ÇOKLU DOĞRUSAL BAĞLANTI HALİNDE ENKÜÇÜK KARELER TEKNİĞİNİN ALTERNATİFİ YANLI TAHMİN TEKNİKLERİ VE BİR UYGULAMA

Yrd. Doç. Dr. Ali Sait ALBAYRAK

Zonguldak Karaelmas Üniversitesi
Çaycuma İktisadi ve İdari Bilimler Fakültesi
Sayısal Yöntemler Anabilim Dalı Öğretim Üyesi
asalbayrak@hotmail.com

ÖZET

Bu çalışmanın amacı, beden ağırlığının tahmin edilmesinde yanlı tahmin tekniklerinin [Ridge Regression (RR) ve Principal Component (PC)] enküçük kareler [Least Squares (LS)] tekniğine karşı etkinliğini araştırmaktır. Bu amaçla beden ağırlığı ile açıklayıcı değişkenler arasındaki doğrusal ilişkinin tahmininde LS ve yanlı tahmin tekniklerinin (RR ve PC) göreceli tahmin geçerlilikleri karşılaştırılmaktadır. Araştırmada, bağımsız değişkenler arasındaki yüksek çoklu doğrusal bağlantı problemine dayanarak RR ve PC tekniklerinin LS tekniğine göre daha düşük standart hatalı, durağan ve kuramsal beklentilere uygun tahminler sağlayacağı beklenmiştir.

Anahtar Kelimeler: *Enküçük Kareler Tekniği, Ridge Regresyon Analizi, Temel Bileşenler Regresyonu ve Çoklu Doğrusal Bağlantı.*

ABSTRACT

The purpose of this paper is to examine the effectiveness of applying biased estimation techniques (RR and PC) over Least Squares (LS) technique. For this purpose, the relative predictive validity of three regression techniques was compared using the weight data to study the linear relation of dependent variable to predictor variables. It was hypothesized that, given the high degree of multicolinearity of the predictor variables, biased estimation techniques would provide more stabilized coefficient and less standard errors than would the LS technique.

Keywords: *Least Squares Technique, Ridge Regression, Principal Components Regression and Multicolinearity.*

1. GİRİŞ

Yanlı tahmin teknikleri (RR ve PC), doğrusal regresyon modelinde bağımsız değişkenler arasında güçlü ilişkiler olduğunda enküçük kareler (LS) tekniğine alternatif olarak geliştirilmişlerdir (Shin, 1996:150-157). Çoklu doğrusal bağlantı halinde LS tahminleri yansız olmakta, ancak varyanslar büyüdüğünden tahminler gerçek değerlerinden uzaklaşabilmektedir. RR analizinde, LS tahminlerine küçük bir yanlılık sabiti eklenerek varyanslar küçültülerek daha durağan sonuçlar elde edilebilmektedir. RR tekniğine alternatif kullanılabilecek diğer yanlı tahmin tekniği temel bileşenler (PC) regresyon tekniğidir (Orhunbilge, 2000: 240-251).

Yanlı tahmin tekniklerinin varsayımları LS varsayımlarıyla aynıdır: Doğrusallık, eşvaryans ve bağımsızlık. Ancak yanlı tahmin tekniklerinde güven aralıkları hesaplanmadığından normallik varsayımı yapılmamaktadır (Rawlings, 1998:63-69).

Beden ağırlığının tahmininde kullanılacak açıklayıcı değişkenler arasında yüksek derecede bir örtüşme olacağından, LS yerine yanlı tahmin tekniklerinin kullanılması daha güvenilir bir yaklaşım olarak kabul edilmektedir. Diğer taraftan çoklu doğrusal bağlantılı verilerde örnekten örneğe veya yıldan yıla tahmin edilecek doğruların birbirinden anlamlı bir şekilde farklılaşacağından, durağan ve güvenilir tahmin doğruları elde etmek güçleşmektedir. Bu yüzden, bu tür regresyon doğrularının geçerliliğinin sınanmasında çok büyük örneklere gereksinim duyulmaktadır (Maxwell, 2000:434-458). Bu durumda LS yerine yanlı tahmin tekniklerinin kullanılmasının en uygun yaklaşım olduğu kanıtlanmıştır (Vinod, 1995:287-302). Kısaca, yanlı tahmin teknikleri çoklu doğrusal bağlantılı verilerde daha durağan tahmin doğrularının elde edilmesi ve LS tekniğinin uygun bir şekilde uygulanabilmesi için gerekli örnek büyüklüğünden daha küçük örnekler kullanılması durumunda daha etkili, geçerli ve durağan sonuçlar sağlamaktadır (Vinod, 1995:287-302).

Tracey, Sedlacek ve Miars (1983), SAT puanları ile lise not ortalamalarına göre öğrencilerin üniversite birinci sınıf birikimli not ortalamalarını tahmininde ridge regresyon ve enküçük kareler tekniğinin sonuçlarını karşılaştırdılar. Çalışmalarında çapraz geçerlilik korelasyonlarına göre yanlı regresyon sonuçları ile enküçük kareler tekniğinin sonuçlarının benzer olduğunu gördüler. Yanlı regresyon analizinin LS tekniğine önemli üstünlük sağlayamamasını, değişken sayısı (p) ile örnek hacmi (n) arasındaki düşük orana (p/n) bağlı olduğunu

ileri sürmüşlerdir. Faden (1978) RR'nin LS tekniğine üstünlüğünün p/n oranının yüksekliğine bağlı olduğunu göstermiştir.

Bu amaçla, yirmi kişi üzerinde toplanan verilerden yaralanarak beden ağırlığı ile açıklayıcı değişkenler arasındaki doğrusal ilişkinin tahmininde LS ve yanlı tahmin tekniklerinin göreceli geçerlilikleri karşılaştırılmaktadır. Ayrıca araştırmada, bağımsız değişkenler arasındaki güçlü çoklu doğrusal bağlantı problemine dayanarak yanlı tahmin tekniklerinin LS tekniğine göre daha düşük standart hatalı, durağan ve kuramsal (veya ampirik) beklentilere uygun tahminler sağlayacağı beklenmiştir.

2. ENKÜÇÜK KARELER TEKNİĞİ VE ÇOKLU DOĞRUSAL BAĞLANTI

2.1. Enküçük Kareler Tekniğinin Varsayımları

Regresyon katsayılarının tahmininde genelde Enküçük Kareler (LS) tekniği kullanılmaktadır.[1] LS tekniğinin varsayımlarının sağlanamaması durumunda yapılan tahminler yanlı olmakta ve böylece ilgili anlamlılık testleri geçerliliğini yitirmektedir. Bu varsayımlarla ilgili ayrıntılı tartışmalar Gujarati (1995:319-399) ve Orhunbilge (2000:15-256) kaynaklarında bulunabilir. LS ile ilgili varsayımlar aşağıdaki gibi özetlenebilir:

Hataların beklenen değeri (ortalaması) sıfırdır: $E(e) = 0$.

Hatalar birbirinden bağımsızdır. Yani, birim değerleri arasında sıra korelasyonu yoktur (absence of serial correlation): $Cov(e_i, e_j) = 0$.[2]

Hataların varyansı sabittir (farklı varyanslılığın olmaması): $Var(e_i) = \sigma^2$.

Hatalar (e_i) ile bağımlı değişken (Y) arasında korelasyon yoktur (absence of simultaneous equation bias): $Cov(e_i, Y_i) = 0$.

Hatalar ve bağımsız değişkenler birbirinden bağımsızdır: $Cov(e_i, X_i) = 0$.

Bağımsız değişkenler arasında anlamlı ilişki yoktur (absence of multicolinearity): $Cov(X_i, X_j) = 0$.

---

[1] Minimum varyans (MV) ve maksimum olabilirlik (ML) diğer kullanılan tekniklerdir (Bkz: Kleinbaum, Kupper ve Muller, 1995:49-53 ve 483-518).

[2] Cov, kovaryans kısaltmasını ifade etmektedir.

Doğrusal regresyon analizinde bağımsız değişkenler sabit (fixed, deterministic) olmasına rağmen, bağımlı değişken tesadüfidir (random).

Değişkenler hatasız ölçülmüştür (absence of measurement errors).

Bağımlı ve bağımsız değişkenler arasındaki ilişki doğrusaldır. Fakat bu doğrusallık koşulu kuşkusuz modelin parametreleri için gereklidir (değişkenler doğrusal olmayabilir).

Birim değerleri sayısı (n), değişken sayısından (p) büyük olmalıdır: n>p.

Bağımsız değişkenlerin (X_i) varyansı sıfırdan büyük olmalıdır. Değişkenin tüm gözlem değerleri (birimleri) birbirine eşitse, $X_i = \overline{X}$ olacağından, regresyon doğrusunun eğimi tanımsız olmaktadır.

Model doğru tanımlanmış olmalıdır. İlişkiye uygun fonksiyonun ve dahil edilmesi gereken tüm değişkenlerin dikkate alınması gerekmektedir.

Yukarıdaki varsayımlardan birisinin sağlanamaması durumunda LS tahmincileri yanlı (biased), tutarsız (inconsistent) veya etkisiz (inefficient) olmaktadır. Söz konusu tahminciler aşağıdaki ilk üç koşulu sağlaması durumunda en iyi doğrusal tahminciler (BLUE = Best Linear Unbiased Estimators) olarak kabul edilmektedir. $\hat{y}$, tahmin edilen değeri gösterirse;

Tahmin edilen istatistiğin beklenen değeri bilinmeyen anakütle parametresine eşitse, buna yansız (unbiased) tahmin denilmektedir:

$$E(\hat{y}) = y \tag{1}$$

Diğer tekniklerle elde edilen sonuçlarla kıyaslandığında minimum varyansa sahip tahmine etkili tahmin denilmektedir:

$$Var(\hat{y}_1) < Var(\hat{y}_2) < Var(\hat{y}_3)... \tag{2}$$

Tahmin, örnek terimlerinin doğrusal bir fonksiyonu ise bu tahmine doğrusal tahmin denilmektedir:

$$\overline{Y} = \sum Y / N = (1/N)(Y_1 + Y_2 + ...) = Y_1 / N + Y_2 / N + ... \tag{3}$$

Tahmin, örnek büyüklüğü artarken gerçek değerine yaklaşıyorsa tutarlıdır denir.

2.2. Çoklu Doğrusal Bağlantı Problemi

2.2.1. Çoklu Doğrusal Bağlantı Probleminin Sonuçları

Bağımsız değişkenler arasında güçlü ilişkilerin olmasına bağlantı (collinearity) veya çoklu doğrusal bağlantı (multicolinearity) adı verilmekte ve regresyon analizinde istenmeyen durumu göstermektedir (Orhunbilge, 2000: 240-251). Regresyon analizinde çoklu doğrusal bağlantı aşağıdaki problemlere yol açmaktadır (Gujarati, 1995:319-399; Orhunbilge, 2000:240-251):

Tam çoklu doğrusal bağlantı durumunda regresyon katsayıları belirsiz ve bu katsayıların standart hataları sonsuz olmaktadır.

Çoklu doğrusal bağlantı halinde regresyon katsayılarının varyans ve kovaryansları artmaktadır.

Modelin R^2 değeri yüksek, ancak bağımsız değişkenlerden hiçbiri veya çok azı kısmi t testine göre anlamlıdır.

İlgili bağımsız değişkenlerin bağımlı değişkenle olan ilişkilerinin yönü kuramsal ve ampirik beklentilerle çelişebilmektedir.

Bağımsız değişkenler birbiriyle bağıntılı ise, bunlardan bazılarının modelden çıkartılması gerekebilir. Fakat hangi değişkenler çıkartılacaktır? Modelden yanlış bir değişkenin çıkartılması, modelin hatalı tanımlanmasına (specification error) yol açacaktır. Diğer taraftan, bağımsız değişkenleri modele dahil edip çıkartmakta kullanabileceğimiz basit kurallar bulunmamaktadır (Myers, 1990; Gujarati, 1995).

2.2.2. Çoklu Doğrusal Bağlantının Saptanması

Çoklu doğrusal bağlantının saptanmasında kullanılan birkaç yaklaşım bulunmaktadır (Neter, vd. 1996; Gujarati, 1995). Bunlar aşağıda kısaca açıklanmaktadır.

1. Bu yaklaşımlardan birincisi, basit korelasyon matrisinin incelenmesidir. Yüzeysel olarak, iki bağımsız değişken arasındaki basit korelasyon katsayısı oldukça anlamlı (r>%75) ise, bu durum çoklu doğrusal bağlantı problemine yol açabilir. Buna rağmen, istatistik açıdan anlamlı korelasyonlar her zaman çoklu doğrusal bağlantı problemine yol açmamaktadır. Lawrence Klein'e göre basit korelasyon katsayısı (r), çoklu korelasyon katsayısından (R) küçükse, çoklu bağlantı problemi ortaya çıkmayabilir.
2. Çoklu doğrusal bağlantının saptanmasında kullanılan ikinci yaklaşım, modele yeni bağımsız değişkenler ilave edildiğinde, $R^{2'}$deki

değişimlerin incelenmesidir. R^2'de önemli bir gelişme sağlanamazsa, çoklu doğrusal bağlantı problemi ortaya çıkmış olabilir.

3. Çoklu doğrusal bağlantının saptanmasında kullanılan üçüncü yaklaşım, kısmi korelasyon katsayılarının incelenmesidir. İki değişken arasındaki basit korelasyon katsayısı anlamlı, fakat kısmi korelasyon katsayısı anlamsız ise bu durum çoklu doğrusal bağlantı problemi için bir işaret olabilir. Kısmi korelasyon yaklaşımı her zaman etkili olmamaktadır. Diğer bir anlatımla, kısmi korelasyon katsayıları yüksek olması durumunda bile çoklu doğrusal bağlantı problemi olabilmektedir.
4. Çoklu doğrusal bağlantının saptanmasında kullanılan dördüncü yaklaşım, varyans arttırıcı faktör (VIF=**V**ariance **I**nflation **F**actor) kullanılmasıdır. VIF'lerin hesaplanmasını göstermek için aşağıdaki gibi üç bağımsız değişkenli bir regresyon modelini inceleyelim.

$$y_i = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + e_i \qquad (4)$$

Üç bağımsız değişkenli bir modelin VIF değerleri aşağıdaki adımlarla hesaplanmaktadır.

- Birinci adımda X_1 bağımlı, X_2 ile X_3 bağımsız değişken olarak alınıp modelin R_1^2'si hesaplanır. Böylece X_1 için varyans artırıcı faktör $VIF(X_1) = 1/(1 - R_1^2)$ olarak hesaplanır.
- İkinci adımda X_2 bağımlı, X_1 ile X_3 bağımsız değişken olarak alınıp modelin R_2^2'si hesaplanır. Böylece X_2 için varyans artırıcı faktör $VIF(X_2) = 1/(1 - R_2^2)$ olarak hesaplanır.
- Üçüncü adımda benzer şekilde X_3 bağımlı, X_1 ile X_2 bağımsız değişken olarak alınıp modelin R_3^2'si hesaplanır. X_3 için varyans artırıcı faktör $VIF(X_3) = 1/(1 - R_3^2)$ olarak elde edilir.

Bağımlı değişken ile bağımsız değişkenler arasında ilişki yoksa (R^2 = 0 olacağından) VIF 1'e $[VIF = 1/(1 - R^2) = 1/(1 - 0) = 1]$ eşittir. Bağımlı ve bağımsız değişkenler arasında tam bir ilişki varsa (R^2 = 1 olacağından) VIF $[VIF = 1/(1 - R^2) = 1/(1 - 1) = \infty]$ sonsuz olacaktır. R^2 %90 ise, VIF 10 $[VIF = 1/(1 - 0{,}9) = 10]$ olarak elde edilir. Webster (1992: 683-684) yorum için şu genel kuralı önermektedir: VIF 10'a eşit veya daha büyük $(VIF \geq 10)$ ise, anlamlı çoklu doğrusal bağlantı problemi söz konusudur.

5. Çoklu doğrusal bağlantının saptanmasında kullanılan beşinci yaklaşım, bağımsız değişkenler için tolerans değerinin (TV=**T**olerance **V**alue) hesaplanmasıdır. Tolerans değeri, 1'den belirlilik katsayısının çıkartılmasıyla $(TV = 1 - R^2)$ hesaplanmaktadır. Böylece daha küçük tolerans, daha büyük VIF değeri demektir.

6. Çoklu doğrusal bağlantının saptanmasında kullanılan altıncı yaklaşım yardımcı regresyon eşitliklerinden (auxiliary regression equations) yararlanarak F değerlerinin hesaplanmasıdır. Bunun için üç bağımsız değişkenli aşağıdaki regresyon modelini inceleyelim.

$$y_i = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + e_i \qquad (5)$$

Her bir bağımsız değişken ile geriye kalan diğer bağımsız değişkenler arasındaki ilişki araştırılarak $R_{1.23}$, $R_{2.13}$ ve $R_{3.12}$ çoklu korelasyon katsayıları hesaplanır. Daha sonra bu katsayılardan yararlanarak her bir bağımsız değişken için F değeri hesaplanmaktadır. Örneğin F değeri X_1 değişkeni için aşağıdaki gibi hesaplanmaktadır:

$$F_{1.23} = \frac{R^2_{1.23} / (k-2)}{(1 - R^2_{1.23}) / (n-k+1)} \qquad (6)$$

Formülde n birim sayısını; k ise bağımsız değişken sayısını göstermektedir. Hesaplanan F değeri ($df_1 = k-2$ ve $df_2 = n-k+1$), kritik F değeriyle karşılaştırılmaktadır. $F_{1.23}$, kritik F değerinden büyükse X_1 değişkeni ile X_2 ve X_3 değişkenleri arasındaki ilişkinin anlamlı olduğuna karar verilmektedir.

7. Çoklu doğrusal bağlantının saptanmasında kullanılan yedinci yaklaşım, koşul sayısı (CN = Condition Number) veya koşul endeksi (CI = Condition Index)) değerlerinin hesaplanmasıdır. CN ve CI aşağıdaki gibi hesaplanmaktadır:

   CN = [Maksimum Özdeğer / Minimum Özdeğer]
   CI=[Maksimum Özdeğer / Minimum Özdeğer]$^{1/2}$

CI değeri 10 ile 30 (CN değeri 100 ile 1000) arasında ise orta, 30'dan (veya 1000'den) büyükse çok güçlü çoklu doğrusal bağlantı problemini gösterir (Gujarati, 1995:338).

Çoklu doğrusal bağlantı probleminin saptanmasında kullanılan ve yukarıda açıklanan yaklaşımlardan her biri belirli dezavantajlara sahiptir.

Ayrıca, hangi durumda hangi yaklaşımın kullanılabileceği konusunda da bir öneride bulunulamamaktadır (Gujarati, 1995:339).[3]

2.2.3. Çoklu Doğrusal Bağlantı Probleminin Çözümü

Çoklu doğrusal bağlantı probleminin çözüm yolları aşağıda verilmektedir (Gujarati, 1995:339-344; Netter-Wasserman-Kunter, 1990:411-412; Orhunbilge, 2000:240-251).

1. Bir veya daha çok bağımsız değişken modelden çıkartılabilir. Fakat hangi değişkenler çıkartılacaktır? Böyle bir yaklaşım, modeli yanlış tanımlamaya götürebilir.
2. Farklar alınarak değişkenler dönüştürülebilir. Fakat böyle bir dönüşüm hatalar arasında otokorelasyon problemine yol açabilir. Ayrıca, böyle bir dönüşüm sadece zaman serilerine uygulanabilirken, kesit verilerine uygulanamamaktadır.
3. Bazen yeni gözlem değerlerinin elde edilmesiyle çoklu doğrusal bağlantı problemi ortadan kaldırılabilir. Fakat, her zaman örneği büyütmek mümkün olmamaktadır.
4. Birbiriyle ilişkili olan iki değişken yerine bu iki değişkenin toplamı (tek bir değişken olarak) alınabilir.
5. Birbirinden bağımsız bileşenler türeten "Temel Bileşenler Regresyon Analizi" veya LS tekniğinin düzeltilmiş şekli olan ve yanlı standartlaştırılmış regresyon katsayılarını tahmin eden "Ridge Regresyon" teknikleri kullanılabilmektedir.[4]

2.3. Ridge Regresyon (RR) Analizi

Burada RR tekniği ana hatlarıyla özetlenmektedir. Tekniğin teorik açıklaması Orhunbilge (2000:240-241) tarafından yapılmaktadır. LS

---

[3] Bilindiği gibi çoklu doğrusal bağlantının saptanmasında en yaygın kullanılan yaklaşım bağımsız değişkenler için VIF değerlerinin hesaplanmasıdır. Fakat bu yaklaşımla da çoklu doğrusal bağlantı kesin olarak saptanabildiği iddia edilememektedir. Çünkü, bilindiği gibi herhangi bir kısmi regresyon katsayısının varyansı üç faktöre bağlıdır: σ^2, $\sum x_i^2$ ve VIF_i. Yani, $Var(\hat{b}_i) = (\sigma^2 / \sum x_i^2) \times (1/(1 - R_i^2)) = (\sigma^2 / \sum x_i^2) \times VIF_i$ dir. Bu eşitlikten görüldüğü gibi yüksek bir VIF değeri düşük bir σ^2 veya yüksek bir $\sum x_i^2$ değerleriyle de elde edilmiş olabilir. Diğer bir anlatımla yüksek bir VIF değeri, her zaman yüksek varyansların veya standart hataların elde edilmesi için gerekli ve yeterli bir neden değildir (Bkz: Gujarati:338-339).

[4] SPSS 11.5 istatistik programında "Ridge Regression" için bir makro dosyası bulunmaktadır.

tekniği ile yapılan tahminler bu tekniğin varsayımlarını sağlanması durumunda yansız olmaktadır. Çoklu doğrusal bağlantı halinde ise, regresyon katsayılarının varyans ve kovaryansları artmaktadır. Diğer bir anlatımla, önemli değişkenlere ait regresyon katsayılarının standart hataları büyür ve bu değişkenlerin regresyon katsayılarının kısmi t testleri anlamsız sonuç verir. Çoklu doğrusal bağlantı halinde herhangi bir bağımsız değişken veya birime ait veriler modelden çıkartıldığında veya modele sokulduğunda kısmi regresyon katsayılarında çok önemli değişiklikler olmaktadır. Ayrıca çoklu doğrusal bağlantı halinde kısmi regresyon katsayılarının işaretleri teoriden veya beklenenden farklı olabilmektedir. Kısaca, çoklu doğrusal bağlantılı verilerle hesaplanan standartlaştırılmış regresyon katsayıları durağanlığını veya kararlılığını kaybetmektedir (Darlington, 1978; Faden, 1978).[5] RR tekniği, bu tahminlere küçük bir yanlılık sabiti ekleyerek varyansı azaltmaya yardım etmektedir (Darlington, 1978; Dempster, Schatzoff & Wermuth, 1997; Hoerl & Kennard, 1970; Price, 1977). Genelde, varyans-kovaryans matrisinin köşegen değerlerine küçük bir yanlılık sabiti (k) ilave etmenin dışında, RR ile LS tekniklerinin işleyişi aynıdır. Diğer bir anlatımla RR ile bir taraftan tahminlerin varyansı azaltılmakta, diğer taraftan ise bu katsayı (k) oranında yanlı tahminler elde edilmektedir. Böylece iki alternatif söz konusu olmaktadır: Yansız tahminlerle yüksek varyans veya yanlı tahminlerle düşük varyans.

Tahmin edilecek LS regresyon modelinin aşağıdaki gibi matris notasyonuyla gösterildiğini varsayalım (NCSS Inc, 2001):

$$\underline{\mathbf{Y}} = \mathbf{X}\underline{\mathbf{B}} + \underline{\mathbf{e}} \qquad (7)$$

Burada $\underline{\mathbf{Y}}$, bağımlı değişkeni; $\mathbf{X}$, bağımsız değişkenleri; $\underline{\mathbf{B}}$, tahmin edilecek regresyon katsayılarını ve $\underline{\mathbf{e}}$, hata terimini göstermektedir.

Ridge regresyon analizinde ilk olarak bağımlı ve bağımsız değişkenler ortalamalarından farkları alınıp standart sapmalarına

[5] Regresyon analizi standartlaştırılmış veriler üzerine uygulanması durumunda hesaplanan katsayılara standartlaştırılmış regresyon katsayıları denilmektedir (Bkz: Norusis and SPSS Inc., 1993:314). Standartlaştırılmamış katsayılara dayanarak değişkenlerin göreceli önemlerini karşılaştırmak uygun olmadığından, analizde kullanılan değişkenler kendi ortalamalarından farkları alınıp standart sapmalarına bölünerek standartlaştırılmaktadır. Diğer bir anlatımla değişkenler ölçü birimlerinden arındırılarak varyansları bire eşitlenmektedir. Ancak standartlaştırılmış katsayılar, doğrudan standartlaştırılmamış katsayılardan yararlanarak aşağıdaki eşitlik yardımıyla da hesaplanabilmektedir: $\text{Beta}_k = b_k (s_{x_k} / s_y)$.

bölünerek standartlaştırılmaktadır. Nihai regresyon katsayıları elde edildiğinde ise, katsayılar orijinal ölçü birimlerine dönüştürülmektedir.

LS tekniğiyle regresyon katsayıları $(\hat{\underline{\mathbf{B}}})$ aşağıdaki gibi hesaplanmaktadır:

$$\hat{\underline{\mathbf{B}}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\underline{\mathbf{Y}} \tag{8}$$

Değişkenler standartlaştırıldığından ve **R** bağımsız değişkenler arasındaki korelasyon matrisini gösterdiğinden $\mathbf{X}'\mathbf{X} = \mathbf{R}$ dir. Bu tahminlerin beklenen değerleri anakütle değerlerine eşit olacağından bu tahminler yansızdır. Yani;

$$\mathbf{E}(\hat{\underline{\mathbf{B}}}) = \underline{\mathbf{B}} \tag{9}$$

Tahminlerin varyans-kovaryans matrisi ise,

$V(\hat{\underline{\mathbf{B}}}) = \sigma^2\mathbf{R}^{-1}$ dir ve y'ler standartlaştırılmış olduğundan, $\sigma^2 = 1$ dir. Buradan;

$$V(\hat{b}_j) = r^{jj} = \frac{1}{1 - R_j^2} = \text{VIF}$$ eşitliği yazılabilir.

R_j^2; X_j değişkeninin bağımlı, analizdeki diğer bağımsız değişkenlerin ise bağımsız değişkenler olarak alınan modelin kareli korelasyon (belirlilik) katsayısını göstermektedir. Böylece bu varyans VIF değerine eşit olmaktadır. Görüldüğü gibi eşitliğin paydasında yer alan R^2, 1'e yaklaştıkça varyans (ve bu yüzden VIF) değeri büyümektedir. Burada VIF için kritik değer (cut-off value), daha önce belirtildiği gibi, 10'dur. Diğer taraftan R^2 için kritik değer 0,90 olmaktadır. Bu nedenle analizdeki bağımsız değişkenlerden herhangi birisi bağımlı değişken olarak alınıp geriye kalan diğer değişkenler arasındaki R^2 değeri 0,90 veya daha büyük ise, çoklu doğrusal bağlantı sorunu söz konusu olmaktadır.

Ridge regresyon analizinde korelasyon matrisinin köşegen değerlerine küçük bir yanlılık sabiti eklenerek, yanlı standartlaştırılmış regresyon katsayıları aşağıdaki gibi hesaplanmaktadır:

$$\tilde{\underline{\mathbf{B}}} = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{Y} \text{ veya } \tilde{\underline{\mathbf{B}}} = (\mathbf{R} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{Y} \tag{10}$$

Burada k, 1'den küçük pozitif sayısal bir değerdir (genelde $k \leq 0{,}3$ tür). Bu tahminin yanlılık büyüklüğü beklenen değeri aşağıdaki gibidir.

$$\mathbf{E}(\underline{\widetilde{\mathbf{B}}} - \underline{\mathbf{B}}) = [(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{X} - \mathbf{I}]\underline{\mathbf{B}} \qquad (11)$$

ve kovaryans matrisi aşağıdaki eşitlikle elde edilmektedir:

$$\mathrm{V}(\underline{\widetilde{\mathbf{B}}}) = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1} \qquad (12)$$

Buradan, belirli bir k değeri için, yanlı tahminlere ait ortalama hata karelerinin beklenen değerinin enküçük kareler tekniğiyle elde edilen tahminlerden daha küçük olduğu gösterilebilmektedir. Fakat optimum k sabiti gerçek regresyon katsayılarının (tahmin edilecek) bilinmesine bağlıdır. Optimum çözümü garanti edebilecek bir yaklaşım henüz bulunamamıştır. Burada k, 0 ile 1 aralığında değerler alabilmektedir. k, 1'e yaklaştıkça tahminlerin yanlılığı artmakta, fakat varyansları azalmaktadır.

Optimum k sabitini araştırmak için yanlı standartlaştırılmış regresyon katsayıları ile k arasında hesaplanan ve yanlı regresyon grafiği (Ridge Trace) adı verilen grafiklerden yaralanılmaktadır (Hoerl vd., 1970:69-82). Bu grafiklerde yanlı regresyon katsayıları k'nın bir fonksiyonu olarak gösterilmektedir. Optimum k değeri, yanlı standartlaştırılmış regresyon katsayılarının durağanlaştığı bölgeden seçilmektedir. Genelde standartlaştırılmış regresyon katsayıları ilkönce küçük k değerleriyle çok anormal bir biçimde değişmekte ve daha sonra durağanlaşmaktadır. Regresyon katsayılarının durağanlaştığı bu bölgede olası enküçük k değeri optimum k değeri olarak seçilmektedir. Optimum k değerinin seçiminde kullanılan diğer kriterler arasında katsayıların kuramsal beklentilere uygunluğunu, durağanlığını, makul büyüklüğünü, kabul edilebilir hata kareleri toplamını ve minimum VIF'leri (bağımsız değişkenler için birlikte 1'e yaklaşan VIF değerlerini) sağlayan k sabiti yaklaşımları sayılabilir (Anderson, 1998). Uygulamada tüm kriterlerin birlikte değerlendirilmesi en sağlıklı yaklaşım olarak kabul edilmektedir.

2.4. Temel Bileşenler Regresyon (PC) Analizi

Matris notasyonuyla verilen aşağıdaki regresyon modelini dikkate alalım:

$$\underline{\mathbf{Y}} = \mathbf{X}\underline{\mathbf{B}} + \underline{\mathbf{e}} \qquad (13)$$

$\underline{\mathbf{Y}}$**,** bağımlı değişkeni; **X** bağımsız değişkenleri; $\underline{\mathbf{B}}$**,** regresyon katsayılarını ve $\underline{\mathbf{e}}$**,** hata terimini göstermektedir.

Temel bileşenler (PC) analizinde ilk olarak hem bağımlı hem de bağımsız değişkenler ortalamadan farkları alınıp standart sapmalarına bölünerek standartlaştırılmaktadır (NCSS Inc., 2001). Nihai regresyon katsayıları gösterildiğinde ise, değişkenler orijinal ölçeğine dönüştürülmektedir. Daha önce belirtildiği gibi LS tekniğiyle regresyon katsayıları aşağıdaki eşitlikle hesaplanmaktadır:

$$\hat{\underline{\mathbf{B}}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\underline{\mathbf{Y}} \qquad (14)$$

Değişkenler standartlaştırıldığından $\mathbf{X}'\mathbf{X} = \mathbf{R}$ dir. Burada **R**, bağımsız değişkenler için korelasyon matrisini göstermektedir. PC regresyon analizini gerçekleştirmek için bağımsız değişkenler temel bileşenlere dönüştürülür (NCSS Inc., 2001). Bu matematik olarak aşağıdaki gibi yazılmaktadır:

$$\mathbf{X}'\mathbf{X} = \mathbf{P}\mathbf{D}\mathbf{P}' = \mathbf{Z}'\mathbf{Z} \qquad (15)$$

Eşitlikte PC modelini tanımlayan **D**, $\mathbf{X}'\mathbf{X}$ özdeğerlerinin köşegen matrisini; **P**, $\mathbf{X}'\mathbf{X}$ özvektör matrisini ve **Z**, veri matrisini (**X** yapısına benzer) göstermektedir. Temel bileşenler (**P**) ortogonal olduğundan $\mathbf{P}'\mathbf{P} = \mathbf{I}$ dır.

Böylece orijinal değişkenlerin (**X**) ağırlıklı ortalamalarını ifade eden yeni değişkenler (**Z**) türetilmektedir. Bu durum, regresyon analizine başlamadan önce değişkenlerin logaritmik veya karekök dönüşümlerinin alınmasından farklı bir şey değildir. Bu yeni değişkenler temel bileşen olduğu için, bu bileşenler arasındaki korelasyonlar sıfırdır. X1, X2, X3 gibi değişkenlerle analize başlanmışsa, Z1, Z2 ve Z3 gibi dönüştürülmüş değişkenler elde edilmektedir.

Çok küçük özdeğerler hesaplanması durumunda çok güçlü çoklu doğrusal bağlantı (severe multicolinearity) sorunu söz konusu olmaktadır. Bu sorunun üstesinden gelebilmek için düşük özdeğerlerle özdeşleşen bileşenler analizden çıkartılmaktadır. Genelde göreceli olarak özdeğeri küçük bir veya iki (en çok) temel bileşen elde edilmektedir. Örneğin üç değişkenli bir modelde özdeğeri küçük sadece bir temel bileşen elde edilmişse, Z3 (üçüncü temel bileşen) analizden çıkartılmaktadır (NCSS Inc., 2001).

Böylece **Y** değişkeni bağımlı ve Z1 ve Z2 bileşenleri ise bağımsız değişkenler olarak alınan modelde artık çoklu doğrusal bağlantı söz konusu olmamaktadır. Daha sonra sonuçlar **X** ölçeğine geri

dönüştürülerek **B**'nin tahminleri elde edilir. Bu tahminler yanlı olacaktır. Fakat bu yanlılığın büyüklüğü varyansın azaltılmasıyla dengeleneceği umulmaktadır. Diğer bir anlatımla PC tahminlerinin hata kareleri ortalamasının LS tahminlerinden daha küçük olması beklenmektedir. Matematik olarak, asal bileşenlerin özel doğası gereği, regresyon katsayılarının tahmini aşağıdaki gibidir (NCSS Inc., 2001):

$$\underline{\hat{\mathbf{A}}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\underline{\mathbf{Y}} = \mathbf{D}^{-1}\mathbf{Z}'\underline{\mathbf{Y}} \qquad (16)$$

Artık bu eşitlik, değişik bağımsız değişken seti üzerine uygulanan LS regresyonudur. Böylece $\hat{\mathbf{A}}$ ve $\hat{\mathbf{B}}$ gibi iki regresyon katsayıları seti arasındaki ilişkiler ise aşağıdaki gibi yazılmaktadır:

$$\underline{\hat{\mathbf{A}}} = \mathbf{P}'\underline{\hat{\mathbf{B}}} \qquad \text{ve} \qquad \underline{\hat{\mathbf{B}}} = \mathbf{P}\underline{\hat{\mathbf{A}}} \qquad (17)$$

$\underline{\hat{\mathbf{A}}}$'nın ilgili elementi sıfıra eşitlenerek ilgili temel bileşen analizden çıkartılabilir. Böylece PC regresyonu aşağıdaki adımlarla özetlenebilir (NCSS Inc., 2001):

X matrisi için PC analizi uygulanarak temel bileşenler (**Z**) elde edilir.

$\hat{\mathbf{A}}$ katsayılar matrisinin enküçük kareler tahminlerini elde etmek için **Y** ile **Z** arasında LS regresyonu uygulanır.

$\hat{\mathbf{A}}$'nın en son elementi sıfıra eşitlenir.

Son aşamada $\underline{\hat{\mathbf{B}}} = \mathbf{P}\underline{\hat{\mathbf{A}}}$ eşitliği kullanılarak hesaplanan katsayılar orijinal ölçeğine dönüştürülmektedir.

Son olarak, RR tekniğinde k yanlılık sabitinin seçiminde yaşanan belirsizliğin aksine, PC analizinde elimine edilecek temel bileşenlerin sayısı göreceli olarak daha kesindir. Diğer bir anlatımla çoklu doğrusal bağlantı sorununu çözmek için özdeğeri göreceli olarak küçük olan bileşenlerin analizden çıkartılması gerektiği hususu daha kesindir.[6] Görüldüğü gibi PC tekniğinde, RR tekniğindeki k yanlılık sabitinin seçimi gibi bir belirsizlik söz konusu değildir.

---

[6] Bir korelasyon matrisinin tipik özdeğerinin büyüklüğü 1 olduğundan, özdeğeri 1'den çok küçük olan bileşenler analizden çıkartılmaktadır ve genelde verilecek karar açık ve nettir. Bilindiği gibi standartlaştırılmış değişkenler için bir korelasyon matrisinin özdeğeri bağımsız değişken sayısının toplamına eşittir. Örneğin, üç bağımsız değişkenli bir korelasyon matrisinin özdeğerleri toplamı 3'tür. Korelasyon matrisinin tipik özdeğeri ise bu korelasyon matrisinin özdeğerleri toplamının ortalamasına (yani 3/3=1) eşittir.

3. UYGULAMA: BEDEN AĞIRLIĞI VERİLERİ İÇİN LS, RR VE PC TEKNİKLERİYLE ELDE EDİLEN SONUÇLARIN KARŞILAŞTIRILMASI

Araştırmada kullanılan veriler, İstanbul Üniversitesi İstanbul Tıp Fakültesi Hastanesine şişmanlık şikayeti ile başvuran hastalar arasından tesadüfi olarak elde edilmiştir. Diğer bir anlatımla araştırmanın anakütlesi İstanbul Tıp Fakültesine şişmanlık şikayeti ile başvuran hastalar, kullanılan örnekleme tekniği basit tesadüfi bir örneklemedir. Araştırmada dört değişken kullanılmaktadır: Beden ağırlığı (kg), deri alanı (cm^2), uyluk kemiğinin çevresinin uzunluğu (cm) ve belden yukarı ölçülen kasların çevrelerinin uzunluğu (cm). Araştırmada 20 kişiden elde edilen ölçüm sonuçları (veriler tablosu) Tablo 1'de verilmektedir.

Araştırmada 20 kişi üzerinde ölçülen veriler ve LS ve RR teknikleri için yazılan makro komutları Tablo 1'de verilmektedir. LS, RR ve PC teknikleriyle elde edilen sonuçlar ise Tablo 2'de özetlenmektedir.[7] LS analizine göre beden ağırlığı ile açıklayıcı değişkenler arasındaki doğrusal ilişkinin %89,5 olduğu ve beden ağırlığında meydana gelen değişimlerin yaklaşık %80,1'i bağımsız değişkenler tarafından açıklandığı anlaşılmaktadır [Tablo 2: 1.1].

Ayrıca LS tekniğiyle elde edilen ANOVA tablosu incelendiğinde %5 veya %1 geleneksel anlamlılık düzeylerinde modelin anlamlı (güvenilir) olduğu (F=21,488; p=,000) görülmektedir [Tablo 2: 1.2]. Modelin yüksek belirlilik (R^2=%80,1) ve korelasyon katsayısının (R=%89,5) aksine katsayılar tablosu incelendiğinde hiçbir değişkenin kısmi t testine göre anlamlı olmadığı görülmektedir [Tablo 2: 1.3]. Bu durum daha önce belirtildiği gibi, çoklu doğrusal bağlantı probleminin bir göstergesidir.

Aynı sonuca çoklu doğrusal bağlantı ile ilgili istatistiklerin incelenmesiyle de varılmaktadır [Tablo 2: 1.4]. Burada bağımsız değişkenler için hesaplanan korelasyon matrisi, VIF ve TV değerleri, $R_{1.23}$, $R_{2.13}$, $R_{3.12}$ ile korelasyonların özdeğerleri için hesaplanan CN (koşul sayısı) ve CI (koşul endeksi) değerleri verilmektedir (Tablo 2: 1.4). Korelasyon matrisi incelendiğinde X1 ve X2 değişkenleri arasında çok yüksek bir ilişki (%92,4) görülmektedir. X1, X2 ve X3 bağımsız değişkenleri için hesaplanan VIF'ler sırasıyla yaklaşık olarak 755, 601 ve 111 olarak elde edilmiş ve bu değerler kritik VIF değerinin (10) oldukça üstündedir.

[7] LS ve RR sonuçları SPSS 11.5 istatistik paket programında makro komutlar yazılarak elde edilmiştir.

Bağımsız değişkenlerden her biri bağımlı değişken olarak alınıp geriye kalan diğer bağımsız değişkenler arasındaki ilişkiler incelendiğinde bu ilişkilerin 0,90'dan büyük olduğu görülmektedir. Hesaplanan korelasyonlara ait özdeğerlerden yararlanarak elde edilen CN (koşul sayısı = 3028,97) ve CI (koşul endeksi = 55,036) değerleri kritik değerlerden (1000 ve 30) büyüktür [Tablo 2: 1.4]. Bu istatistiklerin tamamı verilerde çok güçlü çoklu doğrusal bağlantı problemi olduğunun bir kanıtıdır.

Ridge regresyon analizi sonuçlarının birinci kısmında belirli yanlılık katsayılarına göre R^2, standartlaştırılmış ridge regresyon katsayıları ile bu katsayılara ait VIF değerlerindeki değişimler verilmektedir [Tablo 2: 3.1]. Bu verilerden yararlanarak optimum k sabiti standartlaştırılmış ridge regresyon katsayılarının durağanlaşmaya başladığı ve bu katsayılara ait VIF değerlerinin birlikte 1'e yaklaştığı noktada (k=%2) seçilir.

Tablo 1: LS ve RR Analizleri İçin SPSS Makro Komutları[8]

```
Title "LS ve RR Analizi".
Data List Free /N Y X1 X2 X3.
Variable Labels
            Y          "Beden Ağırlığı, Kg"
            /X1        "Deri Alanı, Metre Kare"
            /X2        "Uyluk Kemiğinin Çevresinin Uzunluğu, Cm"
            /X3        "Ölçülen Kasların Çevrelerinin Uzunluğu, Cm".
Begin Data.
1          75,57     1,81     109,47    73,91
2          144,78    2,30     126,49    71,63
3          118,74    2,86     131,83    93,98
4          127,64    2,77     137,92    78,99
5          81,91     1,78     107,19    78,49
6          137,79    2,38     136,91    60,20
7          172,09    2,92     148,59    70,10
8          161,29    2,59     132,33    77,72
9          135,26    2,06     126,75    58,93
10         122,56    2,37     135,89    62,99
11         161,29    2,89     143,76    76,20
12         172,72    2,83     144,02    71,88
13         74,29     1,74     118,11    58,42
14         113,03    1,83     112,27    72,64
15         81,28     1,36     108,46    54,10
16         151,76    2,74     138,18    76,45
17         143,51    2,58     140,46    65,28
18         161,29    2,81     148,84    62,48
19         93,98     2,11     122,43    68,83
20         133,99    2,34     129,54    69,85
End Data.
Subtitle "(1) LS Analizi".
Regress Variables = Y1 X1 X2 X3
            /Statistics Coeff  Outs CI R ANOVA Collin Tol
            /Dependent = Y
            /Method = Enter X1 X2 X3.
Subtitle "(2) Korelasyon Analizi".
Correlation Variables = X1 X2 X3.
Subtitle "(3) RR Analizi".
Include File "C:\Program Files\SPSS\Ridge Regression.sps".
Ridgereg Dep = Y
            /Enter X1 X2 X3.
*Rapor ve ilgili grafiklerin elde edilebilmesi için komutlar bu şekilde yazılmalıdır.
Subtitle "(4) RR Analizi".
Ridgereg Dep = Y
            /Enter X1 X2 X3
            /k = 0.02.
Subtitle "(5) RR Analizi".
Ridgereg Dep = Y
            /Enter X1 X2 X3
            /Start = 0
            /Stop  = 1
            /Inc   = 0.05.
```

* Not: k, 0-1 aralığında tanımlı yanlılık sabitidir ve regresyon çıktısı için gereklidir.

[8] Sonuçlar SPSS 11.5 istatistik programıyla elde edilmiştir.

Tablo 2: LS, RR ve PC Regresyon Teknikleriyle Elde Edilen Sonuçlar

(1) LS ve Çoklu Doğrusal Bağlantı

(1.1) Model Özeti (LS)

R	R-Kare	Düzeltilmiş-R Kare	Tahminin Standart Hatası	D-Watson (d)
,895	,801	,764	15,756	2,202

(1.2) ANOVA (LS)

Kaynak	SS	DF	MS	F	p
Regresyon	16003,875	3	5334,625	21,488	,000
Hata	3972,212	16	248,263		
Toplam	19976,087	19			

(1.3) Tahmin Edilen Regresyon Katsayıları ve Güven Aralıkları

P	$\hat{b}$	SH	$\hat{\beta}$	t	p	b İçin %95 Güven Aralığı	
						Alt	Üst
Sabit	768,420	653,842		1,175	,257	-617,663	2154,503
X1	304,366	212,716	4,383	1,431	,172	-146,571	755,303
X2	-7,404	6,667	-3,036	-1,111	,283	-21,538	6,729
X3	-5,619	4,113	-1,605	-1,366	,191	-14,338	3,101

(1.4) LS Çoklu Doğrusal Bağlantının Saptanması

p	Korelasyon Matrisi			VIF	TV	R^2 ve Diğer X'ler
	X1	X2	X3			
X1	1			754,923	,001	$R^2_{1.23}$ = 0,999
X2	,924	1		601,325	,002	$R^2_{2.13}$ = 0,999
X3	,475	,085	1	111,104	,009	$R^2_{3.12}$ = 0,991

VIF değerleri 10'dan büyük olduğundan verilerde çoklu doğrusal bağlantı söz konusudur.

Korelasyonların Özdeğerleri

No	Özdeğer	Göreceli Yüzde	Birikimli Yüzde	Koşul Sayısı	Koşul Endeksi
1	2,066	68,88	68,88	1,00	1
2	0,933	31,10	99,98	2,22	1,490
3	0,001	0,02	100,00	3028,97	55,036

Bazı endeks sayıları 1000'den (veya koşullu endeks değerleri 30'dan) büyük olduğu için verilerde çok önemli çoklu doğrusal bağlantı problemi söz konusudur.

(2) k=%2 Yanlılık Sabitiyle Ridge Regresyon Raporu

Çoklu R	,884
R-Kare	,782
Düzeltilmiş R-Kare	,741
Standart Hata	16,515

ANOVA Tablosu

Kaynak	DF	SS	MS
Regresyon	3	15612,213	5204,071
Hata	16	4363,874	272,742
F Değeri	19,081		
F Anl. (p)	,000		

Tablo 2: (Devam)

Modeldeki Değişkenler

Değişken	$\hat{b}$	SH	$\hat{\beta}$	$\hat{b}$ / SH($\hat{b}$)	VIF
Sabit	-47,910	40,329	,000	-1,188	
X1	37,661	8,326	,542	4,523	1,053
X2	,929	,291	,381	3,193	1,041
X3	-,473	,410	-,135	-1,155	1,003

(3) Ridge Regresyon (RR) Analizi

(3.1) k, R-Kare, Standartlaştırılmış Ridge Regresyon Katsayılar ve VIF'ler

		Standartlaştırılmış Ridge Regresyon Katsayıları			Varyans Arttırıcı Faktör (VIF)		
k	R^2	X1	X2	X3	X1	X2	X3
0,000	0,801	4,3828	-3,0360	-1,6051	754,9231	601,3251	111,1043
0,001	0,788	2,0243	-0,9320	-0,7038	124,3590	99,2328	19,0402
0,004	0,782	0,9928	-0,0131	-0,3093	16,2573	13,1545	3,2518
0,005	0,781	0,8909	0,0773	-0,2702	11,1139	9,0587	2,4991
0,006	0,779	0,8194	0,1406	-0,2428	8,1013	6,6594	2,0574
0,007	0,779	0,7665	0,1874	-0,2224	6,1864	5,1342	1,7760
0,008	0,779	0,7257	0,2232	-0,2067	4,8941	4,1047	1,5855
0,009	0,778	0,6933	0,2516	-0,1942	3,9811	3,3773	1,4503
0,010	0,777	0,6669	0,2747	-0,1840	3,3121	2,8442	1,3509
0,020	0,772	0,5423	0,3808	-0,1351	1,0528	1,0410	1,0032
0,030	0,768	0,4976	0,4157	-0,1169	0,6026	0,6783	0,9202
0,040	0,764	0,4739	0,4319	-0,1066	0,4395	0,5445	0,8796
0,050	0,759	0,4588	0,4406	-0,0997	0,3618	0,4788	0,8520
0,080	0,747	0,4330	0,4494	-0,0867	0,2727	0,3969	0,7931
0,090	0,743	0,4274	0,4498	-0,0836	0,2594	0,3828	0,7767
0,100	0,739	0,4225	0,4496	-0,0808	0,2493	0,3715	0,7612
0,200	0,703	0,3910	0,4350	-0,0610	0,2047	0,3072	0,6342
0,300	0,670	0,3700	0,4155	-0,0477	0,1835	0,2682	0,5386
0,400	0,640	0,3527	0,3967	-0,0375	0,1674	0,2381	0,4635
0,500	0,613	0,3376	0,3791	-0,0294	0,1540	0,2135	0,4034
0,800	0,545	0,3000	0,3344	-0,0128	0,1227	0,1603	0,2802
1,000	0,507	0,2797	0,3101	-0,0059	0,1070	0,1357	0,2274

(4) Temel Bileşenler (PC) Regresyon Raporu[9]

(4.1) Regresyon Katsayıları Raporu

Değişken	$\hat{b}$	SH	$\hat{\beta}$	VIF
Sabit	-77,520			
X1	28,863	3,955	0,416	0,236
X2	1,229	,181	0,504	0,399
X3	-0,313	,394	-0,089	0,920

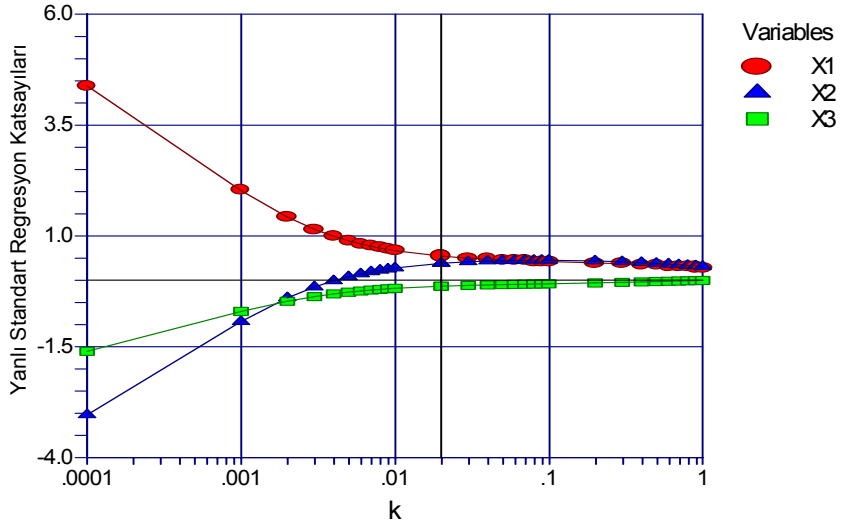
(4.2) Varyans (ANOVA) Analizi

Kaynak	DF	SS	MS	F	p
Sabit	1	328902,30	328902,3	18,942	0,000
Model	3	15587,29	5195,764		
Hata	16	4388,79	274,299		
Toplam (Düzeltilmiş)	19	19976,09	1051,373		
Standart Hata (SH)		16,562			
R-Kare		0,781			

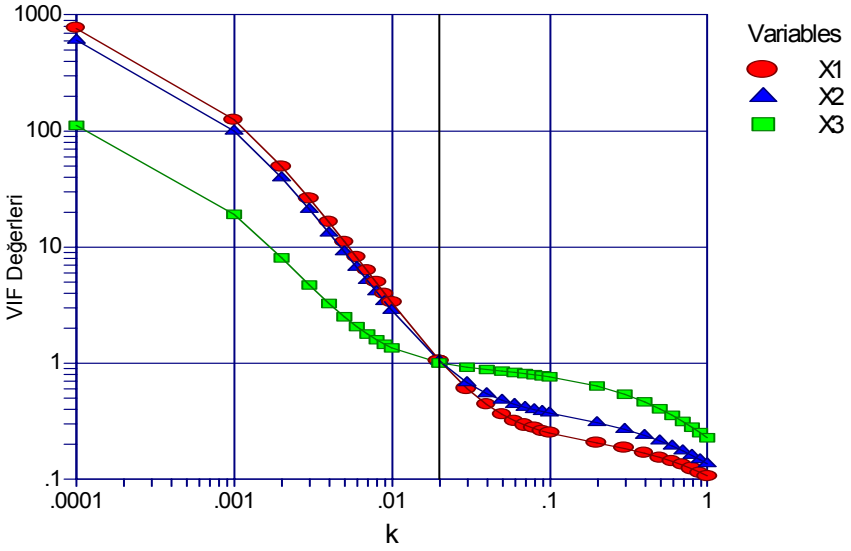
[9] Temel Bileşenler Regresyon analizinin sonuçları NCSS 2004 istatistik paket programıyla elde edilmiştir.

Tablo 2: (Devam)

(3.2) Logaritmik Ölçekli Ridge Regresyon Grafikleri



(a) Ridge Grafiği



(b) VIF Grafiği

Yanlılık sabiti (k) ile yanlı standartlaştırılmış regresyon katsayıları arasındaki grafiğe bakıldığında çok küçük (k=0,02) bir yanlılık sabitinden sonra regresyon katsayılarının durağan hale geldiği görülmektedir [Tablo 2: Grafik 3.2-a]. Yanlılık sabiti ile modelin R^2 değerleri arasındaki ilişki incelendiğinde k, 0 ile 1 aralığında bir değişim gösterirken; R^2, %68,2 ile %80,1 aralığında değişmektedir [Tablo 2: 3.1].

Logaritmik ölçekli grafikte standartlaştırılmış ridge regresyon katsayılarının nasıl durağanlaştığı çok açık bir biçimde görülmektedir [Tablo 2: Grafik 3.2-a]. Ayrıca değişkenlerin VIF değerleri hangi yanlılık sabiti ($k \cong 0{,}02$) ile 1'e birlikte yaklaştıkları bir sonraki grafikte görülmektedir [Tablo 2: Grafik 3.2-b].

Tablo 3: LS, RR ve PC Teknikleriyle Elde Edilen Sonuçların Karşılaştırılması

	$\hat{b}_i$			$\hat{\beta}_i$			SH			VIF		
P	LS	RR	PC	LS	RR	PC	LS	RR	PC	LS	RR	PC
Sabit	768,42	-47,91	-77,52									
X1	304,37	37,66	28,86	4,38	0,54	0,42	212,72	8,49	3,96	754,92	1,05	0,24
X2	-7,40	0,93	1,23	-3,04	0,38	0,50	6,68	0,29	0,18	601,33	1,04	0,40
X3	-5,62	-0,47	-0,31	-1,61	-0,14	-0,09	4,11	0,41	0,39	111,10	1,00	0,92

$$LS \mapsto y_i' = 768{,}42 + 304{,}37\,x_1 - 7{,}404\,x_2 - 5{,}620\,x_3 \mapsto R^2 = 0{,}801 \mapsto F = 21{,}448 \mapsto SH = 15{,}756$$

$$RR \mapsto y_i' = -47{,}91 + 37{,}661\,x_1 + 0{,}929\,x_2 - 0{,}473\,x_3 \mapsto R^2 = 0{,}782 \mapsto F = 19{,}081 \mapsto SH = 16{,}515$$

$$PC \mapsto y_i' = -77{,}52 + 28{,}863\,x_1 + 1{,}229\,x_2 - 0{,}313\,x_3 \mapsto R^2 = 0{,}781 \mapsto F = 18{,}942 \mapsto SH = 16{,}562$$

LS tekniğinde olduğu gibi RR sonuçları da geleneksel anlamlılık düzeylerinde anlamlı (F = 19,1 ve p = 0,000) olduğu görülmektedir [Tablo 2: 4]. Diğer bir anlatımla beden ağırlığı (Y), açıklayıcı değişkenler tarafından anlamlı bir şekilde açıklanmaktadır. Beklendiği gibi, tahmincilerin standart hataları da küçülmüştür [Tablo 2: 4 ve Tablo 3]. LS tekniği ile yanlı tahmin tekniklerinin sonuçları arasındaki en önemli çelişki ise, X2 değişkenine ait regresyon katsayısının işaretinde görülmektedir [Bkz: Tablo 3]. Diğer bir anlatımla, LS tekniğinde kuramsal beklentilerin aksine beden ağırlığı (Y) ile uyluk kemiğinin çevresinin uzunluğu (X2) değişkeni arasında negatif yönlü bir ilişki söz konusu iken, yanlı tahmin teknikleriyle elde edilen sonuçlarda kuramsal beklentilere uygun (pozitif yönlü) bir ilişki elde edilmiştir. Görüldüğü gibi, yanlı tahmin tekniklerinin (RR ve PC) sonuçları birbirini doğrulamaktadır.

Enküçük kareler (LS) regresyonu, temel bileşenler (PC) regresyonu ve ridge regresyon (RR) teknikleriyle elde edilen sonuçlar Tablo 3'te özetlenmektedir. Tablo 3 dikkatle incelendiğinde yanlı tahmin teknikleriyle (RR ve PC) elde edilen sonuçların birbiriyle örtüştüğü, fakat bu sonuçlar ile enküçük kareler (LS) tekniğiyle elde edilen sonuçlar arasında bazı tutarsızlıkların ve katsayı büyüklüklerinde önemli düzeyde farklılıkların olduğu açık bir biçimde görülmektedir.

4. SONUÇ

Bağımsız değişkenlerin çoklu doğrusal bağlantı ile ilgili istatistikleri incelendiğinde verilerde çok güçlü çoklu doğrusal bağlantı probleminin varolduğu anlaşılmaktadır. Bu nedenle LS tekniğiyle elde edilen sonuçlar geçerliliğini kaybetmektedir. Diğer bir anlatımla, çoklu doğrusal bağlantılı verilerde regresyon katsayılarının standart hataları, büyüklükleri ve işaretleri uygun bir biçimde tahmin edilememektedir. Verilerde çoklu doğrusal bağlantı olması durumunda RR ve PC teknikleri LS tekniğine göre daha durağan ve kuramsal (veya ampirik) beklentilere uygun sonuçlar sağlamaktadır. Her ne kadar LS ve yanlı tahmin tekniklerinden (RR ve PC) birisinin seçimi, yanlı veya yansız tahmincilerden birisinin seçimi anlamına gelse de, gerçekte durum bu değildir. Bilindiği gibi, pratik anlamda, LS tahmincileri sadece modelin hatasız tanımlanması durumunda yansızdırlar. Bu nedenle pratikte LS tahmincilerinin genelde yanlı olacağı kabul edilmektedir. Kısaca yanlı tahmin teknikleriyle, çoklu doğrusal bağlantı sorununu azaltmak amacıyla, birbiriyle anlamlı ilişki içinde olan açıklayıcı değişkenler birlikte analiz edilebilmektedir.

Araştırmamızda çoklu doğrusal bağlantı probleminin bir sonucu olarak tahmincilerin standart hatalarının yüksek ve X2 değişkenine ait katsayının işareti kuramsal ve ampirik beklentilerle çeliştiği görülmektedir. Ayrıca regresyon katsayılarının büyüklükleri de çoklu doğrusal bağlantıdan olumsuz bir şekilde etkilenmektedir. Bu olumsuzlukları düzeltmek için uygulanan yanlı tahmin tekniklerinin birbiriyle tutarlı ve kuramsal beklentilere uygun sonuçlar verdiği görülmektedir. Optimum yanlılık sabitini araştırmak amacıyla VIF ve ridge grafiklerinden yararlanarak, yanlı regresyon katsayılarının durağanlaştığı ve bu katsayılara ait VIF değerlerinin birlikte 1'e yaklaştığı bölgede yaklaşık bir k değeri seçilerek iterasyonlara başlanmaktadır (Tablo 2: Grafik 3.2-a ve Grafik 3.2-b). İterasyonlar sonucunda RR için seçilen optimum 0,02 yanlılık sabiti (k = %2) ile elde edilen sonuçlarda X2 değişkenine ait katsayının işareti değişmekte ve tahmincilerin standart hataları küçülmektedir. Ayrıca şunu vurgulamakta yarar var ki, RR tekniği için F testi geçerli değildir (tahminler yanlı olduğundan). Burada F istatistiği sadece bir endeks görevi gördüğünden dikkatle yorumlanmaktadır.

Sonuç olarak, beden ağırlığını açıklayan değişkenler arasında çoklu doğrusal bağlantı olduğundan, yanlı tahmin teknikleri enküçük kareler tekniğine göre daha tutarlı, geçerli, durağan ve kuramsal beklentilere uygun tahminler sağladığı görülmektedir.

KAYNAKÇA

Anderson, Björn (1998); *Scandinavian Evidence on Growth and Age Structure*, ESPE 1997 Conference at Uppsala University.

Darlington, R. B. (1978); "Reduced Variance Regression," *Psychological Bulletin*, 85, s. 1283-1255.

Dempster, A. P., M. Schatzoff, and N. Wermuth (1977); "A Simulation Study of Alternatives to Ordinary Least Square," *Journal of American Statistical Association*, 72, s. 77-91.

Draper, N. R., and H. Smith (1981); *Applied Regression Analysis*, John Willey, NY.

Faden, V. B. (1978); *Shrinkage in Regression and Ordinary Least Squares Multiple Regression Estimators*, Yayınlanmamış Doktora Tezi, University of Maryland.

Gujarati, D. N. (1995); *Basic Econometrics*, 3rd Ed., McGraw-Hill, New York.

Hoerl, A. E. and Kennard R.W. (1970); "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics,* 12, s. 69-82.

Kleinbaum D. G., Lawrence L. Kupper and Keith E. Muller (1988); *Applied Regression Analysis and Other Multivariable Methods*, Duxbury Press, New Jersey.

Maxwell, Scott E. (2000); "Sample Size in Multiple Regression Analysis," *Psychological Methods*, Vol. 5, No: 4, s. 435-458.

Myers, R. H. (1990); *Classical and Modern Regression with Applications*, Massachusetts: PWS-Kent Publishing Company, Boston.

Neter, J., W. Wasserman and M. Kunter (1990); *Applied Linear Statistical Models*, 3rd Ed., New Jersey.

NCSS Inc. (2001); *NCSS User Guide 2001*, Kaysville, NCSS Inc.

Orhunbilge, Neyran (2000); *Uygulamalı Regresyon ve Korelasyon Analizi*, Avcıol-Basım Yayın, İstanbul.

Price, B. (1979); "Ridge Regression: Application to Nonexperimental Data," *Psychological Bulletin*, 84, s. 759-766.

Rawlings, J. O. (1998); *Applied Regression Analysis: A Research Tool*, California.

Shin, Kilman (1996); *SPSS Guide for DOS Version 5 and Windows 6.1.2*, 2nd Ed., Irwin, Chicago.

SPSS Inc, (1999); *SPSS® Base 10 Application Guide*, Chicago: SPSS Inc.

Tracey, T. J., W. E. Sedlacek and R. D. Miras (1983); "Applying Ridge Regression to Admissions Data by Race," *College and University*, 58, s. 313-318.

Vinod, H. D. (1995); "Double Bootstrap for Shrinkage Estimators," *Journal of Econometrics*, 68, s. 287-302.