …, g)'lerin ağırlıklandırılmış ortalaması olarak tanımlanır. Burada her bir $\lambda^{(2)}_{g}$, aşağıdaki "2-tahmin edici model" den tahmin edilir.

$$y = \alpha^{(2)}_{g} + \gamma^{(2)}_{1.g} S_1 + \lambda^{(2)}_{g} \chi_g + \varepsilon^{(2)}_{g} \qquad g = 1, 2, ..., P$$

S_2, S_1'in kontrol edilmesi ile regresyon katsayılarının ortalaması olduğu için, sabit ve her bir X için modeldeki S_1'e ait katsayıların ihmal edilmesi ile ikinci bileşen; $S_2 = \sum_{g=1}^{p} \lambda^{(2)}_{g} \chi_g$

olarak yazılır. Buradan "2-Bileşen İBR" modeli S_1 ve S_2 bileşenlerine dayalı olarak; $\hat{Y} = \alpha^{(2)} + b^{(2)}_{1} S_1 + b^{(2)}_{2} S_2$

şeklinde yazılır. Bu yeni modelde S_1 ve S_2, X değişkenlerine ait bileşenler vektörüdür. Buradan α, ve sırası ile S_1 ve S_2 ye ait regresyon katsayıları olan $b^{(2)}_{1}$ $b^{(2)}_{2}$, yeniden tahmin edilir. Bu durumda eşitlik, X'lere ait düzenlileştirilmiş etkileri belirlemek üzere; X'lerin kombinasyonu;

$$\hat{Y} = \alpha^{(2)} + \sum_{g=1}^{P} \lambda^{*(2)}_{g} \chi_g$$

şeklinde yazılır. Burada $\lambda^{*(2)}_{g} = b^{(2)}_{1} \lambda^{(1)}_{g} + b^{(2)}_{2} \lambda^{(2)}_{g}$ olup, bunlar "2-Bileşen İBR" modelinde X_g'ler için düzenlileştirilmiş katsayılardır.

3. Adım: Üçüncü ve K. Bileşenin Tahmini

Bu işlem, kalan K adet bileşen için $\lambda^{(K)}_{g} \chi_g$ (g = 1, 2, …, g)'lerin ağırlıklı ortalaması olarak tanımlanır. Burada her bir $\lambda^{(K)}_{g}$, aşağıdaki "2-tahmin edici model" den elde edilir.

$$Y = \alpha^{(K)}_{g} + \gamma^{(K)}_{1.g} S_1 + \gamma^{(K)}_{2.g} S_2 + ... + \gamma^{(K-1)}_{(K-1).g} S_{K-1} + \lambda^{(K)}_{g} \chi_g + \varepsilon^{(K)}_{g}$$
$$g = 1, 2, ..., P$$

ve S_K; $S_K = \sum_{g=1}^{p} \lambda^{(K)}_{g} \chi_g$

olarak yazılır. S_1 den S_{K-1}'e kadar olan bileşenler için katsayıların ve sabitlerin ihmal edilmesi ile (çünkü S_K, S_1 den S_{K-1} ya kadar olan katsayıların kontrolü ile regresyon katsayılarının ortalamasıdır), "K-Bileşen İBR" modeli; S_1 ,…, S_K bileşenlerine dayalı olarak;

$$\hat{Y} = \alpha^{(K)} + b^{(K)}_{1} S_1 + b^{(K)}_{2} S_2 + ... + b^{(K)}_{K} S_K$$

şeklinde yazılır. Burada S_1 ,…, S_K, X değişkenlerine ait bileşenler vektörüdür.

$b^{(K)}_{j}$ (j =1, 2, …, K), "K-Bileşen İBR" modelinde S_j bileşenlerine ait regresyon katsayılarıdır. Buradan eşitlik, X'lere ait düzenlileştirilmiş etkileri belirlemek üzere;

$$\hat{Y} = \alpha^{(K)} + \sum_{g=1}^{P} \lambda^{*(K)}_{g} \chi_g$$

şeklinde yazılabilir. Burada;

$\lambda^{*(K)}_{g} = b^{(K)}_{1} \lambda^{(1)}_{g} + b^{(K)}_{2} \lambda^{(2)}_{g} + ... + b^{(K)}_{K} \lambda^{(K)}_{g}$ katsayıları, Xg'ler için "K-Bileşen İBR" modelinde düzenlileştirilmiş katsayılardır.

M-Kat Geçerlilik

Çapraz geçerlilik yöntemi, ilişkili bileşenler analizi için M-kat geçerlilik olarak bilinir. Tipik olarak birden fazla çalıştırılması için M-kat geçerlilik gereklidir. Her aşamada çapraz geçerlilik için performans istatistikleri sağlanır. Böylece bir örnek için her aşama bir gözlem olarak düşünülür. Bu durum her bir çapraz geçerlilik istatistiği için standart hataların ve ortalamaların hesaplanmasını sağlar. Genel kural olarak örnekleme için 50-100 aşama önerilmektedir.[4,5]

İndirgeme İşlemi

Değişken seçme basitçe indirgeme prosedürünü kullanır. Bu prosedür ise M-kat çapraz geçerlilikle ilişkili olarak çalışır. (İlişkili bileşen) K'nın özel bir değeri için başlangıçta model mümkün olan bütün açıklayıcı değişkenlerle tahmin edilir. Daha sonra bu model M-kat geçerlilik ile değerlendirilir. Bunun için R^2 ve doğruluk değerleri hesaplanır. Daha sonra bu modelden eğitim örneği üzerine tahmin yapılır. Model katsayıları standardize edilir. Standart doğrusal regresyon mode-

linde olduğu gibi bu süreç standardize edilmemiş katsayılar ile standardize edilmiş katsayıların çarpımı ve bunların cevap değişkenlerine ait standart sapmaya bölümünü içerir.[4]

En düşük standardize edilmiş etkiye sahip olan açıklayıcı değişken eğitim örneğinden çıkarılır. Ve bu süreç yeniden tekrar edilir. Bu durumda, elde edilen R^2 bir öncekinden çıkarılır. Bu süreç, K'nın 1 ile maksimum sayısı olan açıklayıcı değişken arasındaki tüm değerleri için R^2 elde edene kadar devam eder. Elde edilen her model için çapraz geçerlilik performansı değişkenlerin maksimum sayısından bir tahmin edici değişken arasında süreçte belirlenir. Ve program doğrusal regresyon ya da logistik regresyon için en düşük çapraz geçerlilik R^2 değerini dikkate alarak optimum tahmin edici sayısını belirler. Optimum tahmin edici sayına karar verildiğinde bu indirgeme süreci bütün eğitim örneği için uygun tahmin edici değişkenlerin sayısının belirlenmesi için yeniden çalıştırılır. Kullanıcılar, alternatif olarak, modellerin çapraz geçerlilik değerlerini karşılaştırarak tahmin edici değişken sayını belirleyebilir. Bu süreç klasik regresyon analizi programlarındaki değişken seçme yöntemlerinden birisi olan geriye doğru eliminasyon yöntemine benzerlik göstermektedir.[4]

Final İBR Modeli

Bütün bilgilerin sağlanabilmesi için düzenlileştirilmiş son İBR modeli, optimum K ve P sayısı ile bütün eğitim örneğinde tahmin edilir. Çapraz geçerlilik ve indirgeme süreci optimum K ve P değerlerinin seçimi için ve tahmin edici değişkenlerin önemliliği için ek bir bilgi sağlamaktadır.[3,4]

İBR'nin Doğrusal Ayırma Analizi ve Logistik Regresyon için Genişletilmesi

Cevap değişkeni ikili (binary) olduğu durumlarda logistik regresyon ve doğrusal ayırma analizi kullanılmaktadır. İBR modeli de bu iki analiz için benzer işlem sürecini izlemektedir.[3,5]

BULGULAR

Değişken arası Pearson Korelasyon Katsayıları Tablo 2'de verilmiştir. Tablo 2'de görüldüğü üzere, en yüksel korelasyonlar sırasıyla DNA hasarı ile; LDL (0,643), CRP (0,497) ve SKB (-0,412) arasında izlenmiştir. Klinik olarak yorumlandığında; hastalarda LDL düzeyindeki artışa bağlı olarak % 64,3 oranında DNA hasarında da artış meydana gelmesi beklenmektedir. Bunun yanı sıra cevap değişkeni olan DNA ile CRP pozitif korelasyonlu bulunurken, SKB ile negatif korelasyonlu bulunmuştur.

TABLO 2. Korelasyon katsayıları.

	YAS	BMI	SKB	DKB	TKOL	LDL	HDL	TRIGL	CRP	PTH	ALB	CAT	MDA	GPX	SOD	DNA
YAŞ	1															
BMI	,199	1														
SKB	,058	,271	1													
DKB	-,280	,084	***,577****	1												
TKOL	,027	,129	,108	-,082	1											
LDL	-,180	,121	-,197	,245	,154	1										
HDL	,020	,081	,101	-,184	-,145	-,129	1									
TRIG	-,162	-,046	-,068	,033	***,460****	***,430****	-,179	1								
CRP	-,162	,018	-,359	,065	,156	***,396****	,055	,075	1							
PTH	-,253	-,129	-,188	-,076	-,237	,044	-,091	-,047	,147	1						
ALB	-,198	-,302	-,087	-,010	-,236	,308	-,057	,285	-,258	,139	1					
CAT	,284	,033	,025	-,142	,216	-,100	,120	,100	,264	-,193	-,387	1				
MDA	-,005	,076	,054	-,175	,370	,046	-,367	,087	,017	-,273	,019	,222	1			
GPX	-,171	-,291	-,237	-,236	,022	,100	,337	-,031	,175	,129	,303	,144	-,015	1		
SOD	-,318	-,089	-,051	-,150	-,143	-,276	,330	-,281	-,333	,031	-,197	,053	-,093	,158	1	

** p<0.01 ; * p< 0.05.

Çalışmada, ilişkili bileşen regresyonunun yanı sıra karşılaştırma amaçlı olarak Standart regresyon, Ridge regresyon ve Temel Bileşenler regresyonu analizleri için Açıklayıcı değişkenlere ait Regresyon katsayıları ve Standart hataları Tablo 3'te verilmiştir.

Tablo 3'te görüldüğü üzere Standart regresyon analizi sonuçlarına göre R^2 %84 ve %5 düzeyinde anlamlı olarak bulunmuş olmasına rağmen, açıklayıcı değişkenlere ait regresyon katsayılarından hiçbiri istatistik olarak anlamlı bulunmamıştır. Regresyon katsayılarının çoğunda standart sapmalar yüksek yani regresyon katsayıları aşırı değişkenlik göstermektedir. Diğer yandan regresyon analizi göreceli olarak küçük örnekte yapıldığından ve her ne kadar Tablo 2'deki korelasyon katsayıları orta düzeyde olsa da modelin yaklaşık %84'lük R^2 ile aşırı uyum gösterme olasılığı yüksektir. Bununla birlikte bu durum çoklu bağlantı sorununa da işaret etmektedir. Dolayısıyla aşırı uyum, çoklu bağlantı problemlerinin yanı sıra değişken sayısının gözlem sayısına yaklaşması nedeniyle bu verilere İlişkili Bileşenler Regresyonu yapılması öngörülmüştür.

İlişkili bileşenler regresyonu analizinde ilk olarak açıklayıcı değişkenlerin tümü modele dahil edilmiş ve buna ilişkin Regresyon katsayıları (b_i) ve standart hataları (S_{bi}) Tablo 3'te verilmiştir. Daha önce de belirtildiği üzere, bütün değişkenlerin modele dahil edilmesi (doymuş model) ile elde edilen sonuçlar standart regresyon analizi sonuçlarıyla eşdeğerdir. İlişkili bileşenler regresyonu düzenlileştirme için uygun bileşen sayısını belirlemede çapraz geçerlilik R^2 (belirleme katsayısı) istatistiğini kullanmaktadır. M-kat çapraz geçerlilik testinde m'nin 5 ile 10 arasında belirlenmesinin uygun olacağı düşünülmüştür. Gözlem sayısı 25 olduğundan, kolay bölünebilme amacıyla m=5 ve asgari döngü (round) sayısı 100 alınarak, İBR analizi yapılmış ve sonuçlar Tablo 3'te verilmiştir. Tablo 3'te görüldüğü üzere; katsayılar tahmin edilirken, örneğin LDL'ye ait döngü sayısı 500 iken, CRP ye ait döngü sayısı 416 olarak gözlenmiştir.

Açıklayıcı değişkenlere ait regresyon katsayılarının bir kısmında çapraz geçerlilik sonucunda işaret değişikliği gözlenmiştir. Örneğin GPX ve TRIG'e ait regresyon katsayıları negatif iken, pozitif işaretli bulunmuş. Diğer yandan CAT ve MDA'ya ait regresyon katsayıları pozitif iken de negatif işaretli bulunmuştur. Bu analiz sonucunda Tablo 3'te görüldüğü üzere, doymuş regresyon

TABLO 3. Regresyon analizi sonuçları.

	Standart Regresyon		Ridge Regresyon	Temel Bileşenler Regresyonu	İlişkili Bileşen Regresyonu P =15	İlişkili Bileşen Regresyonu m= 5, Çapraz Geçerlilik	
	$b_i \pm S_{bi}$	p	$b_i \pm S_{bi}$	$b_i \pm S_{bi}$	b_i	b_i	Döngü Sayısı
Regr. Sabiti	-5,711	0,150	-5,4097	-4,5500	-5,711	0,724	
YAS	-0,001 ±0,011	0,973	-0,0004 ± 0,011	0,0032 ± 0,012	-0,001	-0,006	246
BMI	-0,005 ±0,033	0,883	-0,0053 ± 0,034	-0,0064 ± 0,038	-0,005	-0,012	179
SKB	-0,030 ±0,013	0,051	-0,0278 ± 0,012	-0,0103 ± 0,007	-0,030	-0,008	***352***
DKB	0,034 ±0,018	0,094	0,0314 ± 0,017	0,0104 ± 0,013	0,034	0,003	177
TKOL	0,006 ±0,003	0,112	0,0052 ± 0,003	0,0037 ± 0,003	0,006	0,001	147
LDL	0,011 ±0,007	0,167	0,0109 ± 0,007	0,0112 ± 0,008	0,011	0,010	***500***
HDL	0,102 ±0,046	0,055	0,0969 ± 0,045	0,0393 ± 0,032	0,102	0,012	186
TRIG	-0,003 ±0,002	0,210	-0,0027 ± 0,002	-0,0020 ± 0,002	-0,003	0,001	197
CRP	0,002 ±0,105	0,988	0,0097 ± 0,100	0,1507 ± 0,067	0,002	0,078	***416***
PTH	0,002 ±0,001	0,098	0,0015 ± 0,001	0,0008 ± 0,001	0,002	0,001	***298***
ALB	0,767 ±0,502	0,161	0,7530 ± 0,495	1,0167 ± 0,511	0,767	0,342	***348***
CAT	0,007 ±0,038	0,852	0,0055 ± 0,038	-0,0134 ± 0,042	0,007	-0,020	220
MDA	0,012 ±0,089	0,892	0,0063 ± 0,088	-0,0673 ± 0,087	0,012	-0,046	241
GPX	-0,034 ±0,038	0,387	-0,0325 ± 0,038	-0,0306 ± 0,043	-0,034	0,023	228
SOD	-0,094 ±0,122	0,462	-0,0872 ± 0,118	0,0421 ± 0,106	-0,094	-0,049	225
	$R^2 \pm SR2$ =,840±,555 p=0,043		$R^2 \pm S_R{}^2$=,832±,569	$R^2 \pm S_R{}^2$=,841±,555	$R^2 \pm S_R{}^2$=,840±,555	R^2 = 0,655	

TABLO 4. Final modellere ait katsayılar.

	Final Model P=5, K= 5				Final Model P=5, K=1			
	Katsayılar		Bileşen Ağırlıkları		Katsayılar		Bileşen Ağırlıkları	
	Standardize Edilmemiş	Standardize Edilmiş	Standardize Edilmemiş	Standardize Edilmiş	Standardize Edilmemiş	Standardize Edilmiş	Standardize Edilmemiş	Standardize Edilmiş
Regr. Sabiti	-2,014				-0,859			
LDL	0,012	0,368	0,628	0,847	0,013	0,382	0,021	0,477
CRP	0,117	0,351	0,794	0,113	0,099	0,296	0,166	0,369
SKB	-0,006	-0,151	0,955	0,065	-0,010	-0,245	-0,016	-0,306
PTH	0,001	0,188	0,905	0,042	0,001	0,195	0,002	0,243
ALB	0,609	0,317	0,200	0,025	0,433	0,225	0,728	0,281
	R^2 = 0,655				R^2 = 0,643			

modelinde aşırı uyumu, m-kat (m=5) çapraz ge çerlilik testi yapılarak İBR modeliyle bu aşırı uyum elemine edilebilmiş ve döngü sayısına bakılarak modele en fazla katkı sağlayan değişkenlerin; LDL, CRP, SKB, PTH ve ALB olduğu belirlenmiştir.

Bu değişkenlere göre P=5 ve K=5 olmak üzere yeniden İBR yapılmış ve analiz sonuçları Tablo 4'te verilmiştir. Bu sonuca göre modele ait R^2 %65,5 olarak bulunmuştur. Bunun yanı sıra açıklayıcı değişkenlerden CRP ve ALB'ye ait regresyon katsayıları belirgin değişim göstermiştir. Ve son olarak P=5 ve K=1 alınarak İBR yapılmış ve R^2 %64,3 olarak bulunmuştur.

K=1 olan model ile K=5 olan model arasında R^2 bakımından belirgin değişim gözlenmediğinden K=1 olan modelin final model olarak benimseneceği söylenebilir. Bu modele göre de bileşenin standardize edilmiş ve edilmemiş bileşen ağırlık değerleri sırasıyla; 0.595 ve 0.802 olarak bulunmuştur (Tablo 5).

TABLO 5. Final modellere ait bileşen ağırlığı.

		Bileşen ağırlığı	
Model		Standardize Edilmemiş	Standardize Edilmiş
p=5, k=1	1	0,595	0,802
	1	0,628	0,847
	2	0,794	0,113
p=5, k=5	3	0,955	0,065
	4	0,905	0,042
	5	0,200	0,025

Örnek genişliğinin düşük olması, korelasyon katsayılarının orta düzeyde olması ve değişken sayısının fazla olması nedeniyle doymuş regresyon modelinde aşırı uyum olduğundan m-kat çapraz geçerlilik testi yapılarak İBR modeliyle bu aşırı uyum elemine edilebilmiştir. Diğer yandan 5 ve 1 bileşenli modellerin daha iyi sonuçlar verdiği ve adı geçen bu iki bileşen modeli arasında da fark olmadığını söylemek mümkündür.

TARTIŞMA ve SONUÇ

Son zamanlardaki gelişmeler; yüksek boyutlu verilerde, özellikle açıklayıcı değişken sayısının örnek genişliğinden fazla olması durumunda regresyon modelleri ile tahminlemedeki güvenilirliğin nasıl arttırılacağı üzerine yoğunlaşmıştır.

Tibshirani (1996) çalışmasında, Lasso ve Ridge regresyonun performanslarını incelemiş ve n>p (n: gözlem sayısı, p: açıklayıcı değişken sayısı) olduğu durumlarda Lasso regresyonun performansının Ridge regresyonun performansından daha yüksek olduğunu belirtmiştir.[6] Benzer şekilde, Zou ve Hastie (2005), p>n olduğu durumlarda Elastik ağ'ın kullanılabileceğini önerirken, bu koşullarda Lasso'un tatminkar bir değişken seçme yöntemi olmadığı vurgulamıştır ve Elastik ağ'ın Lasso'dan daha çok modele açıklayıcı değişken dahil etme eğiliminde olduğunu ve bunun özellikle açıklayıcı değişkenler arasında yüksek korelasyon olduğu durumda gerçekleştiğini vurgulamıştır.[7] Diğer yandan Fu (1998) değişik küçültme veya daraltma parametrelerini (shrinkage parameter) (γ) kullanarak Bridge regresyon, Standart reg-

resyon (γ=0), Lasso regresyon (γ=1) ve Ridge regresyon (γ =2) yöntemlerini bir simülasyon çalışması ile karşılaştırmış ve çalışma sonucunda Bridge regresyonun diğerlerinden daha iyi performans gösterdiğini belirtmiştir.[8]

Bu çalışmada; yüksek boyutlu verilerde, özellikle değişken sayısının gözlem sayına eşit olduğu veya geçtiği durumlarda kullanılabilecek yöntemlere alternatif olarak önerilebilecek yöntem olan İlişkili Bileşen regresyonu tanıtılmıştır. Ayrıca standart doğrusal regresyon, Ridge regresyon ve Temel Bileşenler regresyonu ile karşılaştırmalı olarak değerlendirilmiştir. Bu sonuçlara göre; İlişkili Bileşen regresyonunun belirtilen koşullarda, (değişken sayısının gözlem sayısına yaklaşması ve geçmesi durumunda) standart regresyona göre daha iyi performans gösterdiği söylenebilir. Ayrıca regresyon katsayılarındaki değişkenlik ve aşırı uyum gibi problemleri de elemine etmede başarılı olduğu söylenebilir. Ancak, yöntemin performansını değerlendirmek ve diğer yöntemlerle karşılaştırmalı olarak incelemek üzere simülasyon çalışmalarının gerekliliği unutulmamalıdır.

KAYNAKLAR

1. Magidson J. Correlated component regression: A prediciton/classification methodology for possibly many features. JSM Proceedings of the American Statistical Assocation. Section on Statistical Learning and Data Mining 2010;4372-86.
2. Ari E, Kaya Y, Demir H, Cebi A, Alp HH, Bakan E, et al. Oxidative DNA damage correlates with carotid artery atherosclerosis in hemodialysis patients. Hemodial Int 2011; 15(4):453-9.
3. Magidson J, Wassmann K. The role of Proxy genes in predictive models: an application to early detection of prostate cancer. JSM Proceedings of the American Statistical Association, Biometrics Section 2010;2739-53.
4. Magidson J, Bennet G. Correlated Component Regression (CCR)-A brief methodogical description (Contains a brief technical overview of the CCR Methodology). Technical Papers 2011;1-7.
5. Magidson J. Summary of Correlated Component Regression (CCR). COREXpress User's Guide: Manual for COREXpress by Statistical Innovations Inc. Belmont, MA: Statistical Innovations Inc; 2011. p.64-87.
6. Tibshirani R. Regression Shrinkage and selection via the Lasso. J R Statist Soc B 1996;58(1):267-88.
7. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Statist Soc B 2005;67(2):301-20.
8. Fu WJ. Penalized regression: the bridge versus the lasso. AmericanJournal of Computational and Graphical Statistics 1998;7(3): 397-416.