

**T.C.
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

**SİSTOLİK KAN BASINCINI ETKİLEYEBİLECEK
FAKTÖRLERİN RİDGE REGRESYON ANALİZİ İLE
İNCELENMESİ VE ÇOKLU BAĞLANTI PROBLEMİ**

YÜKSEK LİSANS TEZİ

Esra PAMUKÇU

Anabilim Dalı: İstatistik

Programı: Uygulamalı İstatistik

OCAK- 2010

**T.C
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

**SİSTOLİK KAN BASINCINI ETKİLEYEBİLECEK
FAKTÖRLERİN RİDGE REGRESYON ANALİZİ İLE
İNCELENMESİ VE ÇOKLU BAĞLANTI PROBLEMİ**

YÜKSEK LİSANS TEZİ

Esra PAMUKÇU

07233102

Tezin Enstitüye Verildiği Tarih : 29 Aralık 2009

Tezin Savunulduğu Tarih : 13 Ocak 2010

Tez Danışmanı : Yrd. Doç. Dr. Cemil ÇOLAK (F.Ü)

Diğer Jüri Üyeleri : Yrd. Doç. Dr. Sinan ÇALIK (F.Ü)

Yrd. Doç. Dr. Hasan BULUT(F.Ü)

Ocak-2010

ÖNSÖZ

Yanlı tahmin tekniklerinden biri olan Ridge Regresyon yönteminin En Küçük Kareler Yöntemi üzerine etkinliğini araştırmak amacıyla oluşturulan bu çalışmada genel hatlarıyla regresyon yöntemi, çoklu bağlantı problemi ve Ridge Regresyon yöntemi üzerinde durulmuştur. İleriye yönelik olarak toplanan verilere Ridge Regresyon yöntemi uygulanmış ve En Küçük Kareler üzerine etkinliği tartışılmıştır.

Bu çalışmanın oluşmasında ilgi ve yardımlarını esirgemeyen, özel olarak vakitlerini ayırıp sorularımı cevapsız bırakmayan başta danışman hocam sayın Yrd. Doç. Dr. Cemil ÇOLAK olmak üzere, bölüm başkanımız sayın Yrd. Doç. Dr. Sinan ÇALIK, sayın Yrd. Doç. Dr. Nurhan HALİSDEMİR'e ve diğer tüm bölüm hocalarıma, ayrıca veri alım sürecinde yardımlarını esirgemeyerek her türlü kolaylığı sağlayan Fırat Üniversitesi Tıp Fakültesi Kardiyoloji Bölümü öğretim üyesi sayın Doç. Dr. Mehmet AKBULUT'a ve asistanları Dr. Adil BAYDAŞ ve Dr. Zülfiye KUZU'ya saygı ve şükranlarımı bir borç bilirim.

Ayrıca iş hayatıyla beraber bu zorlu yolculukta maddi manevi hiçbir yardımı esirgemeyen kıymetli ailelerim, EMİR ve PAMUKÇU ailelerine, eşim Hilmi Emrah PAMUKÇU'ya teşekkürlerimi sunarım.

Esra PAMUKÇU
ELAZIĞ-2010

İÇİNDEKİLER

	<u>Sayfa No</u>
ÖNSÖZ.....	II
İÇİNDEKİLER.....	III
ÖZET.....	V
SUMMARY.....	VI
ŞEKİLLER LİSTESİ.....	VII
TABLOLAR LİSTESİ.....	VIII
SEMBOLLER LİSTESİ.....	IX
KISALTMALAR LİSTESİ.....	X
1.GİRİŞ.....	1
2. MATERYAL VE METOD.....	3
2.1. Basit Doğrusal Regresyon Modeli.....	3
2.2. En Küçük Kareler Yöntemi.....	5
2.3. Çoklu Doğrusal Regresyon Modeli.....	7
2.3.1. Çoklu Regresyonda Hipotez Testleri.....	10
2.3.1.1. Regresyon Katsayılarının Anlamlılığı İçin F Testi.....	10
2.3.1.2. Regresyon Katsayılarının Anlamlılığı İçin t Testi.....	11
2.3.1.3. Çoklu Korelasyon Katsayısının Anlamlılığının Sınanması.....	12
2.3.2. Çoklu Regresyon Modelinin Varsayımları.....	14
2.3.2.1. Hata Terimlerinin Ortalamasının Sıfır Olması.....	16
2.3.2.2. Hata Terimlerinin Varyansının Sabit Olması.....	16
2.3.2.3. Hata Terimlerinin Normal Dağılması.....	19
2.3.2.4. Hata Terimlerinin Bağımsız Olması.....	20
2.3.2.5. Gözlem Sayısının Fazla Olması.....	21
2.3.2.6. Bağımsız Değişkenler Arasında İlişki Olmaması.....	21
2.4. Çoklu Doğrusal Bağlantı Problemi.....	21
2.4.1. Giriş.....	21
2.4.2. Çoklu Bağlantının Kaynakları.....	22
2.4.3. Çoklu Bağlantının Etkileri.....	23
2.4.3.1. Çoklu Bağlantının EKK Kestirimlerine Olan Etkileri.....	23
2.4.3.2. Bağımlı Değişkenin Kestirimlerine Olan Etkileri.....	26
2.4.3.3. Hipotez Testlerine Olan Etkileri.....	27
2.5. Çoklu Bağlantıyı Belirleme Teknikleri.....	27

	<u>Sayfa No</u>
2.5.1. Çoklu Bağlantının $X'X$ Korelasyon Matrisiyle Belirlenmesi.....	28
2.5.2. Çoklu Bağlantının Belirlilik Katsayısı ile Belirlenmesi.....	29
2.5.3. Çoklu Bağlantının Kısmi Korelasyon Katsayıları İle İncelenmesi.....	29
2.5.4. Çoklu Bağlantının VIF(Variance Inflation Factors: Varyans Şişirme Faktörü) İle Belirlenmesi.....	29
2.5.5. Çoklu Bağlantının Tolerans Değerleri İle Belirlenmesi.....	30
2.5.6. Çoklu Bağlantının F Testi İle Belirlenmesi.....	30
2.5.7. Çoklu Bağlantının $X'X$ Matrisinin Özdeğerleri İle Belirlenmesi.....	31
2.5.8. Çoklu Bağlantının F Ve t Testi Karşılaştırması İle Belirlenmesi.....	31
2.6. Çoklu Bağlantıyı Giderme Yöntemleri.....	32
2.7. Ridge Regresyon.....	33
2.7.1. Ridge Kestircisinin Özellikleri.....	36
2.7.2. Ridge Parametresinin Seçimi.....	40
2.7.3. Ridge İzi.....	41
3. UYGULAMA.....	43
3.1. Gereç ve Yöntem.....	43
3.1.1. Araştırmanın Tasarımı ve Modeli.....	43
3.1.2. Araştırmanın Örneklem Büyüklüğü.....	44
3.1.3. Veri Toplama Yöntemi ve Süresi.....	45
3.1.4. Araştırmada Karşılaşılan Zorluklar.....	45
3.2. Bulgular.....	45
4. SONUÇLAR VE TARTIŞMA.....	60
5. ÖNERİLER.....	63
KAYNAKLAR.....	64
EKLER.....	67
ÖZGEÇMİŞ.....	73

ÖZET

Bu araştırmanın amacı, sistolik kan basıncını tahmin etmede yanlış tahmin tekniği Ridge Regresyon'un En Küçük Kareler tekniği üzerindeki etkinliğini araştırmaktır. Bu amaçla çoklu bağlantı problemi üzerinde durulduktan sonra sistolik kan basıncı ile bağımsız değişkenler arasındaki doğrusal ilişkinin tahmininde En Küçük Kareler regresyonu ve Ridge Regresyon tekniklerinin göreceli tahmin geçerlilikleri karşılaştırılmıştır. Araştırmada, bağımsız değişkenler arasındaki yüksek çoklu doğrusal bağlantı problemine dayanarak, Ridge Regresyon tekniğinin En Küçük Kareler regresyonuna göre daha düşük standart hatalı, daha durağan, tutarlı ve kuramsal beklentilere uygun tahminler sağladığı belirlenmiştir.

Anahtar Kelimeler: Regresyon Analizi, Çoklu Bağlantı Problemi, Ridge Regresyon

SUMMARY

The Investigation of Factors That Could Affect Systolic Blood Pressure With Ridge Regression Analysis and Multicollinearity Problem

The purpose of this study is to investigate the effectiveness of biased estimation technique Ridge regression over Least Squares technique in predicting systolic blood pressure. For this purpose, initially the problem of multicollinearity was examined then, the relative predictive validity of Least Squares and Ridge regression techniques was compared in predicting linear relation between systolic blood pressure and independent variables. In this study, by virtue of high degree of multicollinearity problem between independent variables, it was determined that Ridge regression technique would provide more stabilized coefficients and less standard error than would the Least Squares technique.

Key words: Regression Analysis, Multicollinearity Problem, Ridge Regression

ŞEKİLLER LİSTESİ

	<u>Sayfa No</u>
Şekil 2-1. Değişen varyanslılık.....	17
Şekil 2-2. Bir tahmin edici de varyans ve yanlılık.....	34
Şekil 2-3. Ridge kestiricisinde varyans ve yanlılık.....	39
Şekil 2-4. Ridge izi örnek şekil.....	42
Şekil 3-1. Ridge izi.....	51
Şekil 3-2. VIF grafiği.....	52

TABLÖLAR LİSTESİ

	<u>Sayfa No</u>
Tablo 2-1. Çoklu regresyon modelinde verilerin gösterimi.....	9
Tablo 2-2. Regresyonda değişkenliklerin varyans analiz tablosu.....	14
Tablo 2-3. Varyans dengeleme dönüşümleri.....	1
Tablo 3-1. Modeldeki değişkenler ve ölçü birimleri.....	43
Tablo 3-2. Tanımlayıcı istatistikler.....	46
Tablo 3-3. EKK çoklu regresyon sonuçları.....	47
Tablo 3-3a. EKK için ANOVA tablosu.....	47
Tablo 3-4. Korelasyon matrisi.....	48
Tablo 3-5. EKK çoklu bağlantının saptanması.....	49
Tablo 3-6. Korelasyonların özdeğerleri.....	50
Tablo 3-7. Standartlaştırılmış Ridge regresyon katsayıları.....	53
Tablo 3-8. VIF değerleri.....	55
Tablo 3-9. k analiz tablosu.....	56
Tablo 3-10. $k = 0,02$ için Ridge ve EKK karşılaştırması.....	57
Tablo 3-11. Ridge regresyon katsayıları.....	58
Tablo 3-12. $k = 0,02$ için ANOVA tablosu.....	59

SEMBOLLER LİSTESİ

$diag$	:Diagonal matris
$E()$	: Beklenen değer
e_i	: Artık (örnek)
iz	: Bir matrisin izi
k	: Parametre sayısı
k	:Yanlılık sabiti
K	: Koşul sayısı
Kov / Cov	: Kovaryans
n	: Gözlem sayısı
N	: Ana kütle
$N()$	: Normal dağılım
R^2	: Belirlilik katsayısı
Var	: Varyans
α_i	: Temel bileşenler kestiricisi
$\hat{\beta}$	: En küçük kareler tahmin edicisi
$\hat{\beta}^*$	: Ridge kestiricisi
β_i	: Çoklu doğrusal regresyon katsayısı
ε_i	:Artık (anakütle)
λ_i	: i. özdeğer

KISALTMALAR

EKK	: En Küçük Kareler
VIF	: Variance Inflation Factors (Varyans Şişirme Faktörü)
SPSS	: Statistical Package for the Social Sciences(Sosyal Bilimler İçin İstatistiksel Paket Program)
NCSS	:Number Cruncher Statistical System
HKO	:Hata Kareler Ortalaması
HKT	:Hata Kareler Toplamı
HDL	:High Density Lipoprotein(Yüksek Yoğunluklu Lipoprotein)
LDL	:Low Density Lipoprotein (Düşük Yoğunluklu Lipoprotein)
HGB	:Hemoglobin
HCT	:Hematokrit
MRFIT	:Multiple Risk Factor Intervention Trial Research Group(Çoklu Risk Faktörlerine Müdahale Deney Araştırma Grubu)
ANOVA	:Tek yönlü varyans analizi
LR	:Lojistik regresyon analizi
EAA	:Esnek ayırma analizi
YSA	:Yapay sinir ağları

1. GİRİŞ

Bir çok olay, ister ekonomik, ister iktisadi, ister tıbbi bir olay olsun, birden fazla sebebin ortak bir sonucu olarak ortaya çıkabilir. İlgilenilen olayın açıklamasında biri bağımlı diğeri bağımsız olmak üzere iki tip değişken bulunur.

Çoğu istatistik araştırmasının temel amacı da bağımlı değişkenin bağımsız değişken(ler) cinsinden tahmin edilmesini olanaklı kılan ilişkileri bulmaktır. Öyleyse yeni bir ürünün olası satışlarının fiyata, bir hastanın kilosunun perhiz yaptığı hafta sayısına, ailelerin sosyal harcamalarını ailenin gelirine göre v.b. tahmin etmek için araştırmalar yapılmaktadır.

Tahminler yapabilmek için, bağımlı (Y) ve bağımsız (X) değişken(ler) arasındaki ilişki mutlaka ölçülmelidir. Eğer bir korelasyon varsa, X bilindiğine göre Y 'nin tahmin edilmesi için bir regresyon denklemi geliştirilebilir. Regresyon, sonuçların tahmin edilmesini ve değişkenler arasındaki ilişkileri açıklamak için kullanılan bir yöntemdir. Basit ve çoklu regresyon yöntemleri, değişkenler arasında doğrusal bir ilişki olduğunda kullanılır [2].

Regresyon analizinde üç türlü amaç gözetilebilir:

1. *İddianın test edilmesi*: Örneğin trafik kazaları ile alkol tüketimi arasında bir fonksiyonel ilişki ileri sürülüyorsa, bu iddianın testi regresyon analizi ile araştırılabilir.
2. *Politika tesbiti*: Örneğin bir bölgede yeni bir departman açmayı düşünen bir işletme, o bölgede kendi malına olan talebi, talep fonksiyonu regresyonu ile araştırdıktan sonra buna karar verebilir.
3. *Geleceğe dönük ön tahmin*: Örneğin bir gazetenin aylık tiraj rakamları ile aylık harcamaları arasında doğrusal artan bir regresyon bulunmuşsa, öngörülen daha büyük bir reklam harcaması karşılığında gazetenin muhtemel aylık tirajının ne olacağı bu regresyon yardımıyla tahmin edilebilir [17].

Bu çalışmada; regresyon analizi basit ve çoklu regresyon olarak iki başlık altında incelendikten sonra, bu yöntemlerin varsayımları üzerinde ve bu varsayımlardan sapmalar olduğunda meydana gelebilecek sorunlar ve çözümleri üzerinde durulacaktır. Çalışmamızın ana temasını ise çoklu doğrusal regresyon modelinde, açıklayıcı değişkenler arasındaki bağımsızlık varsayımının bozulmasıyla ortaya çıkan çoklu bağlantı problemi ile

bu sorunun çözüm yöntemlerinden biri olan Ridge Regresyon yöntemi oluşturacaktır. Ridge regresyon yönteminin etkinliğini test edebilmek amacıyla sistolik kan basıncını etkileyebilecek faktörler üzerine bir uygulama yapılacaktır.

2. MATERİYAL VE METOT

2.1. Basit Doğrusal Regresyon Modeli

Regresyonda bir bağımlı değişken ve bir ya da daha çok bağımsız değişken vardır. Bağımsız değişkenin birden fazla olma durumu ilerleyen bölümlerde çoklu regresyon olarak incelenecektir. Bu bölümde bir bağımlı değişken ve bir bağımsız değişken varken bu iki değişken arasındaki ilişkiyi belirleyecek modelin nasıl olduğunun ortaya çıkartılması açıklanmaya çalışılacaktır. n tane birimin her birinden bağımlı değişken (Y) ve bağımsız değişken (X) değerleri saptanmış olsun. Bu durumda $(Y_1, X_1), (Y_2, X_2), \dots, (Y_N, X_N)$ olmak üzere n tane gözlem çifti olacaktır. Acaba Y ve X değişkenleri arasındaki ilişki nasıldır? Bu ilişkiyi matematiksel olarak ifade edilebilir mi? Bu soruların yanıtlarını verebilmek için $(Y_i, X_i) \quad i = 1, 2, \dots, N$ gözlem çiftlerini koordinat eksenlerine işaretlemek gerekir.

Bu işleme regresyonda **serpme diyagramının** hazırlanması denir. N tane gözlem çiftinin herbiri için serpme diyagramında kesişim noktaları bulunduğunda N tane nokta oluşacaktır. Bu noktaların konumuna bakılarak modelin nasıl olduğuna karar verilir. Eğer noktalar bir doğru etrafında toplanıyorsa doğrusal bir model kullanılmalıdır [33].

X ve Y serileri verilmiş olsun. Bunlar arasındaki ilişkinin doğrusal biçimde olduğu varsayıldığında, basit regresyon modeli ortaya çıkar. β_0 ve β_1 sabit sayılar olmak üzere fonksiyonel ilişkiyi açık olarak şöyle yazabiliriz:

$$Y_i = \beta_0 + \beta_1 X_i \quad i = 1, 2, \dots, N \quad (2.1)$$

Yukarıdaki ilişki bu haliyle deterministik bir ilişkidir. Yani X bağımsız değişkeni, Y bağımlı değişkenini kesin bir şekilde belirlemektedir. Oysa gerçek dünyada iki değişken arasında bu türden kesin ilişkilerle nadiren karşılaşılır [17]. Örneğin yukarıdaki ilişkide Y bir hastanın iyileşme süresini, X tedavide kullanılan ilacın dozunu gösteriyorsa, yukarıdaki ilişki bu haliyle eksik gösterilmiş olacaktır. Çünkü en azından bir hastanın iyileşme süresini belirleyen, tedavide kullanılan ilacın dozunun yanında başka faktörlerde olabilir. Tedavi süresi, hastalığın ağırlık derecesi, hastanın yaşı, hastalığın başlangıcından tedaviye kadar geçen süre... v.b. Öte yandan iyileşme süresinin (Y) gözlemlenmesinde ya da ölçülmesinde hatalar yapılmış olabilir.

Bütün bu sayılan fakat yukarıdaki fonksiyona dahil edilmeyen faktörler rassal faktörler olup önceden bilinemezler. İşte istatistiksel bir ilişki kurulurken sözü edilen bu rassal etkiler modele bir rassal terim olarak ilave edilir. Bu rassal terimi ε_i ile gösterelim. O zaman X ve Y arasındaki,

$$\text{Gerçek ilişki} \quad Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad i = 1, 2, \dots, N \quad (2.2)$$

$$\text{Gerçek regresyon doğrusu} \quad E(Y_i) = \beta_0 + \beta_1 X_i \quad i = 1, 2, \dots, N \quad (2.3)$$

olur.

Bu şekilde, Y de meydana gelen değişmeler, deterministik ve rassal değişmelerin toplam etkisiyle ortaya çıkmaktadır. Bu gerçek ilişki ve gerçek regresyon doğrusu, Y ve X e ait tüm anakütle değerleri elde edilmediği sürece bilinemezler. Ancak aynı model, X ve Y anakütlelerinden alınmış bir grup gözlem değeri (örneklem) üzerinden şu şekilde tanımlanabilir.

$$\text{Tahmin edilen ilişki} \quad \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i \quad i = 1, 2, \dots, n \quad (2.4)$$

$$\text{Tahmin edilen regresyon} \quad \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \quad i = 1, 2, \dots, n \quad (2.5)$$

Burada Y bağımlı değişkeni X bağımsız değişkeni üzerine regres edilmiş (bağlanmış)tır. Regresyonda $\hat{\beta}_0$ ve $\hat{\beta}_1$, sırasıyla gerçek ilişkideki β_0 ve β_1 regresyon parametrelerinin kestirim değerleridir (tahminleridir). Öte yandan regresyondaki e_i 'ler gerçek ilişkideki ε_i 'lerin tahminidir. e_i değerleri regresyon artıkları adını alır. Grafik üzerinde (X_i, Y_i) gözlem değerlerinin $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ ile gösterilen regresyon doğrusundan sapmalarının nedeni işte bu e_i artıklarıdır. Yani, $Y_i = \hat{Y}_i + e_i$ ilişkisi vardır. Regresyon doğrusunun üzerinde yer alan gözlem değerleri için e_i 'ler pozitif, altında kalan gözlem değerleri için e_i 'ler negatif olmakla beraber, e_i 'lerin cebirsel toplamı sıfırdır. Burada verilen, $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ regresyon denklemi sadece bir bağımsız değişken içerdiğinden, **basit regresyon** adını alır. Denklemden görüldüğü gibi, basit regresyonda sayısal değeri bulunması gereken iki katsayı vardır; $\hat{\beta}_0$ ve $\hat{\beta}_1$. X değişkeni 0 değerini aldığı anda regresyon doğrusu Y eksenini $\hat{\beta}_0$ noktasında keser. Bu yüzden $\hat{\beta}_0$ katsayısına **kesme terimi** adı verilir. $\hat{\beta}_1$ katsayısı ise regresyon doğrusunun eğimi olduğundan **eğim katsayısı** adı verilir.

Regresyon katsayılarının tahmini için en popüler yöntem En Küçük Kareler (EKK) yöntemidir. Regresyon doğrusunun gözlem değerlerini iyi bir şekilde temsil edebilmesi için, bu gözlem noktalarını tam olarak ortalaması gerekir. Bu şekilde e_i artıkları minimize edilmiş olacaktır. Bunun için EKK yönteminde gerçek ilişkiye bir terim olarak eklenen ε değişkeni hakkında şu varsayımlar geçerlidir:

1. ε bir rassal değişkendir.
2. ε rassal değişkeninin beklenen değeri sıfırdır. $E(\varepsilon_i) = 0$
3. ε rassal değişkeninin varyansı sabittir. $Var(\varepsilon_i) = \sigma^2$ sabit
4. ε rassal değişkeni normal dağılıma sahiptir. $\varepsilon_i \sim N(0, \sigma^2)$
5. ε rassal değişkeninin farklı terimleri arasındaki korelasyon sıfırdır. $Kov(\varepsilon_i, \varepsilon_j) = 0$
6. ε rassal değişkeni açıklayıcı değişkenlerden bağımsızdır. $Kov(\varepsilon_i, X_i) = 0$

Bu şartlar altında, artık kareleri toplamını minimize eden $\hat{\beta}_0$ ve $\hat{\beta}_1$ değerleri tespit edilerek regresyon katsayılarının bulunması EKK yöntemi olarak bilinmektedir [17].

2.2. En Küçük Kareler Yöntemi (EKK)

Regresyon modelinin belirlenmesi için serpilme diyagramı incelendiğinde doğrusal bir eğilim görülüyorsa X 'in Y 'ye göre matematik modelinin doğrusal olduğuna kesin olmasa da karar verilebilir. Ancak gözlem noktaları arasından çok sayıda doğru geçebilir. Bu doğrulardan en uygunu, tüm doğrusal fonksiyonlar arasından Y gözlem değerine en yakın tahmini $\hat{Y}$ değerini minimum hata ile veren doğrusal fonksiyon olacaktır. Yani;

$$\varepsilon = Y - \hat{Y} = Y - \beta_0 - \beta_1 X \quad (2.6)$$

minimum olacak şekilde bir fonksiyon seçilmelidir. Tüm gözlem noktaları için bu durum geçerli olduğuna göre

$$\sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2 = \text{minimum} \quad (2.7)$$

olmalıdır. Bu ifadeyi minimum yapacak şekilde çözüm yapılarak parametre kestirimlerinin elde edilmesine EKK regresyonu denmektedir. Bu ifadenin minimum olabilmesi için β_0 ve β_1 parametrelerine göre birinci mertebeden kısmi türevlerin sıfıra eşit olması gerekir.

$$\frac{\partial \varepsilon}{\partial \beta_0} = 2 \sum (-1)(Y - \beta_0 - \beta_1 X) = - \sum Y + n\beta_0 + \beta_1 \sum X = 0 \quad (2.8)$$

$$\frac{\partial \varepsilon}{\partial \beta_1} = 2 \sum (-X)(Y - \beta_0 - \beta_1 X) = - \sum XY + \beta_0 \sum X + \beta_1 \sum X^2 = 0 \quad (2.9)$$

Negatif terimleri eşitliğin sağ tarafına geçirirsek,

$$\sum Y = n\beta_0 + \beta_1 \sum X \quad (2.10)$$

$$\sum XY = \beta_0 \sum X + \beta_1 \sum X^2 \quad (2.11)$$

eşitlikleri elde edilir. Bu eşitliklere normal denklemleri adı verilir. β_0 ve β_1 bilinmeyenlerine bağlı bu iki bilinmeyenli denklemin çözümü $\sum X$; X değişkeninin toplamı, $\sum X^2$; X değişkeninin kareler toplamı, $\sum Y$; Y değişkeninin toplamı, $\sum XY$; iki değişkenin çarpımlarının toplamı yerine konularak yapılabilir.

Bilindiği gibi aritmetik ortalamadan farkların cebirsel toplamı sıfırdır. O halde yukarıdaki denklemlerde X ve Y değerleri yerine bunların aritmetik ortalamadan farklarını yazarak bazı kısaltmalar yapabiliriz. (2.10) ve (2.11) denklemlerinde,

$$\sum (Y - \bar{Y}) = 0 \text{ ve } \sum (X - \bar{X}) = 0 \quad (2.12)$$

ifadeleri yerine yazılabilir. Böylece,

$$\sum (Y - \bar{Y}) = n\beta_0 + \beta_1 \sum (X - \bar{X}) \quad (2.13)$$

$$\sum (X - \bar{X})(Y - \bar{Y}) = \beta_0 \sum (X - \bar{X}) + \beta_1 \sum (X - \bar{X})^2 \quad (2.14)$$

denklemleri elde edilir. (2.13)'den $n\beta_0 = 0$ ve (2.14)'den

$$\beta_1 = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2} \quad (2.15)$$

elde edilir. n gözlem sayısı sıfır olamayacağına göre $n\beta_0 = 0$ için β_0 'nın sıfır olması gerekir. Aritmetik ortalamadan farklar kullanıldığı için regresyon doğrusunun dikey eksenini kesmesi mümkün değildir. O halde β_0 parametresini elde edebilmek için fonksiyonun geçtiği $(\bar{X}, \bar{Y})$ noktası denklemde yerine yazarsak

$$\hat{\beta}_0 = \bar{Y} - b\bar{X} \quad (2.16)$$

yardımla β_0 hesaplanır. Böylece regresyon denklemi elde edilmiş olur.

β_1 parametresinin tanımı yukarıdaki formüle göre şöyle ifade edilebilir. Her iki değişkenin birlikte değişmesinden $[\sum(X - \bar{X})(Y - \bar{Y})]$, X değişkeninin kendi içindeki değişiklik $[\sum(X - \bar{X})^2]$ arındırıldığı için, β_1 parametresi X 'deki bir birimlik değişimin Y de ne kadarlık bir değişme yarattığını gösterir.

β_1 formülünün cebirsel açılımı yapılarak bir başka formül;

$$\hat{\beta}_1 = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}} \quad (2.17)$$

elde etmek mümkündür Bu formülden yararlanıldığında, aritmetik ortalamadan farklar alınmamakta, iki bilinmeyenli denklem çözümüne gerek kalmamaktadır [24].

2.3. Çoklu Doğrusal Regresyon Modeli

Sağlık alanındaki bağımlı değişkenler çoğunlukla iki ya da daha fazla bağımsız değişken tarafından etkilenebilirler. Biyolojik sistem karmaşık bir etkiler yumağıdır. Gözlenen bir sağlık değişkeninin değeri çok sayıda değişkenin etkilemesi sonucu ortaya çıkmaktadır. Bunlardan bazıları önemli etkiye sahip olan değişkenler (major factors) diğerleri önemsiz etkiye (minor factors) sahip olan değişkenlerdir. Bir değişkeni etkileyen iki ve daha fazla bağımsız değişken arasındaki neden- sonuç ilişkilerini doğrusal bir modelle açıklamak ve bu bağımsız değişkenlerin etki düzeylerini belirlemek için yararlanılan yöntem **çoklu doğrusal regresyon** analizi denir [25]

Çoklu doğrusal regresyonda, bağımlı değişkeni etkileyen birden çok bağımsız değişken söz konusu olup bu tür çalışmalarda, araştırmacıların iki genel amacı vardır:

1. Bağımlı değişkeni etkilediği düşünülen bağımsız değişkenlerden hangisi ya da hangilerinin bağımlı değişkeni daha çok etkilediğini bulmak.
2. Bağımlı değişkeni etkilediği belirlenen değişkenler yardımıyla bağımlı değişken değerini kestirebilmek [6].

Bir örnek ile açıklamak gerekirse, anne karnındaki bir bebeğin doğum ağırlığını (bağımlı değişken) önceden kestirebilmek amacıyla, annenin gebelik süresince beslenme durumu dikkate alınarak bu iki değişken arasında bir regresyon denklemi oluşturulsun. Eğer annenin beslenme durumunun, bebeğin doğum kilosunu açıklamakta yetersiz kaldığı görülürse, bağımlı değişkene etki ettiği düşünülen, gebelik haftası, anne yaşı, gebelik sayısı, canlı doğum sırası, annenin sosyo-ekonomik-kültürel özellikleri,... gibi bağımsız değişkenler de modele eklenerek bir çoklu regresyon modeli oluşturulabilir. Bu örnekteki birinci amaç, bulunacak çoklu regresyon denklemi yardımıyla bebeğin doğum kilosunu en çok hangi faktörün etkilediğini bulmak, ikinci amaç ise bebeğin doğum kilosunu önceden belirleyerek risk altındaki gebeliklere zamanında müdahale edebilmektir.

Genel olarak Y bağımlı değişkeninin k bağımsız değişkenli;

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + \varepsilon_i \quad i = 1, 2, \dots, N \quad (2.18)$$

denklemi, k bağımsız değişkenli çoklu doğrusal regresyon modeli olarak tanımlanır. β_j , $j=0,1,2,\dots,k$ parametreleri regresyon katsayıları olarak adlandırılır. β_j parametreleri, $i \neq j$ olmak üzere tüm X_i bağımsız değişkenleri sabit olduğunda, X_j deki her bir birimlik değişime karşılık Y deki beklenen değişimi gösterir. Bu nedenle β_j parametreleri kısmi regresyon katsayıları adını da alırlar [21].

Burada ana kütle ile ilgili gözlem sayısı her değişken için N dir. ε hata terimi olup, ortalaması sıfır standart sapması σ olan normal dağılıma sahiptir. Ancak anakütlenin tamamı için gözlemlerin yapılamaması nedeniyle “ n ” boyutlu örneklem için çoklu regresyon modeli, $\hat{\beta}$ değerleri β parametrelerinin kestirim (tahmin) değerleri olmak üzere şu şekilde ifade edilir.

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \dots + \hat{\beta}_k X_{ik} + e_i \quad i = 1, 2, \dots, n \quad (2.19)$$

Bu modele ilişkin veri seti, tablo halinde aşağıdaki gibi gösterilebilir:

Tablo 2-1: Çoklu regresyon modelinde verilerin gösterimi [21].

<i>Gözlem</i>	<i>Y</i>	<i>X₁</i>	<i>X₂</i>	...	<i>X_k</i>
<i>1</i>	<i>y₁</i>	<i>x₁₁</i>	<i>x₁₂</i>	...	<i>x_{1k}</i>
<i>2</i>	<i>y₂</i>	<i>x₂₁</i>	<i>x₂₂</i>	...	<i>x_{2k}</i>
<i>3</i>	<i>y₃</i>	<i>x₃₁</i>	<i>x₃₂</i>	...	<i>x_{3k}</i>
.	.	.	.	...	.
.	.	.	.	...	.
.	.	.	.	...	.
<i>n</i>	<i>y_n</i>	<i>x_{n1}</i>	<i>x_{n2}</i>	...	<i>x_{nk}</i>

Bu denklemin matris notasyonu ile gösterimi:

$$Y = X\hat{\beta} + e \quad (2.20)$$

ile verilir. Burada Y : $n \times 1$ boyutlu bağımlı değişken vektörü, X : $n \times (k + 1)$ boyutlu girdi matrisi, $\hat{\beta}$: $(k + 1) \times 1$ boyutlu katsayılar vektörü, e : $n \times 1$ boyutlu artık (hata) vektörü olmak üzere yukarıdaki eşitlik gözlemler cinsinden aşağıdaki gibi yazılır [6].

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1k} \\ 1 & X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

$\hat{\beta}$ kestirim deęerleri varsayımları saęlandığı zaman EKK yöntemi ile yapılabilir. $\hat{\beta}$ kestiricisi

$$\hat{\beta} = (X'X)^{-1}X'Y' \text{ dir. [7]} \quad (2.21)$$

2.3.1 Çoklu Regresyonda Hipotez Testleri

Çoklu regresyon denklemi elde edildikten sonra, çeşitli hipotezler test edilebilir. Öncelikle varyans analizi yapılarak, bağımlı deęişkenin bağımsız deęişkenler tarafından açıklanıp açıklanmadığı, dięer bir deyişle bağımlı deęişkenle bağımsız deęişkenler kümesi arasında doğrusal bir ilişki olup olmadığı test edilir [6].

Kestirimlerin anlamlı olup olmadığına karar verebilmek için F testi ve t testi gibi testler kullanılır. Bu testler hem regresyon katsayılarının, hem de çoklu korelasyon katsayısının anlamlılığı için kullanılabilir.

Bu testlerin yanısıra, modeldeki deęişkenler arasındaki ilişkinin derecesini yani kestirimlerin anlam derecesini belirleyen gösterge olarak çoklu korelasyon katsayısı kullanılabilir [14].

2.3.1.1. Regresyon Katsayılarının Anlamlılığı için F Testi

Regresyon analizinde F testi, birden çok bağımsız deęişkenin bağımlı deęişken üzerinde etkili olup olmadığını anlamak için kullanılabilir. Dięer bir ifade ile bağımlı deęişken Y' nin açıklayıcı deęişkenlerin hepsiyle doğrusal olarak bağımlı olup olmadığı, F testi ile sınanabilir. Bununla beraber F testi sonucunda Y ile açıklayıcı deęişkenler arasında ilişki olduğu sonucuna varılmasına rağmen, bu ilişkinin hangi deęişken nedeniyle meydana geldiği belirlenemez.

Modeldeki deęişkenler arasındaki ilişkinin anlamlı olup olmadığını belirlemek için hipotezler şu şekilde formüle edilir;

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k \quad (2.22)$$

$$H_1: \beta_1 \neq \beta_2 \neq \dots \neq \beta_k \neq 0$$

H_0 yokluk hipotezi, bütün parametrelerin sıfır olduğunu dolayısı ile bağımlı değişkenin bağımsız değişkenler tarafından açıklanamadığını, yani modelin istatistiksel olarak anlamlı olmadığını ifade eder. H_1 varlık hipotezi ise, bu parametrelerden en az ikisinin sıfırdan farklı olduğunu dolayısıyla modelin istatistiksel olarak anlamlı olduğunu ifade eder [8].

Çoklu korelasyon katsayısı (R) 'nin karesi olan **Belirlilik Katsayısı (R^2)** yardımıyla yapılan F testi için F formülü;

$$F = \frac{R^2 / (k-1)}{1-R^2 / (n-k)} = \frac{R^2}{(1-R^2)} \cdot \frac{(n-k)}{(k-1)} \quad (2.23)$$

Burada;

k : kestirilecek parametre sayısı

$k - 1$: bağımsız değişken sayısı

n : gözlem sayısıdır [14].

Belli bir α anlam düzeyinde F tablosundan $F_{(k-1, n-k)}$ ' e karşılık gelen değer bulunarak bu değer hesapladığımız F istatistiği ile karşılaştırılır. Eğer α önem düzeyinde $F_{(k-1, n-k)} > F$ ise sıfır hipotezi red edilir ve modeldeki bağımsız değişkenlerin bağımlı değişkeni açıkladığına yani modelin istatistiksel olarak anlamlı olduğunu karar verilir.

2.3.1.2 Regresyon Katsayılarının Anlamlılığı için t Testi

t testi modeldeki bağımlı değişken ile bu değişkeni açıklayan bağımsız değişkenler arasındaki ilişkiyi gösteren β parametrelerinin tek tek test edilmesinde kullanılır. t testi $n < 30$ ise ve anakütle varyansı bilinmiyorsa kullanılır. $n > 30$ ve anakütle varyansı biliniyorsa Z testinden faydalanılır [31]. Örnek olarak; modelden bir veya daha fazla değişkeni ilave ederek veya çıkartarak modelin daha etkin hale getirilmesi istenebilir. Regresyon modeline bir değişken ilave edilmesi genellikle regresyonun toplam karelerinin artmasına ve artıkların toplam karelerinin azalmasına yol açar. Modele ilave bağımsız değişken kullanmaya gerek olup olmadığına karar verebilmek için regresyon kareler toplamındaki artışın yeterli olup olmadığına karar vermek gereklidir. Ayrıca bir bağımsız değişkenin ilave edilmesi $\hat{Y}$ beklenen değerindeki varyansı da artırır. Bu yüzden sadece

bağımlı değişkeni açıklayabileceğine inanılan bağımsız değişken seçimine dikkat edilmelidir. Zira, önemsiz bir değişkenin ilave edilmesi modelin kullanışlılığını azaltabilen artık kareler ortalamasının artmasına sebep olacaktır.

β_j gibi tek bir regresyon katsayısının önemi için hipotez testleri:

$$H_0: \beta_j = 0 \quad (2.24)$$

$$H_1: \beta_j \neq 0$$

Hipotezler için test istatistiği:

$$t = \frac{\hat{\beta}_j}{\sqrt{1 - C_{jj}}} = \hat{\beta}_j \sqrt{\frac{1 - R_j^2}{\sigma^2}} \quad j = 1, 2, \dots, k \quad (2.25)$$

Burada C_{jj} standartlaştırılmış korelasyon matrisinin tersinin köşegen elemanlarıdır. t istatistiği belli bir α anlamlılık düzeyi ve $n - k$ serbestlik derecesine göre tablo değeri ile karşılaştırılır. Eğer t istatistiği, tablo değerinden büyükse sıfır hipotezi reddedilir. Bu ise X_k değişkeninin model için anlamlı olduğuna işaret eder. Tersine eğer sıfır hipotezi kabul edilirse bu X_k değişkeninin modelden çıkarılabileceğine işaret eder. Bu yönüyle t testine kısmi veya marjinal test de denilebilir [21].

2.3.1.3. Çoklu Korelasyon Katsayısının Anlamlılığının Sınanması

Regresyonda, birden fazla bağımsız değişkenin söz konusu olduğu olaylarda, bağımlı değişkenin gözlem değerleri ile tahmin değerleri arasındaki Pearson korelasyon katsayısına çoklu korelasyon katsayısı denir [32]. Ancak uygulamalarda yorumun daha kolay yapılabilmesi için çoklu korelasyon katsayısının karesi olan çoklu belirlilik katsayısı R^2 hesaplanır. R^2 , bağımsız değişkenlerin, bağımlı değişkeni yüzde ne kadar açıklayabildiklerinin gösteren bir ölçüttür. Bu ölçüt genel olarak 0.80 ve üstü olması halinde yeterli kabul edilir. 1'e ne kadar yaklaşırsa tanımlayıcılık gücü de o kadar artar. Buna ek olarak bağımlı değişkende olan değişimin ne kadarının hangi bağımsız değişken tarafından ortaya çıkarıldığı ve her aşamada modele giren bağımsız değişkenin ρ ve R^2 'ye ne kadar katkıda bulunduğu da incelenebilir. Model ne kadar iyi olursa ρ ve R^2 'de o kadar büyük olur [28].

Dağılımı normal ve çoklu korelasyon katsayısı sıfır olan bir anakütleden çekilen örneklerle dayanılarak hesaplanan belirlilik katsayılarından oluşan dağılımın ortalaması;

$$E(R^2) = \frac{k - 1}{n - 1} \quad (2.26)$$

olarak ifade edilebilir. Buradan görülmektedir ki bağımsız değişken sayısı $k - 1$, gözlenen birim sayısına yaklaştıkça değişkenler arasında hiç bir ilişki olmasa da R^2 'nin değeri 1'e yaklaşmaktadır. Bu bakımdan regresyon katsayılarının anlamlı olup olmadığını anlamak için korelasyon katsayılarının anlamlılığı test edilmelidir.

Bu amaçla bağımlı değişken Y , bağımsız değişkenler X_i olduğunda regresyonda değişkenlikler şöyle tanımlanır;

$$\text{Toplam değişkenlik: } \sum_{i=1}^n (Y_i - \bar{Y})^2$$

$$\text{Regresyonla açıklanan değişkenlik: } \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

$$\text{Regresyonla açıklanamayan değişkenlik: } \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad [14].$$

İşte bu değişkenlikler yardımıyla aşağıdaki varyans analiz tablosu oluşturulabilir.

Tablo2-2: Regresyon değişkenlikleri için varyans analiz tablosu

Değişimin kaynağı	Kareler toplamı	Serbestlik derecesi	Kare ortalama	F oranı
Regresyonla açıklanan	$\sum (\hat{Y}_i - \bar{Y})^2$	k-1	$\frac{\sum (\hat{Y}_i - \bar{Y})^2}{k - 1}$	$F: \frac{\frac{\sum (\hat{Y}_i - \bar{Y})^2}{k - 1}}{\frac{\sum (Y_i - \hat{Y}_i)^2}{n - k}}$
Hata	$\sum (Y_i - \hat{Y}_i)^2$	n-k	$\frac{\sum (Y_i - \hat{Y}_i)^2}{n - k}$	
Toplam	$\sum (Y_i - \bar{Y})^2$	n-1		$= \frac{\sigma_{\hat{Y}_i - \bar{Y}}^2}{\sigma_{Y_i - \hat{Y}_i}^2}$

Böylece elde edilen F oranı yardımıyla aşağıdaki hipotezler test edilebilir.

$$H_0: \rho = 0 \quad (2.27)$$

$$H_1: \rho \neq 0$$

$k - 1$ ve $n - k$ serbestlik derecelerine göre belirli bir α anlam düzeyinde F tablo *değeri* hesaplanan F istatistiğinden küçükse, sıfır hipotezi reddedilir. Bu çoklu korelasyon katsayısının anlamlı olduğunu ve modele dayalı olarak yapılan parametre tahminlerinin güvenilir olduğunu ifade eder. Aksi durumda sıfır hipotezi kabul edilerek çoklu korelasyon katsayısının anlamlı olmadığına karar verilir.

2.3.2. Çoklu Regresyon Modelinin Varsayımları

Basit doğrusal regresyonda olduğu gibi, çoklu doğrusal regresyonda da parametre tahminleri yapılırken öncelikle bazı varsayımlar göz önünde bulundurulmaktadır. Varsayımlardan sapmalar beraberinde bazı problemleri ortaya çıkarmakta ve bu durumun model üzerinde bazı olumsuz etkileri olabilmektedir. Bu varsayımlar:

1. Hata terimlerinin aritmetik ortalaması sıfır olmalı
2. Hata terimlerinin varyansı sabit olmalı
3. Hata terimleri normal bir dağılıma sahip olmalı
4. Hata terimleri birbirinden bağımsız olmalı
5. Gözlem sayısı parametre sayısından büyük olmalı
6. Bağımlı değişken ile bağımsız değişkenler arasında doğrusal bir ilişki olmalı
7. Bağımsız değişkenler arasında ilişki olmamalı

Genel olarak tahmin edicilerin aşağıdaki ilk üç koşulu sağlaması durumunda en iyi doğrusal tahmin ediciler olduğu kabul edilmektedir: (BLUE: Best Linear Unbiased Estimators)

1. Tahmin edilen istatistiğin beklenen değeri bilinmeyen anakütle parametresine eşitse,

$$E(\hat{Y}) = Y$$

buna yansız (unbiased) tahmin denilmektedir.

2. Diğer yöntemlerle elde edilen sonuçlarla kıyaslandığında en küçük varyansa sahip ise

$$Var(\hat{Y}_1) < Var(\hat{Y}_2) < Var(\hat{Y}_3) < \dots$$

buna etkili tahmin denilmektedir.

3. Tahmin, örnek terimlerinin doğrusal bir fonksiyonu ise

$$\bar{Y} = \sum Y/N = (1/N)(Y_1 + Y_2 + \dots) = Y_1/N + Y_2/N + \dots$$

buna doğrusal tahmin denilmektedir.

4. Tahmin örneklem büyüklüğü artarken gerçek değerine yaklaşıyorsa tutarlıdır denir.

Yukarıdaki varsayımlardan birisinin sağlanamaması durumunda EKK tahmin edicisi yanlı, tutarsız veya etkisiz olmaktadır [5].

İlerleyen kısımlarda bu varsayımlara ve bozulmaları halinde ortaya çıkacak bu problemlere değinilecektir.

2.3.2.1. Hata Terimlerinin Aritmetik Ortalamasının Sıfır Olması

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \dots + \hat{\beta}_k X_{ik} + e_i \quad i = 1, 2, \dots, n \quad (2.28)$$

regresyon modelinde, gözlem değerlerinin herbir değeri için hata terimi çeşitli değerler alabilir. Bunlardan regresyon doğrusunun üzerinde yer alan gözlem değerleri için e_i ' ler pozitif, altında kalan gözlem değerleri için e_i ' ler negatif olmakla beraber, e_i ' lerin cebirsel toplamının sıfır olması varsayımdır. Bu varsayıma göre Y değişkeninin beklenen değeri:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$$E(Y) = E(\beta_0 + \beta_1 X) + E(\varepsilon)$$

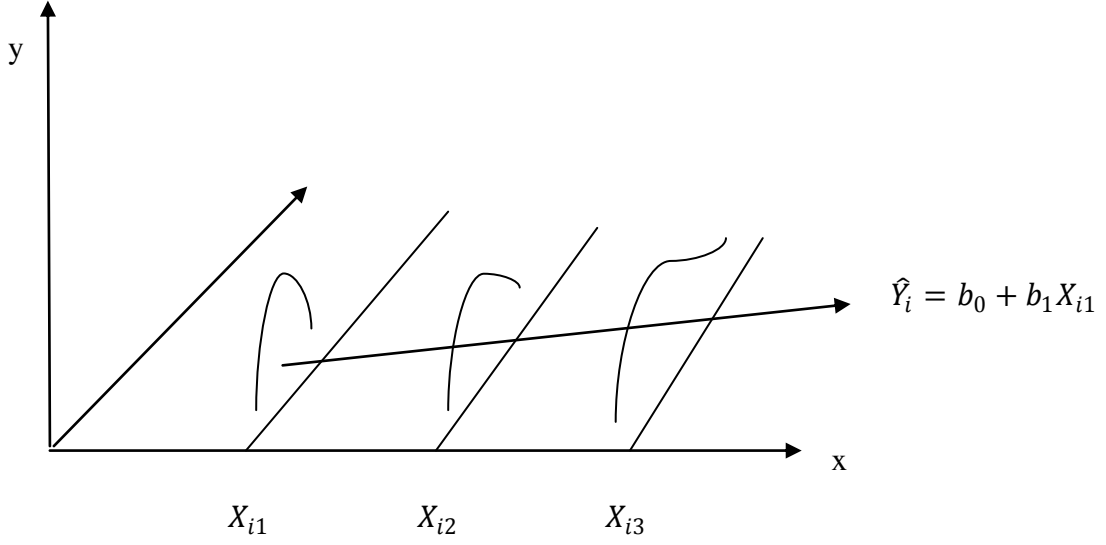
$$= \beta_0 + \beta_1 E(X) \quad (2.29)$$

olur ki bu ifade X ve Y arasında doğrusal bir ilişki olduğunu gösterir [31]. Bu varsayım sayesinde örnekten hareketle tahmin edilen regresyon doğrusu anakütle doğrusunun iyi bir tahmini olabilmektedir. Bu varsayımın sağlanmaması durumunda regresyon modeliyle yapılan parametre tahminleri gerçek değerlerinden, e_i 'lerin pozitif olması durumunda daha büyük, negatif olması durumunda daha küçük olurlar. Yani parametre kestirimleri sapmalı kestirimler olarak elde edilir [8].

2.3.2.2. Hata Terimlerinin Varyansının Sabit Olması

Homoscedasticity veya eş varyanslılık olarak da bilinen bu varsayımın bozulması durumunda değişen varyanslılık bir diğer adıyla Heteroscedasticity sorunu ortaya çıkmaktadır.

Bu varsayım sayesinde X 'in bütün değerleri için hata terimleri kendi ortalamaları etrafında aynı dağılımı gösterirler. Bu ise doğrusal regresyon modelinde, kestirimlerin standart hatalarının küçük olmasını dolayısıyla kestirimlerin daha isabetli olmasını sağlar [8]. Değişen varyanslılık Şekil 2-1'de gösterilmektedir.



Şekil 2-1: Değişen Varyanslılık [6].

Artık varyansları özellikle bağımlı değişkenin değerlerine bağlıdır. Özellikle bağımlı değişkenin değerlerine ilişkin değişim aralığının çok büyük olduğu durumlarda (örneğin 0-1000) bu sorunla karşılaşmak olasıdır. Diğer taraftan bağımlı değişkenin; varyansı o dağılımın ortalamasının bir fonksiyonu olan bir olasılık dağılımına uyması da değişen varyanslılığı ortaya çıkarır. Binom ve Poisson dağılımları bu özelliğe sahip iki yaygın olasılık dağılımlarıdır.

Eğer değişen varyanslılık sorunu araştırılmaz ve giderilmesi için gerekli önlemler alınmazsa elde edilen regresyon katsayıları yansız olmasına karşılık büyük standart hataya sahip olacaklardır. Bu ise parametrelere ilişkin güven aralıklarının genişlemesine ve katsayılara ilişkin testlerin düşük duyarlılıkta olmasına neden olacaktır [6].

Bu sorunun araştırılması için kullanılan yöntemlerden bazıları Grafik Yöntemi, Glejser testi, Spearman'ın Sıra Korelasyon Testi, Goldfield Quandt Testi, Breusch Pagon Testi'dir [31].

Bu testlerden herhangi birisiyle değişen varyanslılık durumu ortaya konur ise, bu sorunu giderebilmek için değişkenler üzerinde yapılacak dönüşümlerden yararlanmak sık başvurulan bir yöntemdir. Sadece Y , sadece X veya hem Y hem de X değişkenlerine

yapılabilen bu dönüşümlere **varyans dengeleme dönüşümleri** denir. Bazı varyans dengeleme dönüşümleri tablodaki gibidir.

Tablo 2-3: Varyans dengeleme dönüşümleri

Açıklama ve Y değişkeninin olasılık dağılışı	Dağılımın ortalaması açısından Y nin varyansı	Dönüşüm	Artık durumu
Y 'ler Poisson dağılışına uyan sayımlar ise	μ	$\sqrt{y}$	Sağa ya da sola megafon
Y 'ler Poisson dağılışına uyan sayımlar ve Y 'ler sıfıra yakın ya da çok küçükse	μ	$\sqrt{y + \sqrt{y + 1}}$ $\sqrt{y + 0.5}$ $\sqrt{y + 1}$	“
Y 'lerin dağılım genişliği çok büyük ve tüm Y_i 'ler pozitif ise	μ^2	$\log(y)$	“
Yukarıdakine ek olarak Y_i 'lerin bazıları sıfıra eşit ise	μ^2	$\log(y + 1)$	“
Y 'lerin sıfıra yakın olacak şekilde toplandığı ve pozitif olduğu durumlarda	μ^4	$\frac{1}{y}$	“
Yukarıdakine ek olarak bazı Y_i 'ler sıfır ise	μ^4	$\frac{1}{y + 1}$	“
Binom oranları için $0 \leq y_i \leq 1$	$\frac{\mu(\mu + 1)}{n}$	$\sin^{-1}(\sqrt{y})$	Elips biçimi

$1/y$ türü dönüşümler daha çok bağımlı değişkenin; bekleme süresi, yaşam süresi,..vb gibi değişkenleri aldığı durumlarda kullanılır. Değişen varyanslılık sorununu giderebilmenin bir diğer yolu da ağırlıklı en küçük kareler yöntemini kullanmaktır [6].

2.3.2.3 Hata Terimlerinin Normal Dağılımı

Aralık tahmini ve regresyon katsayılarıyla korelasyon katsayılarının testlerinin yapılabilmesi için hataların (e_i) lerin dağılımının, ortalaması sıfır standart sapması sabit olan bir normal dağılım olduğu kabul edilir. Yapılan testlerin güvenilebilir olması için bu varsayımın yerine gelmiş olması gerekir. Dolayısıyla hataların normal dağılıp dağılmadığı, üstünde önemle durulması gereken bir durumdur.

Normallikten sapma olup olmadığının araştırılmasında,

1. Momentlere dayanan çarpıklık (α_3) ve basıklık ölçüleri (α_4) hesaplanarak (α_3) = 0 ve (α_4) = 3 (normal dağılım) olup olmadığı incelenebilir
2. Q-Q nokta grafik yöntemine bakılabilir
3. Ki-kare uygunluk testi
4. Kolmogorov-Smirnov testi
5. Shapiro-Wilk testi
6. Anderson-Darling testi uygulanabilir [24].

Hataların normal dağılıp dağılmadığı bu yöntemlerden biri kullanılarak araştırılabilir. SPSS, NCSS gibi istatistik yazılımlarının çoğunda hataların normal dağılıp dağılmadığı grafik yöntemlerle görsel olarak verilmektedir. Bu grafiklerde hatalara ilişkin nokta dağılımının doğru üzerinde ya da çevresinde kümelenmiş olmaması ya da belirli şekiller göstermemesi hataların normal dağılmadığının bir göstergesidir. Hataların normal dağılmamasının nedeni aykırı değerler olabileceği gibi etkili gözlemlerin varlığı da olabilir. Böyle bir durum söz konusu ise aykırı veya etkili gözlemlerin veriden çıkarılması gerekebilir. Ayrıca örnek birim sayısı artarken normal dağılıma uygunluk söz konusu olabileceği gibi unutulmuş önemli bir açıklayıcı değişkenin modele ilave edilmesi de normalliği sağlayabilir [6].

Hataların diğer varsayımları sağlamaları halinde (özellikle eşit varyanslılık ve hataların bağımsız olması) normal dağılımdan sapmaların genellikle ciddi sonuçlar ortaya çıkarmadığı kabul edilmektedir [24].

2.3.2.4 Hata Terimlerinin Bağımsız Olması (Otokorelasyon Olmaması)

Bu varsayıma göre $i \neq j$ olmak üzere iki gözleme ait hata terimleri ε_i ve ε_j birbirinden bağımsız olmalıdır. Yani;

$$Kov(\varepsilon_i, \varepsilon_j) = 0 \text{ 'dır.} \quad (2.30)$$

Bu varsayımın bozulması otokorelasyon (öz ilişki) olarak bilinmektedir. Otokorelasyon birçok nedenle ortaya çıkmaktadır. En sık rastlandığı durum zaman serileridir. Nedenleri:

1. Önemli bir açıklayıcı değişkenin modele alınmamış olması
2. Gözlenen birim sayısının yetersiz olması
3. Modelin uygun seçilmemiş olması
4. Bağımsız değişkenlerin arasında ilişki olması

Otokorelasyonun regresyon analizine etkileri ise şunlardır:

1. EKK yöntemiyle elde edilen regresyon katsayıları tarafsızdır. Ancak standart hataları minimum olamaz
2. Örnek regresyon denkleminin standart hatası ve regresyon katsayılarının standart hataları olması gerekenden düşük çıkabilir.
3. Aralık tahmini ve istatistik testler bağımsızlık ve tesadüfilik varsayımına dayandıkları için geçerliliklerini kaybeder.

Otokorelasyonun varlığı için grafik yöntemi ve Durbin-Watson ile Von-Neumann testleri kullanılan yöntemlerdir. Grafikler görsel olarak yorum yapmada kolaylık sağlamasına rağmen bu yöntemle kesin karar vermek her zaman mümkün olmayabilir. Bu nedenle analitik testleri uygulamak daha kesin sonuçlar verecektir [24].

Otokorelasyon varlığı durumunda en çok kullanılan yöntem, en basit genelleştirilmiş en küçük kareler yöntemi iki aşamalı Cochrane-Orcutt (CO2) prosedürüdür. Bunun dışında modele yeni bir bağımsız değişken ilave edilebilir, gözlem sayısı artırılabilir, model yeniden tanımlanabilir veya model üzerinde çeşitli dönüşümler yapılabilir [16].

2.3.2.5. Gözlem Sayısının Fazla Olması

Çoklu doğrusal regresyon analizinde gözlem sayısının yetersiz olması başta çoklu bağlantı olmak üzere çeşitli problemlere sebep olabilir. Bu nedenle n gözlem sayısı, k regresyon modelindeki parametre sayısı olmak üzere $n > k$ olmalıdır.

2.3.2.6. Bağımlı Değişken ile Bağımsız Değişkenler Arasında Doğrusal İlişki Olması

Çoklu regresyon analizi, faktör analizi, yapısal eşitlik modeli ve diskriminant analizi gibi korelasyon katsayılarına dayanan çok değişkenli tekniklerin varsayımlarından birisi de doğrusallık varsayımdır. Doğrusal olmayan etkileşimler için hesaplanacak doğrusal korelasyonlar gerçek ilişkiyi her zaman daha düşük gösterecektir. Değişkenler arasında doğrusallık varsayımı sağlanmadığında bağımlı ve bağımsız değişkenlerin bazen her ikisine bazen sadece birine dönüşüm uygulamak koşuluyla doğrusallık varsayımı sağlanmaktadır [16].

2.3.2.7. Bağımsız Değişkenlerin İlişkili Olmaması

Bu varsayımın bozulması halinde Çoklu Bağlantı problemi ortaya çıkmaktadır. Bu çalışmamız, çoklu bağlantı problemi ve giderme yöntemlerinden biri olan Ridge regresyon yöntemiyle ilgili olduğundan bu başlık, ayrı bir bölüm olarak ele alınarak daha detaylı incelenecektir.

2.4. Çoklu Doğrusal Bağlantı Problemi

2.4.1. Giriş

Çoklu doğrusal regresyon modelinde bağımsız değişkenlerin birbirleriyle ilişkili olmaması varsayımının bozulması çoklu doğrusal bağlantı problemini ortaya çıkarmaktadır. Eğer bağımsız değişkenler arasında hiçbir ilişki yoksa yani bağımsız değişkenlerin ikişerli basit korelasyon değerleri sıfıra eşitse değişkenlerin dik (orthogonal)

olduğu söylenir. Ancak uygulamaların çoğunda bağımsız değişkenler arasında ilişki olmaması çok az rastlanan bir durumdur. Genellikle değişkenler arasında az da olsa bir ilişki söz konusudur. Eğer bu ilişki doğrusal bir ilişki ise çoklu doğrusal bağlantı ortaya çıkmaktadır. Bağlantının doğrusal olmaması durumunda çoklu doğrusal bağlantıdan söz edilemez. Bir başka ifadeyle çoklu doğrusal bağlantı bağımsız değişkenler arasında doğrusal bağlantılarla ilişkili olup doğrusal olmayan ilişkilerle ilgisi yoktur [8].

Çoklu bağlantı X ; $n \times (k+1)$ boyutlu veri matrisini göstermek üzere, $X_1, X_2, \dots, X_k$ kolonlarının doğrusal bağımsızlığı olarak da tanımlanabilir. O halde lineer (doğrusal) bağımlılık tanımı gereğince $t_1, t_2, \dots, t_k$ en az biri sıfırdan farklı olan skalerler olmak üzere;

$$\sum_{j=1}^k X_j t_j = 0 \quad (2.31)$$

ise $X_1, X_2, \dots, X_k$ vektörleri doğrusal bağımlıdır ve bu durumda **tam çoklu bağlantıdan** söz edilir. Yukarıdaki ifadeden de anlaşılacağı gibi herhangi bir X_i vektörü diğerleri cinsinden yazılabilir. Bu durumda $X'X$ matrisinin rankı $(k+1)$ 'den küçük olur ve $(X'X)^{-1}$ hesaplanamaz. Eğer,

$$\sum_{j=1}^k X_j t_j \cong 0 \quad (2.32)$$

ise o zaman **güçlü çoklu bağlantıdan** söz edilir. Bu durumda $(X'X)^{-1}$ hesaplanabilecek ancak bu durumda regresyon sonuçları üzerinde olumsuz etkileri olacaktır [6].

Bu olumsuz sonuçlar;

1. EKK tahminleri tahmin edilen parametrelerin gerçek değerlerinden oldukça farklıdır.
2. Tahminlerde yansızlık vardır, ancak tahminleri mutlak değerleri oldukça büyüktür. Veride ki çok küçük değişikliklerle tahmin edilen parametrelerin işaretleri değişir.
3. Şiddetli çoklu iç ilişki altında parametre tahminleri karasız olma eğilimi gösterecektir. Tahminlerin geçerliliğini görmek için yeni örneklemeler kullanıldığında tahminler şiddetle etkilenecek değişirler.

4. Ayrıca çoklu iç ilişki varlığında farklı EKK bilgisayar algoritmaları belirlenen model parametreleri için farklı tahminler ve işaretler verebilir [1].

2.4.2. Çoklu Bağlantının Kaynakları

Çoklu bağlantının kaynaklarının belirlenmesi bize çözüm arayışında bazı ipuçları verebilir. Çoklu bağlantı aşağıda belirtilen kaynaklardan bir veya bir kaç tanesinin ortak sonucu olarak ortaya çıkabilir. Bunlar:

1. **Örnekleme yöntemleri:** Veri toplama yöntemi; araştırmacının bilerek veya bilmeyerek bağımsız değişkenler uzayından bir alt uzayı örnekleme alması durumunda çoklu doğrusallığa neden olur [21]. Bu durumda gerçekte modelin kendisinde çoklu doğrusal bağlantı yoktur; ancak bağımsız değişkenlerden eksik ya da yetersiz bir alt kümenin seçilmesinden kaynaklı bir çoklu bağlantı söz konusudur.
2. **Aşırı tanımlanmış model:** Gözlem sayısının parametre sayısından az olması ($n < k$) durumu olarak da açıklanabilir. Daha çok tıbbi araştırmalarda karşılaşılan bu modellerde geçerli örnek birimi sayısı azdır. Bu noktadan kaynaklanan bir çoklu doğrusal bağlantıdan kurtulabilmek için önemine göre bazı değişkenleri modelden çıkartmak veya gözlem sayısını artırmak gerekebilir.
3. **Model ve anakütle üzerindeki fiziksel kısıtlar:** Bu durum, gerçekte anakütlede var olan ilişkinin örnekleme de ortaya çıkması olarak açıklanabilir. Kitledeki zorunluluklar daha çok bağımsız değişkenlerin kimyasal veya üretim proseslerinden ortaya çıkar. Örneğin bir kimyasal reaksiyonun gerçekleşmesi için belli içeriklerin sabit oranlarda olması vb... [34].

Bu üç ana başlığa ilaveten, araştırmacı kaynaklı bağımsız değişkenlerin seçimi de çoklu doğrusal bağlantıya sebep olabilir. Bir örnekle açıklamak gerekirse, kadının yaşı , gebelik sayısı ve doğum sayısı ayrı değişkenlermiş gibi ele alınsa da aslında her üçü de birbirleriyle yakından ilişkilidir. Çünkü kadının yaşı arttıkça gebelik sayısı ve gebelik sayısı arttıkça da doğum sayısı artacaktır. Bu açıdan bir çalışmada araştırmacı tarafından her üçünün de ayrı ayrı değişkenlermiş gibi modele alınması çoklu bağlantıya sebep olabilir [28].

2.4.3. Çoklu Bağlantının Etkileri

2.4.3.1. Çoklu Bağlantının EKK Kestirimlerine Olan Etkileri

Çoklu bağlantının, regresyon katsayılarının EKK kestirimleri üzerinde oldukça önemli etkileri vardır. Bu etkileri daha iyi ve daha kolay bir şekilde görebilmek için iki bağımsız değişkenli bir doğrusal regresyon modelini ele alalım

$$\hat{Y} = \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + e \quad (2.33)$$

olmak üzere vektörel olarak

$$\begin{aligned} e'e &= \sum_{i=1}^n e_i^2 = (Y - \hat{Y})'(Y - \hat{Y}) \\ &= (Y - X\hat{\beta})'(Y - X\hat{\beta}) \\ &= Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta} \end{aligned} \quad (2.34)$$

yazılabilir. $\hat{\beta}'$ ya göre kısmi türev alınarak sıfıra eşitlenirse;

$$\frac{\partial \sum_{i=1}^n e_i^2}{\partial \hat{\beta}'} = -2X'Y + 2X'X\hat{\beta} = 0 \quad (2.35)$$

eşitliğinden EKK normal denklemleri;

$$X'X\hat{\beta} = X'Y \quad (2.36)$$

olarak elde edilir. Bu ise,

$$\begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} r_{1y} \\ r_{2y} \end{bmatrix} \quad (2.37)$$

şeklinde yazılabilir. Burada r_{12} , X_1 ve X_2 değişkenleri arasındaki korelasyonu, r_{1y} ve r_{2y} ise bağımsız değişkenler ile Y arasındaki korelasyonu göstermektedir. $X'X$ matrisinin tersi C ile gösterilmek üzere

$$C = (X'X)^{-1} = \begin{bmatrix} 1 & -r_{12} \\ \frac{(1-r_{12}^2)}{(1-r_{12}^2)} & \frac{(1-r_{12}^2)}{(1-r_{12}^2)} \\ -r_{12} & 1 \\ \frac{(1-r_{12}^2)}{(1-r_{12}^2)} & \frac{(1-r_{12}^2)}{(1-r_{12}^2)} \end{bmatrix} \quad (2.38)$$

ve buradan,

$$\hat{\beta}_1 = \frac{r_{1y} - r_{12}r_{2y}}{(1-r_{12}^2)} , \quad \hat{\beta}_2 = \frac{r_{2y} - r_{12}r_{2y}}{(1-r_{12}^2)} \quad (2.39)$$

olarak bulunur. Eğer X_1 ve X_2 değişkenleri arasında kuvvetli bir ilişki varsa r_{12} korelasyon katsayısı değişkenler arasındaki ilişki derecesine bağlı olarak

$$|r_{12}| \rightarrow 1 \quad (2.40)$$

olur. Bu ise,

$$Var(\hat{\beta}_1) = C_{11} = \frac{1}{(1-r_{12}^2)} \rightarrow \infty \quad (2.41)$$

olmasına neden olur. Yine r_{12} nin +1 veya -1'e yaklaşmasına bağlı olarak

$$Cov(\beta_1, \beta_2) = C_{12} \rightarrow \mp \infty \quad (2.42)$$

yaklaşacaktır [21].

Böylece X_1 ve X_2 arasında var olan kuvvetli ilişki, regresyon katsayılarının EKK tahminlerinin büyük varyans ve kovaryanslara sahip olmasına neden olacaktır. Büyük varyans durumu, regresyon katsayı kestirimlerinin her bir örnek verisi için bir hassasiyet göstergesi olmasından dolayı, farklı örnekler kullanıldığında oldukça farklı katsayı kestirimleri ortaya çıkacaktır [14].

İkiden fazla bağımsız değişkenin bulunduğu modellerde çoklu bağlantı durumundan benzer şekilde etkilenecektir. Bunun için;

$$\hat{\beta} = (X'X)^{-1}X'Y \quad (2.43)$$

EKK kestiricisinin kovaryansını inceleyecek olursak;

$$\begin{aligned} Cov(\hat{\beta}) &= Cov[(X'X)^{-1}X'Y] \\ &= Cov[(X'X)^{-1}X'(X\beta + \varepsilon)] \end{aligned}$$

$$= Cov[\beta + (X'X)^{-1}X'\varepsilon] \quad (2.44)$$

β bir sabit olup kovaryansı etkilemeyeceğinden

$$\begin{aligned} &= Cov[(X'X)^{-1}X'\varepsilon] \\ &= (X'X)^{-1}X'Cov(\varepsilon)[(X'X)^{-1}X']' \end{aligned} \quad (2.45)$$

ε birbirinden bağımsız ve σ^2 sabit varyanslı olduğundan

$$\begin{aligned} &= (X'X)^{-1}X' \cdot \sigma^2 I \cdot [(X'X)^{-1}X']' \\ &= \sigma^2 (X'X)^{-1}X'[(X'X)^{-1}X']' \\ &= \sigma^2 (X'X)^{-1} \end{aligned} \quad (2.46)$$

elde edilir.

$C = (X'X)^{-1}$ matrisinin köşegen elemanları ise;

$$C_{jj} = \frac{1}{1 - R_j^2} \quad j = 1, 2, \dots, k \quad (2.47)$$

R_j^2 : X_j ile kalan $k - 1$ değişkenin regresyonundan elde edilen belirlilik katsayısıdır. X_j ile diğer $k - 1$ değişken arasında çoklu doğrusal bağlantı varsa R_j^2 değeri 1'e yaklaşır.

Dolayısıyla

$$Var(\beta_j) = \sigma^2 C_{jj} = \sigma^2 (1 - R_j^2)^{-1} \quad (2.48)$$

ile parametre kestirimlerinin varyansı

$$Var(\beta_j) > \sigma^2 \quad (2.49)$$

olur. Böylece ikiden fazla bağımsız değişken arasında da çoklu bağlantı varlığının, EKK kestirimlerinin varyansını arttırdığı söylenebilir. Doğrusal çoklu bağlantının artması sonucunda artan varyans, regresyon katsayıları için EKK kestirimlerinin güven aralıklarının genişlemesine neden olur.

Çoklu doğrusal bağlantı, regresyon katsayılarının kestirimlerini işaretçe de etki eder. Bu durumda kestirimlerin işaretleri, bağımsız değişkenle bağımlı değişken arasındaki ilişkiyi

yanlış göstermiş olacaktır. Yani araştırmacının bir önsav olarak pozitif çıkmasını beklediği parametre kestiriminin işareti negatif, negatif çıkmasını beklediği bir parametre kestiriminin işareti pozitif çıkabilir. Sonuçta, bağımsız değişkenler arasındaki çoklu bağlantı, bu değişkenlerin kestirim yeteneklerinin zayıf olmasına neden olur.

2.4.3.2. Bağımlı Değişkenin Kestirimlerine Olan Etkileri

Çoklu doğrusal bağlantı, regresyon parametrelerini değerce ve işaretçe etkilediğinden oldukça ayrı kestirimler ortaya çıkabilir. Regresyon katsayılarının gerçek katsayılardan değerce ve işaretçe çok farklı olması $\hat{Y}$ 'ları etkilediğinden, $\hat{Y}$ kestirimlerinin de standart hataları büyür.

2.4.3.3. Hipotez Testlerine Olan Etkileri

Daha önceki bölümlerde regresyon katsayılarının sınanması ile ilgili başlık altında

$$H_0: \beta_j = 0$$

$$H_1: \beta_j \neq 0 \quad (2.50)$$

şeklinde, parametrelerin ayrı ayrı anlamlılıklarını sınamak için kurulan hipotezlerde, H_0 'ı H_1 'e karşı test etmek için kullanılan t istatistiği;

$$t = \frac{\hat{\beta}_j}{\sqrt{1 - C_{jj}}} = \hat{\beta}_j \sqrt{\frac{1 - R_j^2}{\sigma^2}} \quad j = 1, 2, \dots, k \quad (2.51)$$

dır. Çoklu bağlantı halinde R_j^2 değerinin 1'e yaklaşması ile t değeri de 0'a yaklaşır. t testi sonucunda H_0 'ın reddi zorlaşarak o parametrenin sıfırdan farklı olmadığı ve dolayısıyla ilgili bağımsız değişkenin bağımlı değişkeni etkilemediğine karar verilebilir. Böylece çoklu bağlantı t değerlerini küçülterek değişkenlerin önemliliğinde yanlış bulgulara sebep olabilir.

Sonuç olarak; çoklu bağlantının açıklanan tüm bu sakıncalı etkilerinin giderilebilmesi için çoklu bağlantının varlığının araştırılması ve var olması halinde giderme yöntemlerinin belirlenerek probleme uygulanması gereklidir.

2.5. Çoklu Bağlantıyı Belirleme Teknikleri

Çoklu bağlantıyı ortaya çıkartan bir çok işaret olabilir. En basit haliyle, iki değişken arasındaki basit korelasyon katsayısının 1'e yakın olması (teorik olarak 0.80 ve üstü olması), çoklu bağlantıya bir işaret olabilir.

Regresyon katsayılarının büyüklüğü ve işareti de bazen bir işaret olabilir. Buna göre, regresyon modelindeki katsayılarının büyüklüğü ve işaretlerin beklenenin tersi çıkması da yine çoklu bağlantının işareti olabilir.

Bir bağımsız değişkenin ya da gözlemin modele eklenmesi ya da modelden çıkartılması sonucunda regresyon katsayılarında çok büyük değişikliklerin meydana gelmesi de çoklu bağlantıya işaret olabilir.

Ancak, çoklu bağlantının belirlenmesinin yanı sıra, derecesinin belirlenmesi de anlamlıdır. Bunun için kullanılan bazı yöntemlere aşağıda kısaca değinilmiştir [6].

2.5.1. Çoklu Bağlantının $X'X$ Korelasyon Matrisiyle Belirlenmesi

Çoklu doğrusal bağlantının belirlenmesinde uygulaması en kolay yöntemlerden biridir. Bu yöntemde bağımsız değişkenler;

$$X_{ij} = \frac{X_{ji} - \bar{X}_j}{\sum_{j=1}^p (X_{ji} - \bar{X}_j)^2} \quad (2.52)$$

biçiminde standartlaştırılıp elde edilen $X'X$ standartlaştırılmış korelasyon matrisindeki köşegen dışı r_{ij} elemanları kontrol edilir. Farrar ve Glauber (1967)'de geometrik olarak r_{ij} 'yi x_i ve x_j arasındaki açının kosinüsü olarak tanımlamışlardır [27]. x_i ve x_j açıklayıcı değişkenleri lineer bağımlı olduğunda $|r_{ij}|$ 'nin 1 veya 1'e çok yakın olması ilgili

değişkenler arasında doğrusala yakın bir ilişki olduğunu; yani çoklu doğrusal bağlantının olabileceğini bildirir.

Ancak iki değişken arasındaki kısmi korelasyon katsayısının büyük bir değer almaması çoklu doğrusal bağlantı olmadığı anlamına gelmez. Yani istatistik açıdan anlamlı korelasyonlar her zaman çoklu doğrusal bağlantı problemine yol açmamaktadır. Lawrence Klein'e göre basit korelasyon katsayısı r , çoklu korelasyon katsayısından küçükse çoklu bağlantı problemi ortaya çıkmayabilir [5]. Bilgisayar çıktılarından kolayca elde edilebilen bu değerlerin incelenmesinde hiç bir sakınca olmamasına karşın sadece bu yöntemle çoklu bağlantının varlığına karar vermek sağlıklı olmayabilir.

2.5.2. Çoklu Bağlantının Belirlilik Katsayısı ile Belirlenmesi

Bu yöntemde, modele bağımsız değişkenler ilave edildikçe R^2 'deki değişimler incelenmektedir. R^2 'de önemli bir gelişme sağlanamazsa bu durum çoklu bağlantının bir işareti olabilir [16].

2.5.3. Çoklu Bağlantının Kısmi Korelasyon Katsayıları ile İncelenmesi

İki değişken arasındaki basit korelasyon katsayısı anlamlı fakat kısmi korelasyon katsayıları anlamsız çıkıyorsa bu durum çoklu bağlantı problemine işaret olabilir. Ancak yine bu yöntem de her zaman sağlıklı sonuçlar vermeyebilir. Diğer bir değişle kısmi korelasyon katsayılarının yüksek olması durumunda bile çoklu bağlantı problemi olabilmektedir [5].

2.5.4. Çoklu Bağlantının VIF(Variance Inflation Factors=Varyans Şişirme Faktörü) ile Belirlenmesi

$(X'X)^{-1}$ matrisinin j-inci köşegen elemanına j-inci varyans şişirme faktörü denir ve (VIF_j) ile gösterilir [18]. İlk olarak Farrar ve Glauber (1967) tarafından çoklu bağlantıyı belirlemek için kullanılmış fakat Marquardt(1970) tarafından VIF olarak

adlandırılmıştır [27]. VIF değerlerinin hesaplanmasını göstermek için aşağıdaki gibi üç bağımsız değişkenli bir model ele alınırsa;

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + e. \quad (2.53)$$

Birinci adımda, X_1 bağımsız değişkeni bağımlı değişken olarak alınıp diğer bağımsız değişkenlerle çoklu korelasyon katsayısı (R^2) hesaplanır. Böylece X_1 değişkeni için varyans şişirme faktörü;

$$VIF(X_1) = \frac{1}{1 - R_1^2} \quad (2.54)$$

olarak hesaplanmaktadır. İkinci ve üçüncü adımlarda X_2 ve X_3 için aynı yöntemle

$$VIF(X_2) = \frac{1}{1 - R_2^2} \quad ve \quad VIF(X_3) = \frac{1}{1 - R_3^2} \quad (2.55)$$

değerleri hesaplanır. Bağımsız değişkenler arasında ilişki yoksa ($R^2 = 0$) $VIF = 1$ olacaktır. Bağımsız değişkenler arasında tam bir ilişki varsa ($R^2 = 1$) $VIF = \infty$ olacaktır. Eğer bağımsız değişkenler arasında kuvvetli ilişki varsa ($R^2 = 0.90$) $VIF = 10$ olacaktır. Webster (1992) VIF için şu genel kuralı vermektedir. $VIF \geq 10$ ise anlamlı çoklu bağlantıdan söz edilebilir [16].

2.5.5. Çoklu Bağlantının Tolerans Değerleri İle Belirlenmesi

Çoklu bağlantının belirlenmesinde kullanılan bir diğer yöntem tolerans değerlerini hesaplamaktır. Tolerans değeri;

$$T = 1 - R_j^2 \quad j = 1, 2, \dots, k \quad (2.56)$$

olarak hesaplanır. Böylece çoklu bağlantı halinde VIF değeri büyük dolayısıyla tolerans değeri küçük çıkacaktır [5].

2.5.6. Çoklu Bağlantının F Testi İle Belirlenmesi

Yukarıdaki üç bağımsız değişkenli örnekte F değerlerinin hesabı için, bağımlı kabul edilen her bir bağımsız değişken ile diğer bağımsız değişkenler arasında $R_{1.23}$, $R_{2.13}$, $R_{3.12}$ çoklu korelasyon katsayıları hesaplanır. Daha sonra bu değerlerden faydalanan her bir bağımsız değişken için F değeri aşağıdaki gibi hesaplanmaktadır.

$$F_{1.23} = \frac{R_{1.23}^2 / (k - 1)}{(1 - R_{1.23}^2) / (n - k)} \quad (2.57)$$

Formülde n , toplam örnek birim sayısını, k ise sabit terim dahil tahmin edilecek parametre sayısını göstermektedir. Hesaplanan F değeri belirli bir anlamlılık düzeyinde $F_{k-2, n-k+1}$ değeriyle karşılaştırıldığında, F değeri tablo değerinden büyükse X_1 değişkeniyle diğer değişkenler arasındaki ilişkinin anlamlı olduğuna karar verilir. Benzer yöntemle diğer değişkenler arasındaki ilişki de incelenebilir [16].

2.5.7. Çoklu Bağlantının $X'X$ Matrisinin Özdeğerleri İle Belirlenmesi

Vinod ve Ullah (1981), çoklu bağlantıyı ciddi anlamda çalışan ilk araştırmacı olan Ragnar Frisch (1934)'in çoklu bağlantıyı özdeğerlerle ilişkilendirdiğini söylemişlerdir. Fakat bilgisayar programlarının yetersiz olması nedeniyle $X'X$ 'in özdeğerlerinin sayısal analizi desteklenememiştir [27]. İlk olarak Vinod ve Ullah (1981) koşul sayısını en büyük özdeğerin en küçük özdeğere bölümünün karekökü olarak tanımlamışlar, daha sonra Montgomery ve Peck(1982) ise en büyük özdeğerin en küçük özdeğere bölümü olarak belirtmişlerdir [32].

λ_{min} ve λ_{max} $X'X$ matrisinin minimum ve maksimum özdeğerleri olmak üzere

$$K = \frac{\lambda_{max}}{\lambda_{min}} \quad (2.58)$$

olarak tanımlanan koşul sayısı çoklu bağlantının derecesini de belirlemek için kullanılan çok yaygın bir yöntemdir. Gujarati (1995)'de koşul sayısı için şu genellemeyi yapmıştır. $K > 30$ olması genel olarak bir çoklu doğrusal bağlantının olduğunu gösterir. Ancak

$K < 100$ için bu durum önemli değildir. $100 < K < 1000$ olması çoklu bağlantının güçlü olduğunu, $K > 1000$ olması durumunda ise çoklu bağlantının ciddi boyutlarda olduğunu gösterir [4,7,10].

2.5.8. Çoklu Bağlantının F ve t Testi Karşılaştırması ile Belirlenmesi

Regresyon modelinin anlamlılığı konusunda bilgi veren F istatistiği ile t istatistiği de çoklu bağlantıya işaret edebilir. Özellikle, katsayılara ilişkin t istatistiklerinin tümü anlamsız iken F istatistiğinin anlamlı çıkması çoklu bağlantının varlığı konusunda bir işaret olabilir. Ancak, anlamlı derecede çoklu bağlantı sorunu olan birçok veri seti bu şekilde bir sonuç sergilemediği için, bu yaklaşım kuşkuyla olabilmektedir [21].

Çoklu bağlantının tespitinde kullanılan bütün bu farklı yaklaşımların herbirinin avantajlı ve dezavantajlı yönlerinin olduğu unutulmamalıdır. Bu nedenle sorunun daha açık ve kesin bir şekilde ortaya çıkarılabilmesi için birlikte kullanılmalıdırlar. Örnek olarak özdeğerlerin büyüklükleri bağımlılığın derecesini belirlerken, VIF değeri verideki olası bağımlılıkların her bir regresyon katsayısı üzerindeki olumsuz etkilerini açıklayabilir [32].

2.6. Çoklu Bağlantıyı Giderme Yöntemleri

Daha önceki bölümlerde çoklu regresyon varsayımlarından olan çoklu bağlantı olmaması varsayımının bozulmasının zararlı etkileri üzerinde durulmuştu. Eğer yukarıdaki yöntemler yardımıyla çoklu bağlantı varlığı tespit edilmiş ise, regresyon üzerindeki zararlı etkilerinden dolayı ortadan kaldırma veya etkisini azaltma yoluna gidilmelidir.

Çoklu bağlantının giderilebilmesi için önerilen birçok yöntem vardır. Ancak veriye yeni gözlem eklenmesi, modelin yeniden belirlenmesi veya bazı yanlı kestirim yöntemlerinin uygulanması en yaygın kullanılan yöntemlerdir. Her bir yöntemin yine kendine göre uygulama alanı ve sakıncalı yönleri olabilir. Kısaca;

- Örneklemin evreni iyi temsil edememesi nedeniyle ortaya çıkacak bir çoklu bağlantı durumunda veriye yeni gözlemlerin eklenmesi önerilir. Fakat örneğe birim ilave etmek her zaman mümkün olmayabilir.

- Bir yada daha fazla bağımsız değişken modelden çıkarılabilir. Bu da modelin yeniden tanımlanmasıdır. Fakat hangi değişkenler çıkarılacaktır? Böyle bir yaklaşım modeli yanlış tanımlamaya götürebilir. Ayrıca bu yöntem örneklemin evreni temsil edemediği durumlarda kullanılmamalıdır. Çünkü modele gerçekten katkı yapabilecek bir değişken çoklu bağlantı nedeniyle model dışında bırakılabilir.
- Çoklu bağlantının bağımsız değişkenler arasındaki gerçek ilişki nedeniyle ortaya çıktığı durumlarda, diğer bir değişle sorunun bir örnekleme sorunu olmadığı durumlarda, uygulanabilecek bir yöntem, çoklu bağlantılı değişkenlerin birleştirilerek yeni bir değişken yaratılması ve modele çoklu bağlantılı değişkenler yerine bu değişkenin konulmasıdır. Bu da modeli yeniden tanımlama işlemidir.
- Son olarak da veride hiç bir değişiklik yapmadan, veri ekleyip çıkarmadan mevcut veri üzerinde temel bileşenler regresyonu, Ridge regresyonu gibi bazı yanlı kestirim yöntemlerini uygulamak da bir yöntemdir [5,16].

2.7. Ridge Regresyon

Daha önceki başlıklarda çoklu doğrusal bağlantı probleminden ve sonuçları ile ilgili bilgiler verilmişti. Bunlardan en önemlisi, çoklu doğrusal bağlantı halinde regresyon katsayılarının varyans ve kovaryanslarının artmasıdır. Başka bir değişle çoklu doğrusal bağlantı halinde önemli değişkenlere ait regresyon katsayılarının standart hataları büyüyecek ve bunun sonucunda da bu değişkenlerin regresyon katsayılarının kısmi t testleri anlamsız sonuçlar verecektir. Çoklu doğrusal bağlantı halinde herhangi bir bağımsız değişken veya birime ait veriler modelden çıkartıldığında veya modele sokulduğunda kısmi regresyon katsayılarında çok önemli değişiklikler olacaktır. Ayrıca çoklu doğrusal bağlantı halinde kısmi regresyon katsayılarının işaretleri de teoriden veya beklenenden farklı olabilecektir. Sonuç olarak çoklu doğrusal bağlantılı verilerle hesaplanan standartlaştırılmış regresyon katsayıları durağanlığını ve kararlılığını kaybetmektedir [5].

Çoklu bağlantının bu etkilerini azaltabilmek için kullanılan yanlı tahmin yöntemlerinden biri de **Ridge Regresyon** yöntemidir. Ridge regresyon yöntemi Hoerl ve Kennard tarafından, çoklu doğrusal bağlantı durumunda EKK yönteminin yetersiz kalması nedeniyle geliştirilen bir yöntemdir.

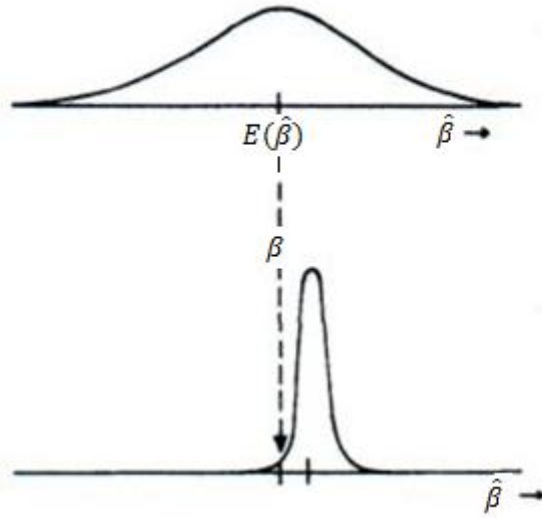
Hoerl ve Kennard (1970 a.b)'de Ridge regresyonu ařağıdaki üç amaç için önermişlerdir:

1. Kuvvetli çoklu doğrusal bağlantının varlığı durumunda, katsayılarda meydana gelen kararsızlıkların grafik üzerinde gösterilmesinde,
2. Çoklu doğrusal regresyon modelinde bağımsız değişkenlerin birbirleriyle ilişkili oldukları durumlarda EKK tahmininden daha küçük varyanslı tahminlerin elde edilmesinde,
3. Modeldeki gereksiz değişkenlerin çıkartılmasında kullanılabilir.[12,13]

Ridge regresyonun, yanlı regresyon yöntemi olmasına karşın EKK yöntemine göre iki önemli etkisi vardır.

1. Bağımsız değişkenlerdeki çoklu doğrusal bağlantıyı gidermek
2. Regresyonda yanlılık karesiyle varyansı değiştirerek hata kareler ortalamasını azaltmak [15]

Bu açıdan bakıldığında ridge tahminleri yanlı olmasına karşın EKK tahminlerinden daha kararlı olabilmektedirler. Bu yaklaşımla ilgili istatistiksel teori henüz tamamlanmış değildir. Böyle olmasına rağmen varyansta sağladığı azalıştan dolayı dikkate değer bir ilgi toplamış ve çok çeşitli alanlarda uygulanabilmiştir. Genelde varyans ve kovaryans matrisinin köşegen değerlerine küçük bir yanlılık sabiti k ilave etmenin dışında, Ridge regresyon ve EKK yöntemlerinin işleyişi aynıdır. Diğer bir değişle Ridge regresyon ile bir taraftan tahminlerin varyansları azaltılmakta, diğer taraftan ise bu katsayı k oranında yanlı tahminler elde edilmektedir. Böylece iki durum söz konusudur. Yansız tahminlerle yüksek varyans, yanlı tahminlerle düşük varyanstır [5]. Şekil 2-2'de bir tahmin edicide varyans ve yanlılık sunulmuştur.



Yansız, büyük varyanslı

$$E(\hat{\beta}) = \beta$$

Yanlı, küçük varyanslı

$$E(\hat{\beta}) \neq \beta$$

Şekil 2-2: Bir tahmin edicide varyans ve yanlılık [19].

Değişkenler arasındaki ilişkiler, EKK katsayı kestirimlerinin varyanslarının artmasına neden olur. Modeldeki her bir terim için VIF değerleri, katsayıların varyansları üzerinde basit korelasyonların toplam etkisini gösterir. Çoklu bağlantı durumunda, korelasyon matrisinin ters matrisindeki diagonal elemanlar olan VIF'ler, her bir kestirimin diğer kestirimlerle olan çoklu korelasyonunda sonsuz hale gelir [19].

Bu durumda $\hat{\beta}$ kestirimleri en küçük varyanslı kestirimler olmaz, çünkü çoklu doğrusal bağlantı $\hat{\beta}$ ve β arasında sapmaya neden olur. L_1 , $\hat{\beta}$ 'nin β 'dan sapması olmak üzere yanlılığın karesi;

$$L_1^2 = (\hat{\beta} - \beta)'(\hat{\beta} - \beta) \quad (2.59)$$

olarak gösterilebilir. L_1^2 'nin beklenen değeri;

$$E(L_1^2) = izVar(\hat{\beta}) = \sigma^2 iz(X'X)^{-1} \quad (2.60)$$

olur. Burada iz , bir kare matrisin esas köşegen üzerindeki elemanlarının toplamıdır. $(X'X)$ matrisinin özdeğerleri $\lambda_{max} = \lambda_1 > \lambda_2 > \dots > \lambda_k = \lambda_{min} > 0$ ile gösterilir. O zaman L_1^2 'nin beklenen değeri;

$$E(L_1^2) = \sigma^2 \sum_{j=1}^k \lambda_j^{-1} \quad (2.61)$$

olarak bulunur [12].

Çoklu bağlantı halinde, önceki bölümlerde belirtildiği gibi bir veya daha fazla özdeğerlerin küçük değer alması $\hat{\beta}'$ 'nin β' 'dan sapmalarının beklenen değerinin büyümesine neden olacaktır. Çoklu doğrusal bağlantının $E(L_1^2)$ yani hata kareler toplamı üzerindeki olumsuz etkisini giderebilmek için Hoerl ve Kennard (1970 a,b)' de ridge kestiricisini aşağıdaki gibi tanımlamışlardır.

$$\hat{\beta}^*(k) = (X'X + kI)^{-1}X'Y \quad \text{veya} \quad \hat{\beta}^*(k) = (R + kI)^{-1}X'Y \quad k > 0 \quad (2.62)$$

Burada değişkenler standartlaştırıldığında R bağımsız değişkenler arasındaki korelasyon matrisini gösterdiğinden $X'X = R'$ 'dir. Ridge tahmin edicisi, k yanlılık parametresine bağlıdır ve k 'nın seçimi tahmin edicinin performansını etkiler. k , 0 ile 1 arasında değerler alabilmektedir. $k=0$ için ridge çözümünün EKK çözümüne eşdeğer olduğu açıktır. Genel olarak regresyon problemi için k 'nın optimal bir değeri vardır. Bu optimal k değeri, EKK yönteminden daha küçük hata kareler ortalaması elde edilmesini sağlayan değerdir [5,12,15].

2.7.1. Ridge Kestiricisinin Özellikleri

1-EKK tahmin edicisinin doğrusal bir dönüşümü ridge tahmin edicisidir:

Şöyle ki;

$$\hat{\beta}^* = (X'X + kI)^{-1}X'Y \quad \text{'den}$$

$$\hat{\beta}^* = (X'X + kI)^{-1}X'X\hat{\beta} \quad (2.63)$$

yazılabilir. Buradan

$$Z = (X'X + kI)^{-1}X'X \quad (2.64)$$

olarak alınırsa,

$$\hat{\beta}^* = Z\hat{\beta} \quad (2.65)$$

olur [18].

2-Ridge kestiricisi yanlıdır:

$D = X'X$ matrisinin özvektörlerinin ortogonal matrisi

$\Lambda = X'X$ matrisinin özdeğerlerinin diagonal matrisi olmak üzere,

$$(X'X) = D' \Lambda D = Z\hat{\beta} \quad (2.66)$$

dönüşümü yapılırsa $i = 1, 2, \dots, n$ için,

$$Z = (X'X + kI)^{-1}X'X = D' \text{diag}(\delta_i)D \quad (2.67)$$

olur. Burada $\delta_i = \frac{\lambda_i}{(\lambda_i + k)^{-1}}$ dir.

Buradan,

$$\hat{\beta}^*(k) = D' \text{diag}(\delta_i)D\hat{\beta} = Z\hat{\beta} \quad (2.68)$$

olur.

$k > 0$ olduğundan

$$E(\hat{\beta}^*) \neq \hat{\beta} \quad (2.69)$$

olur ve $\hat{\beta}^*(k)$ yanlı bir kestiricidir. $k = 0$ olduğunda ridge kestiricisinin EKK kestiricisine eşit olduğu daha önce de belirtilmişti [31].

3- Ridge kestiricisinin hata kareler ortalaması

(2.59) ve (2.61) nolu ifadeler ridge kestiricisine uygulanırsa

$$L_1^2(k) = (\hat{\beta}^* - \beta)'(\hat{\beta}^* - \beta) \quad (2.70)$$

$$E(L_1^2(k)) = E[(\hat{\beta}^* - \beta)'(\hat{\beta}^* - \beta)] \quad (2.71)$$

$$= E[(\hat{\beta} - \beta)'Z'Z(\hat{\beta} - \beta)] + (Z\beta - \beta)'(Z\beta - \beta) \quad (2.71a)$$

$$= \sigma^2 \text{tr}(X'X)^{-1}Z'Z + \beta'(Z - I)'(Z - I)\beta \quad (2.71b)$$

$$= \sigma^2 [\text{İz}(X'X + kI)^{-1} - k \text{İz}(X'X + kI)^{-2}] + k^2 \beta' (X'X + kI)^{-2} \beta \quad (2.71c)$$

$$= \sigma^2 \sum_{i=1}^n \lambda_i / (\lambda_i + k)^2 + k^2 \beta' (X'X + kI)^{-2} \beta \quad (2.71d)$$

$$= \gamma_1(k) + \gamma_2(k) \quad (2.72)$$

olduğu Hoerl ve Kennard tarafından (1970a)'da gösterilmiştir. Burada $\gamma_1(k)$ $\hat{\beta}^*$ parametrelerinin varyans toplamını, $\gamma_2(k)$ ise yanın karesini göstermektedir. $\gamma_1(k)$ ve $\gamma_2(k)$ şöyle de ifade edilebilir. Birinci ifade $\gamma_1(k)$, parametre tahminlerinin toplam varyansıdır. İkinci ifade ise $\gamma_2(k)$ ise $Z\beta$ 'dan β ' ya olan uzaklığın karesidir. $k = 0$ olduğunda $\gamma_2(k) = 0$ olacaktır. Çünkü bu durumda $Z = I$ 'dir. Bu yüzden $\gamma_2(k)$, $\hat{\beta}^*$ 'ın $\hat{\beta}$ 'dan daha fazla kullanıldığı yan karesini ihtiva edecektir [12].

Ridge kestiricisinin hata kareler ortalaması şu şekilde de ifade edilebilir [19,20].

$$HKO(\hat{\beta}^*) = E(\hat{\beta}^* - \beta)^2 = V(\hat{\beta}^*) + [E(\hat{\beta}^*) - \beta]^2 \quad (2.73)$$

$$HKO(\hat{\beta}^*) = Var(\hat{\beta}^*) + (bias \hat{\beta}^*)^2 \quad (2.74)$$

4-Ridge kestiricisinin hata kareler ortalaması minumumdur:

Hoerl ve Kennard, daha önceki başlıklarda gösterilen (2.60) ve (2.72) denklemleri arasındaki ilişkiyi Varlık Teoremi olarak adlandırıp şöyle ifade etmişlerdir:

“ $E(L_1^2(k)) < E(L_1^2)$ olacak şekilde bir $k > 0$ bulmak her zaman mümkündür.” [12].

Bu eşitsizliğe göre çoklu bağlantı durumunda Ridge regresyon yöntemiyle, EKK yöntemine göre küçük hata kareler ortalamasına sahip olan kestirimler yapmak mümkündür.

5-Ridge kestiricisinin varyans ve yanlılığı:

Daha önceki başlıklarda $\gamma_1(k)$ 'nin katsayı kestirimlerinin varyanslarının toplamı olduğunu açıklamıştık. Çoklu bağlantı durumunda regresyon modeli kurulurken Ridge regresyon kullanılıyor ise, ridge kestiricisinin varyansının nasıl bir değişme gösterdiğinin bilinmesi gerekir. Y tesadüfi değişkenimiz için

$$\hat{\beta}^* = Z\hat{\beta}$$

$$= Z(X'X)^{-1}X'Y \quad \text{yazılabilir}$$

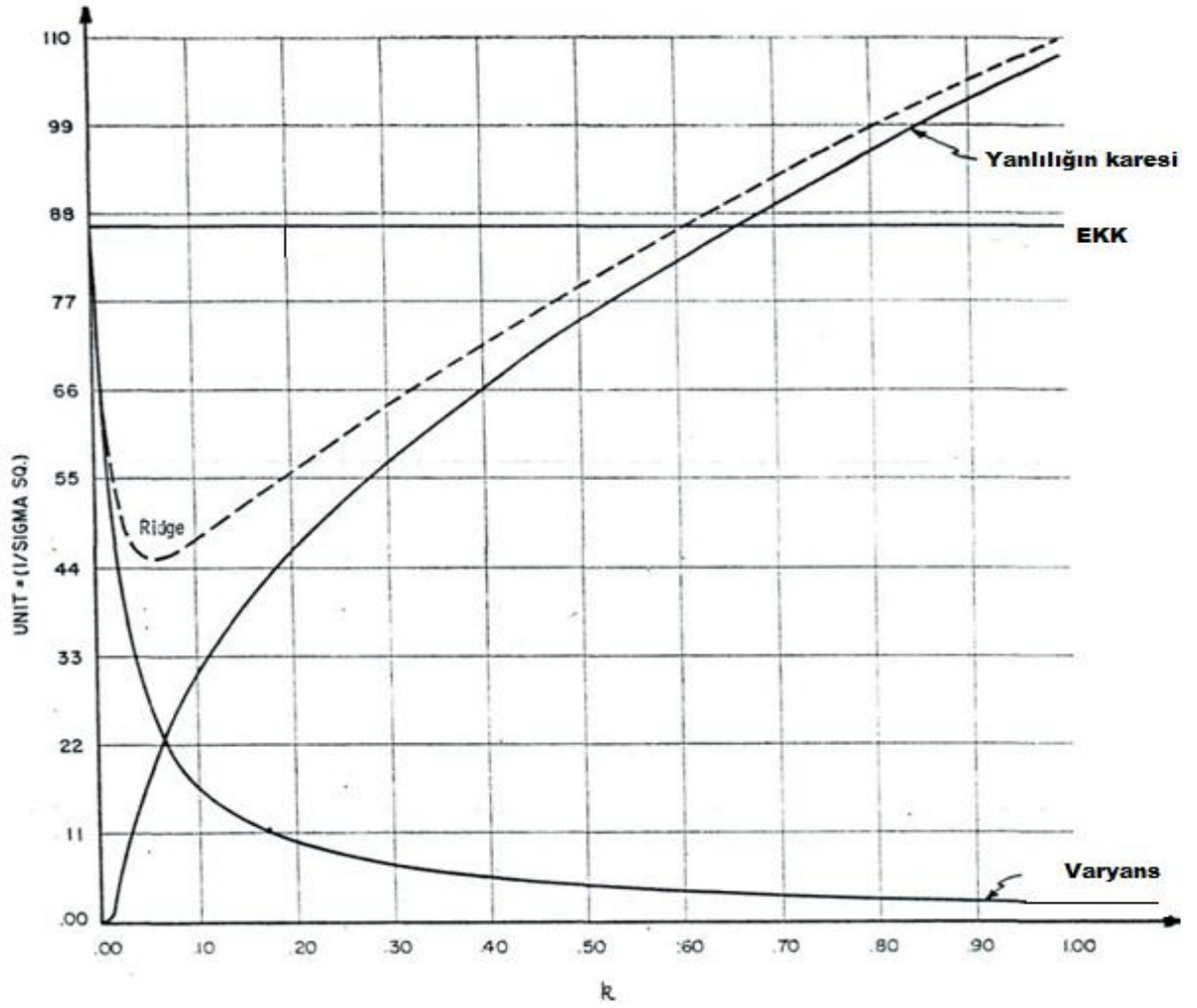
$$Var(\hat{\beta}^*) = Var(Z(X'X)^{-1}X'Y)$$

$$= Z(X'X)^{-1}X'Var(Y)X(X'X)^{-1}Z'$$

$$= \sigma^2 Z(X'X)^{-1}X'X(X'X)^{-1}Z'$$

$$= \sigma^2 Z(X'X)^{-1}Z' \quad \text{olur.} \quad (2.75)$$

Yani tüm $\hat{\beta}^*$ 'ların varyanslarının toplamı matrisinin köşegen elemanlarının toplamıdır [12]. Yanlılık ise Sayfa-34, Şekil 2-2'den de anlaşılacağı gibi $E(\hat{\beta}^*) - \hat{\beta}$ 'dir. Yani daha önceki başlıklarda da izah edildiği üzere $\gamma_2(k)$, $\hat{\beta}^*$ 'ın $\hat{\beta}$ 'dan daha fazla kullanıldığı yan karesidir. Aşağıdaki şekilde ridge parametresi varyans ve yanlılık arasındaki ilişki daha net görülmektedir.



Şekil 2-3: Ridge kestiricisinde varyans ve yanlılık [12].

Şekil 2-3’de Ridge kestiricisinde varyans ve yanlılık verilmiştir. Şekilden de anlaşıldığı gibi $k > 0$ olduğunda ridge kestiricisindeki yanlılık k ile birlikte artmaktadır. Buna karşılık k artarken varyans azalmaktadır. Bu durumla alakalı olarak Hoerl ve Kennard 1970a ‘da aşağıdaki teoremleri önermişlerdir;

Teorem 2-1: $\gamma_1(k)$ toplam varyans, k ’nın sürekli düzgün azalan bir fonksiyonudur.

Teorem 2-2: $\gamma_2(k)$ yanlılığın karesi, k ’nın sürekli düzgün artan bir fonksiyonudur [12].

Ridge regresyon yönteminde yanlılığın karesindeki artıştan daha büyük varyans azalmasını sağlayacak bir k değerinin belirlenmesi halinde Ridge kestiricisindeki HKO,

EKK kestiricisindeki HKO'dan daha küçük bir değer alacaktır. Bu nedenle k 'nın belirlenmesi son derece önemlidir.

2.7.2. Ridge Parametresinin Seçimi

Çoklu doğrusal bağlantı durumunda Ridge regresyon yöntemi ile yapılacak kestirimlerin kararlılığı, ridge parametresi k için optimum değer belirlenmesine bağlıdır. Uygun k değerinin belirlenmesi durumunda ridge kestiricilerinin HKO'sı EKK kestiricisinin HKO'dan daha küçük olacaktır. Ancak en uygun k değeri, k 'nın bilinmeyen β parametre vektörüne ve varyansa (σ^2) bağlı olması nedeniyle kesin olarak belirlenemez [8].

Şimdiye kadar yapılan çalışmalarda ridge parametresi k 'nın belirlenmesi için kesin bir ifade ortaya konulmamasına rağmen k 'nın kestirimi için pek çok öneri geliştirilmiştir. Bunlardan bazıları şunlardır:

1. Hoerl ve Kennard (1970b)'de k 'nın seçimi için Ridge İz'ini (Ridge Trace) önermişlerdir [13]
2. Marquardt ve Snee, VIF'in 1 ile 10 değerleri arasında olduğunda k 'nın seçilebileceğini belirtmişlerdir [19]
3. Anderson 1998'de bağımsız değişkenler için birlikte 1'e yaklaşan VIF değerlerini sağlayan k sabitinin alınabileceğini belirtmiştir.
4. Hoerl, Kennard ve Baldwin (1975)'de;

$$k^* = \frac{k\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}} \quad (2.76)$$

eşitliğini önermişlerdir. Burada k bağımsız değişken sayısını gösterdiği için ridge parametresi k^* ile gösterilmiştir.

5. McDonald ve Galarneau (1975)'de

$$\hat{\beta}^*\hat{\beta}^* = \hat{\beta}'\hat{\beta} - \hat{\sigma}^2 \sum_{j=1}^k \lambda_j^{-1} \quad (2.77)$$

eşitliğini sağlayan k değerinin uygun olacağını belirtmişlerdir [20].

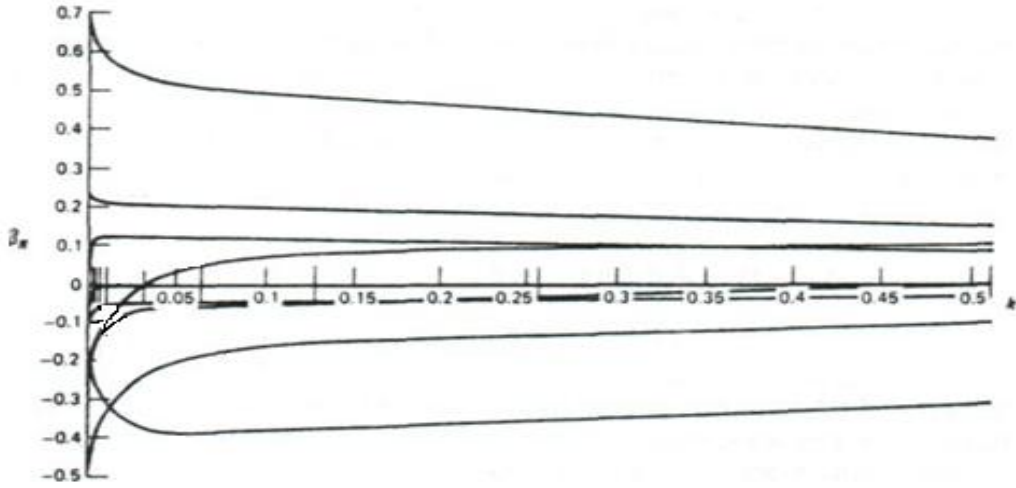
Bunların dışında çoklu bağlantı çok güçlü olduğunda EKK kestirimleri elde edilemeyeceğinden (2.77)'deki formül yerine;

$$k^* = \frac{k \hat{\sigma}^2}{\sum_{i=1}^k \hat{\alpha}_i^2} \quad (2.78)$$

önerilmiştir. Burada $\hat{\alpha}$ temel bileşenler kestiricisidir [3,5,15].

2.7.3. Ridge İzi

Çoklu doğrusal regresyon problemlerinde çoklu bağlantı olduğunda katsayı tahminleri duyarlıdır. Yani veri kümesine birkaç gözlemin ilave edilmesiyle bu tahmin edicilerde değişikliklerin olduğu görülür. Bu nedenle hassaslık analizi Ridge regresyonun amaçlarından biridir. Ridge regresyonun en büyük avantajlarından biri üzerinde çalışılan veri için hassas olan katsayıları görmeye yardımcı olabilen ve Ridge İzi (Ridge Trace) olarak adlandırılan grafiksel gösteriminin olmasıdır. Hoerl ve Kennard'ın (1970b)'de önerdiği ve Ridge regresyonun grafiksel bir gösterimi olan ridge izi, regresyon katsayıları $\hat{\beta}^*$ lar düşey ekseninde k değerleri yatay ekseninde olacak şekilde iki boyutlu uzayda grafik elde edilmesidir. Ridge izi araştırmacıya hangi katsayıların verilere duyarlı olduğu konusunda yardımcı olur. Ridge izi grafiklerinde yanlı regresyon katsayıları, k 'nın bir fonksiyonu olarak gösterilmektedir [13]. Şekil 2-4'de Ridge izi gösterilmiştir.



Şekil 2-4: Ridge izi örnek şekil [21].

Genelde standartlaştırılmış regresyon katsayıları ilk önce küçük k değerleriyle çok anormal bir biçimde değişmekte ve daha sonra durağanlaşmaktadır. Regresyon katsayılarının durağanlaştığı bu bölgede olası en küçük k değeri optimum k değeri olarak seçilmektedir. Ayrıca optimum k değerinin seçiminde kullanılan diğer kriterler arasında katsayıların kuramsal beklentilere uygunluğunu, durağanlığını, makul büyüklüğünü, kabul edilebilir hata kareler toplamını ve minimum VIF'leri sağlayan k sabiti yaklaşımlarını saymak mümkündür. NCSS gibi paket programlarda k 'nın tespiti için k analiz tabloları verilmektedir. Özellikle k karşılaştırmalı VIF tablosundan faydalandığında tüm bağımsız değişkenler için VIF'in 10'dan küçük olduğu ilk değer veya tüm değişkenlerin VIF değerlerinin birlikte 1'e yaklaştıkları ilk değer alınabilir. Uygulamada tüm kriterlerin birlikte değerlendirilmesi en sağlıklı yaklaşım olarak kabul edilmektedir. Sonuç olarak k 'nın seçiminde dikkat edilmesi gereken hususları maddelersek:

1. Seçilen k 'da sistem sabitleşmeli ve ortogonal bir sistem özelliklerini taşımalıdır.
2. Regresyon katsayıları, bağımlı değişken üzerindeki kısmi değişimi temsil ettikleri için mutlak değerce makul büyüklüğe sahip olmalı.
3. Beklentiye göre $k=0$ iken yanlış işarete sahip olduğu düşünülen katsayılar, uygun k için işaret değiştirmiş olmalı.
4. Artık kareler toplamı minimum değerinden fazla uzaklaşmamalıdır [5,15,20,31]

3. UYGULAMA

3.1. Gereç ve Yöntem

3.1.1. Araştırmanın Tasarımı ve Modeli

Bu araştırma hipertansif hastalık hikâyesine sahip olan hastaların sistolik kan basıncı üzerinde etkili olabileceği düşünülen değişkenlerin etkisinin değerlendirilmesine yönelik kesitsel ve analitik bir araştırmadır. Verilerin toplanması ileriye yönelik olarak gerçekleştirilmiştir. Bu bakımdan karşılaştırma amacıyla herhangi bir kontrol grubu oluşturulmamış, bağımlı-ilişkili olabilecek değişkenler arasında çoklu doğrusal bağlantılar aranmıştır. Çoklu bağlantının giderilebilmesi için Ridge regresyon yöntemi kullanılmıştır. Ayrıca tahmin sonuçları, Ridge regresyon ve en küçük kareler regresyonu açısından karşılaştırılmıştır. Çalışmada sistolik kan basıncını tahmin etmek için kullanılan değişkenler ve ölçü birimleri aşağıdaki tabloda özetlenmektedir:

Tablo 3.1: Modeldeki değişkenler ve ölçü birimleri

Değişkenin Türü	Değişkenler	Ölçü Birimi
Bağımsız	Yaş	Yıl
	Boy	m
	Karın Çevresi	m
	Glikoz	mg/dl
	HDL	mg/dl
	LDL	mg/dl
	Üre	mg/dl
	Kreatinin	mg/dl
	Potasyum	mg/dl
	HGB	mg/dl
	HCT	mg/dl
	Diastolik kan basıncı	mm Hg
Bağımlı	Sistolik kan basıncı	mm Hg

3.1.2. Araştırmanın Örneklem Büyüklüğü

Araştırmanın örneklemini; Elazığ Fırat Üniversitesi Fırat Tıp Merkezi Kardiyoloji Polikliniği'ne aşağıda belirtilen zaman aralığında gelen ve hipertansif hastalık hikayesine sahip olan hastalar oluşturmaktadır.

Genel olarak çoklu regresyon analizi için örneklem büyüklüğü hesabı aşağıda özetlenmiştir:

Hair ve arkadaşları (1998) çoklu regresyon analizi için örneklem büyüklüğünü şöyle belirlemişlerdir:

%80 güç ve $\alpha=0.05$ seçilirse;

- R^2 'yi 0.23 ($r=0.48$) ya da daha büyük bulabilmek için örneklem büyüklüğü 50 olmalıdır.
- R^2 'yi 0.12 ($r=0.35$) ya da daha büyük bulabilmek için örneklem büyüklüğü 100 olmalıdır.
- Genel kural olarak, her bağımsız değişken için birey sayısı en az 5 olmalıdır. Fakat arzulanan sayı her bağımsız değişken için 15-20 bireydir.

Green'in (1991) çoklu regresyon analizi için örneklem büyüklüğü ile ilgili görüşü şöyledir:

- Çoklu korelasyon katsayısını test etmek için: $n > 50 + 8m$ (m: bağımsız değişken sayısı)
- Her bir bağımsız değişken ile bağımlı değişken arasındaki ilişkiyi test etmek için: $n > 104 + m$ (m: bağımsız değişken sayısı)

Harris'in (1985) çoklu regresyon analizi için örneklem büyüklüğü ile ilgili görüşü şu şekildedir:

- Eşitlikte 5 ya da daha az bağımsız değişken varsa: $n > 50 + m$ (m: bağımsız değişken sayısı)
- Eşitlikte 6 ya da daha fazla bağımsız değişken varsa her bağımsız değişken için 10 birey gerekir. $n > 104 + m$ (m: bağımsız değişken sayısı) [29].

Tüm bu görüşler ve imkanlar göz önüne alındığında örneklem büyüklüğü (n)=128 alınabilir.

3.1.3. Veri Toplama Yöntemi ve Süresi

Araştırmaya konu olan sistolik kan basıncını etkileyebilecek değişkenler göz önüne alındığında tüm değişkenlere göre kayıtlı veri bulmanın zorluğu nedeniyle 1 Ekim 2009-22 Aralık 2009 tarihleri arasında Fırat Tıp Merkezi kardiyoloji polikliniğine her gün gidilmiş ve hipertansif hastaların fiziksel, ölçümsel ve biyokimyasal değerlerine ulaşarak veriler toplanmıştır.

3.1.4. Araştırmada Karşılaşılan Zorluklar

Araştırma boyunca karşılaşılan en büyük zorluk belirtilen örnek büyüklüğüne ulaşmada sürenin sınırlı olmasıydı. Bunun dışında modeli ve modeldeki değişkenleri belirlemek için uzman hekime ulaşmada çeşitli zorluklarla karşılaşmıştır. Model belirlenirken bağımlı değişkene etkili olabileceği düşünülen bazı değişkenler rutin tahlillerin dışına çıktığı için, bu tahlillerin yapılması ek maliyet, insan gücü ile araç-gereç gerektirdiği için hastane tarafından kabul edilmemiştir. Bu nedenle bazı değişkenler zorunlu olarak modele alınamamıştır. Ayrıca hipertansif hastaların anlık bilgilerini almada bazen yeterli araç-gereç ve ekipmanların eksikliği, gelen hastalarda hipertansif hastalığın primer mi yoksa sekonder mi olduğu noktasında ayrımın tam olarak yapılamaması, antihipertansif ilaç alarak gelen hastaların tansiyon ölçümlerinde normal kan basıncına sahip çıkmaları, hastaların biyokimyasal sonuçlarının alımında hastane otomasyon sisteminde bazı zorluklarla karşılaşılması, eksik tahlil istemi nedeniyle veri kaybı yaşanması diğer zorluklar olarak değerlendirilebilir.

3.2. Bulgular

Elazığ Fırat Üniversitesi Fırat Tıp Merkezi kardiyoloji polikliniğine gelen 128 hastadan elde edilen veriler ekler bölümünde **Ek-A**'da gösterildiği gibidir. Bu veri setine ilişkin tüm gerekli analizler NCSS-2007 paket programı yardımıyla yapılmıştır [11]. Elde edilen sonuçlar aşağıdaki tablolarda verilmiş ve yorumları her tablonun altında özetlenmiştir.

Tablo 3.2. Tanımlayıcı istatistikler

Değişkenler	n	$\bar{X}$	s (Std.Sapma)	Min	Max
Yaş	128	57,984	10,934	20	80
Boy	128	1,640	0,101	1,43	1,9
Karın çevresi	128	1,025	0,109	0,75	1,3
Glikoz	128	111,359	44,911	15	311
HDL	128	50,234	11,054	29	84
LDL	128	137,484	37,666	58	279
Üre	128	42,171	33,030	12	278
Kreatinin	128	1,292	1,381	0,6	13,9
Potasyum	128	4,499	0,438	2,9	5,7
HGB	128	14,202	1,568	9,04	17,5
HCT	128	41,105	4,419	25,4	50,1
Diastolik	128	86,601	14,071	50	130
Sistolik	128	140,390	20,025	100	190

Tablo 3-2'de kayıp olmayan her bir değişken için tanımlayıcı istatistikler verilmiştir. Bu tablo özellikle tüm değişkenlerin aldığı değerleri kontrol etmek için kullanılır. Tablodan 128 adet veri için her bir değişkenin ortalaması, standart sapması, minimum ve maksimum değerleri görülmektedir.

Tablo 3-3. EKK çoklu regresyon sonuçları

Bağımsız Değişkenler	Regresyon Katsayıları	Standart Hata	t-testi	p
Sabit	17,913	31,489	0,569	0,570
Boy	-6,902	13,595	-0,508	0,612
Diastolik	1,041	0,086	12,022	<0.001
Glikoz	0,002	0,027	0,076	0,939
HCT	0,766	1,331	0,575	0,566
HDL	-0,106	0,117	-0,909	0,362
HGB	-1,935	3,722	-0,520	0,604
Karın çevresi	29,826	11,443	2,606	0,010
Kreatinin	3,004	1,854	1,621	0,107
LDL	0,019	0,033	0,602	0,548
Potasyum	0,378	2,961	0,128	0,898
Yaş	0,144	0,118	1,215	0,226
Üre	-0,059	0,079	-0,755	0,451

Tablo 3-3a. EKK için ANOVA tablosu

Değ.Kay.	sd	HKT	HKO	F	p
Sabit	1	2522820	2522820		
Model	12	30588,01	2549	14,41	<0.0001
Hata	115	20342,46	176,891		
Toplam	127	50930,47	401,0273		

Tablo 3-3’de EKK analizi ile elde edilen regresyon katsayıları, standart hataları ve kısmi regresyon katsayıları için t-testi sonuçları verilmiştir. Tablo 3-3a’da ise modele ilişkin ANOVA varyans analiz tablosu verilmiştir. Tablo 3-3a’ya göre $p < 0.05$ olduğu için model anlamlıdır. Ancak Tablo 3-3’e bakıldığında t-testi sonuçlarına göre katsayıların büyük çoğunluğu anlamsız çıkmıştır. F testi anlamlı iken t-testi sonuçlarının anlamsız çıkması çoklu bağlantı probleminin bir sonucudur. Bu nedenle bağımsız değişkenler arasında çoklu bağlantı olup olmadığının incelenmesine karar verilmiştir.

Tablo 3-4. Korelasyon matrisi

	Yaş	Boy	Karın çevresi	Glikoz	HDL	LDL	Üre	Kreatinin	Potasyum	HGB	HCT	Diastolik	Sistolik
Yaş	1												
Boy	-0,030	1											
Kar.çev	-0,056	0,022	1										
Glikoz	0,0207	0,058	0,104	1									
HDL	-0,106	-0,305	0,013	-0,093	1								
LDL	0,034	-0,177	0,140	-0,077	0,154	1							
Üre	-0,000	0,114	-0,183	0,092	-0,126	-0,148	1						
Kreatinin	-0,165	0,119	-0,089	0,078	-0,111	-0,161	0,856	1					
Potasyum	0,027	0,151	-0,088	0,120	-0,015	-0,009	0,116	-0,053	1				
HGB	0,073	0,268	0,163	-0,135	-0,057	0,116	-0,387	-0,332	-0,063	1			
HCT	0,111	0,234	0,182	-0,102	-0,015	0,112	-0,425	-0,375	-0,036	0,975	1		
Diastolik	-0,053	-0,123	0,040	0,062	-0,048	0,046	-0,103	-0,069	-0,114	0,022	0,042	1	
Sistolik	0,012	-0,098	0,196	0,081	-0,088	0,076	-0,040	0,027	-0,110	0,028	0,050	0,741	1

Tablo3-4’de her bir bağımsız değişkenin birbirleriyle ve bağımlı değişkenle aralarındaki basit korelasyon katsayılarından oluşan korelasyon matrisi verilmektedir. Çoklu bağlantıyı belirlemek için kullanılan yöntemlerden biri de korelasyon matrisinin incelenmesidir. Genel olarak X_i ve X_j açıklayıcı değişkenler lineer bağımlı olduklarında $|r_{ij}| \rightarrow 1$ olur. Tablo 3-4’de, HGB-HCT değişkenleri arasındaki korelasyon 0.975, Üre-Kreatinin değişkenleri arasındaki korelasyon 0.856 olduğu için çoklu bağlantı probleminin varlığından bahsedilebilir. Ancak basit korelasyonlar çoklu bağlantının belirlenmesinde tek başına yeterli değildir. Bu nedenle çoklu bağlantıyı belirleyen diğer durumların da incelenmesine karar verilmiştir.

Tablo 3-5. EKK çoklu bağlantının saptanması

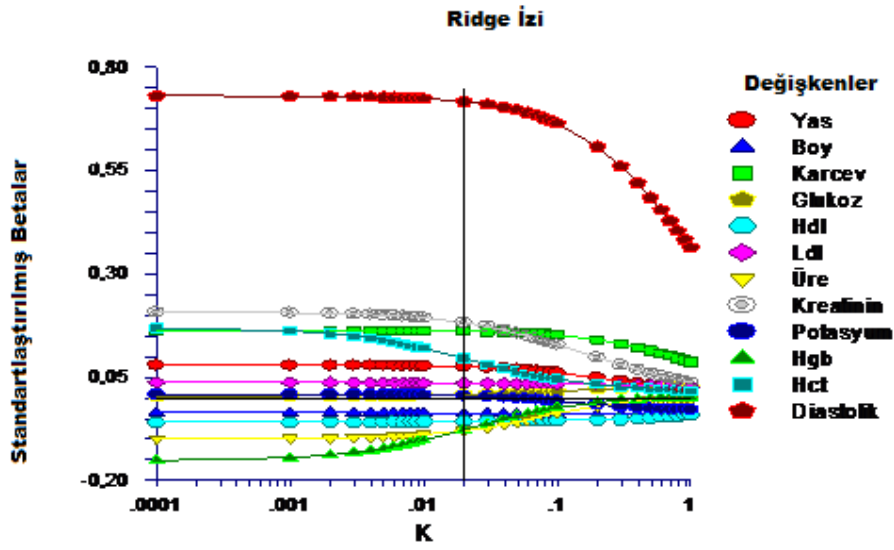
Bağımsız Değ.	VIF	Diğer Değişkenlerle olan R^2	Tolerans
Yaş	1,21	0,17	0,83
Boy	1,36	0,26	0,74
Karın çevresi	1,12	0,11	0,89
Glikoz	1,10	0,09	0,91
HDL	1,21	0,17	0,83
LDL	1,11	0,10	0,90
Üre	4,94	0,80	0,20
Kreatinin	4,71	0,79	0,21
Potasyum	1,21	0,17	0,83
HGB	24,49	0,96	0,04
HCT	24,88	0,96	0,04
Diastolik	1,07	0,06	0,94

Tablo 3-5’de bağımsız değişkenlere ait VIF, R^2 ve tolerans değerleri verilmiştir. Çoklu bağlantıyı belirleme yöntemlerinden biri olan VIF için; $VIF \geq 10$ olduğu zaman çoklu bağlantı problemi ortaya çıkmaktadır. Tablo 3-5’e göre HGB ve HCT değişkenlerine ait VIF değerlerinin 24,49 ve 24,88 olması çoklu bağlantının varlığına işaretler. Her bir değişkenin diğer değişkenlerle olan R^2 değerlerine bakıldığında HGB ve HCT için 0.96, Üre için 0.80 ve Kreatinin için 0.79 olan bu değer, bağımsız değişkenlerin birbirleriyle ilişkili olduğunu vermektedir. Ayrıca $1 - R^2$ olarak hesaplanan tolerans değeri için, çoklu bağlantı varlığında büyük R^2 değeri tolerans değerini küçültecektir. Yine HGB ve HCT değişkenlerinin olan 0.04 tolerans değerleri çoklu bağlantının bir göstergesidir.

Tablo 3-6. Korelasyonların özdeğerleri

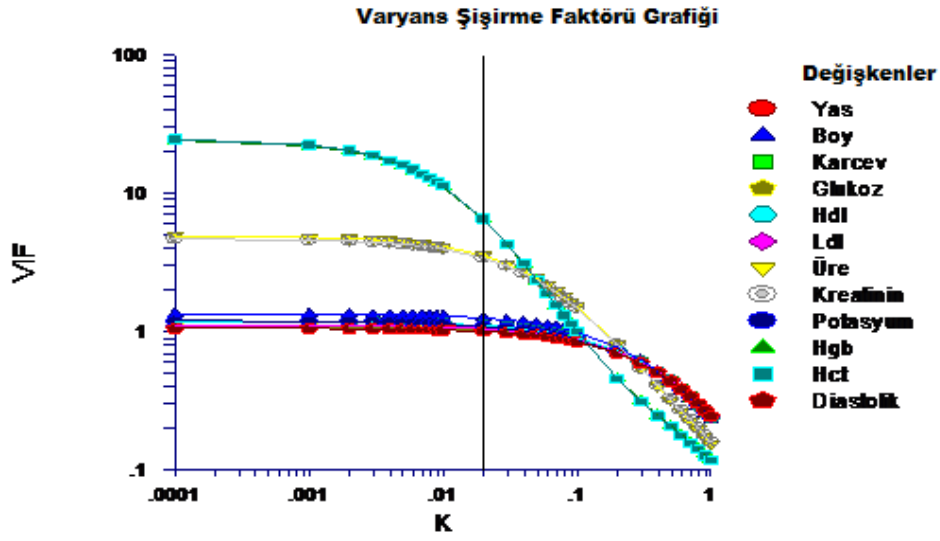
No	Özdeğer	Göreceli Yüzde	Birikimsel Yüzde	Koşul Sayısı
1	2,840	23,67	23,67	1
2	1,725	14,38	38,05	1,65
3	1,196	9,97	48,02	2,37
4	1,158	9,66	57,68	2,45
5	1,075	8,96	66,65	2,64
6	0,995	8,3	74,94	2,85
7	0,891	7,43	82,37	3,19
8	0,786	6,55	88,92	3,61
9	0,665	5,55	94,47	4,27
10	0,534	4,45	98,92	5,32
11	0,108	0,91	99,83	26,12
12	0,020	0,17	100	138,84

Tablo 3-6’da özdeğerler, göreceli yüzdeler, birikimli yüzdeler ve koşul sayısı sunulmuştur. Özdeğerler, $X'X$ korelasyon matrisinin özdeğerleridir ve koşul sayısı bu değerler arasından en büyük özdeğerin diğer özdeğerlere bölünmesi ile hesaplanmıştır. Bazı özdeğerlerin sıfıra yakın olması çoklu bağlantının göstergesidir. 12 nolu özdeğer 0,020462 olduğundan çoklu bağlantı vardır. Bunun yanı sıra Gujarati’nin (1995)’ de koşul sayısı için yaptığı genelleştirme gözönüne alınırsa, $K > 30$ olduğunda çoklu bağlantı vardır. Eğer $K > 100$ olursa çoklu bağlantı güçlüdür. 12 nolu koşul sayısı 138,84 olduğundan çoklu bağlantının güçlü olduğu belirlenmiştir.



Şekil 3-1. Ridge İzi

Şekil 3-1’de Ridge İzi gösterilmiştir. Ridge İzi dikey ekseninde standartlaştırılmış β katsayılarının, yatay ekseninde k ’nın artan değerlerinin bulunduğu ve her bir k ’ya karşılık gelen β değerlerinin izlerinin olduğu bir grafikdir. En soldaki eksen çizgisinde $k = 0$ için β ’nin EKK regresyon değerleri vardır. k arttıkça bu regresyon değerleri başlangıçta çok fazla değişim gösterir sonra gitgide durağanlaşmaya başlar. Yanlılık sabiti k ile yanlı standartlaştırılmış regresyon katsayıları arasındaki grafik incelendiğinde çok küçük ($k=0,02$) bir yanlılık sabitinden sonra regresyon katsayılarının daha durağan hale geldiği tespit edilmiştir.



Şekil 3-2. VIF grafiği

Şekil 3-2’de k ’nın VIF değerleri üzerindeki etkisi görülmektedir. Eğer uygun k değeri seçilmişse tüm VIF değerlerinin 10’dan küçük olması gerekir. Şekil 3-2’den de görüleceği gibi $k=0.02$ için tüm VIF değerleri 10 değerinin altındadır.

Tablo 3-7. Standartlaştırılmış Ridge Regresyon Katsayıları

k	Yaş	Boy	Kar.çev	Glikoz	HDL	LDL	Üre	Kreatinin	Potasyum	HGB	HCT	Diastolik
0,00	0,08	-0,04	0,16	0,01	-0,06	0,04	-0,09	0,20	0,01	-0,13	0,14	0,73
0,01	0,08	-0,04	0,16	0,01	-0,06	0,04	-0,09	0,19	0,01	-0,10	0,12	0,73
0,02	0,08	-0,04	0,16	0,01	-0,06	0,04	-0,08	0,18	0,01	-0,08	0,10	0,72
0,03	0,07	-0,04	0,16	0,01	-0,06	0,03	-0,07	0,17	0,00	-0,06	0,08	0,71
0,04	0,07	-0,04	0,16	0,01	-0,06	0,03	-0,06	0,16	0,00	-0,05	0,07	0,70
0,05	0,07	-0,04	0,16	0,01	-0,05	0,03	-0,06	0,16	0,00	-0,04	0,06	0,70
0,06	0,07	-0,04	0,16	0,01	-0,05	0,03	-0,05	0,15	0,00	-0,04	0,06	0,69
0,07	0,07	-0,04	0,16	0,01	-0,05	0,03	-0,05	0,14	0,00	-0,03	0,05	0,68
0,08	0,06	-0,04	0,16	0,01	-0,05	0,03	-0,05	0,14	0,00	-0,03	0,05	0,68
0,09	0,06	-0,04	0,15	0,02	-0,05	0,03	-0,04	0,13	-0,01	-0,03	0,05	0,67
0,10	0,06	-0,04	0,15	0,02	-0,05	0,03	-0,04	0,13	-0,01	-0,02	0,05	0,67
0,20	0,05	-0,04	0,14	0,02	-0,05	0,03	-0,02	0,10	-0,02	-0,01	0,03	0,61
0,30	0,04	-0,04	0,13	0,02	-0,05	0,03	-0,01	0,08	-0,02	-0,01	0,03	0,56
0,40	0,03	-0,04	0,12	0,03	-0,05	0,03	-0,01	0,07	-0,02	0,00	0,02	0,52
0,50	0,03	-0,04	0,11	0,03	-0,05	0,03	-0,01	0,06	-0,03	0,00	0,02	0,49
0,60	0,03	-0,04	0,11	0,03	-0,05	0,03	0,00	0,05	-0,03	0,00	0,02	0,46
0,70	0,02	-0,04	0,10	0,03	-0,04	0,03	0,00	0,05	-0,03	0,00	0,02	0,43
0,80	0,02	-0,04	0,10	0,03	-0,04	0,03	0,00	0,04	-0,03	0,00	0,02	0,40
0,90	0,02	-0,04	0,09	0,03	-0,04	0,03	0,00	0,04	-0,03	0,00	0,02	0,38
1,00	0,02	-0,03	0,09	0,03	-0,04	0,03	0,00	0,04	-0,03	0,00	0,02	0,36

Tablo 3-7’de $k = 0$ ’dan başlayarak bazı k değerleri için standartlaştırılmış ridge katsayıları verilmektedir. Bu tablo ridge izinde verilen değerlerin tablosudur. Bilindiği üzere çoklu bağlantı problemi EKK kestirimlerini işaret ve büyüklük olarak etkileyebilmektedir. Eğer böyle bir durum söz konusu ise bu problemi ortadan kaldıracabilecek bir k değerinin seçilmesi gerekmektedir. Tablo 3-7’den katsayıların işaret ve büyüklük bakımından kuramsal beklentilere cevap verip vermediklerine bakılabilir. Modeldeki bağımsız değişkenler göz önüne alındığında özellikle yaş, karın çevresi, diastolik kan basıncı gibi değişkenlerin sistolik kan basıncı ile pozitif ilişkili olması beklenmektedir. Bu beklentiler ışığında $k = 0,02$ uygun değer olarak belirlenmiştir.

Tablo 3-8. VIF değerleri

k	Yaş	Boy	Kar. çev	Glikoz	HDL	LDL	Üre	Kreatinin	Potasyum	HGB	HCT	Diastolik
0,00	1,17	1,34	1,11	1,08	1,18	1,10	4,60	4,41	1,18	17,25	17,51	1,05
0,01	1,14	1,31	1,09	1,06	1,15	1,08	4,17	4,01	1,15	11,25	11,41	1,04
0,02	1,09	1,26	1,06	1,03	1,11	1,05	3,59	3,46	1,10	6,52	6,61	1,01
0,03	1,05	1,22	1,03	1,01	1,08	1,03	3,13	3,02	1,06	4,31	4,37	0,99
0,04	1,01	1,19	1,01	0,99	1,05	1,01	2,76	2,67	1,03	3,10	3,14	0,97
0,05	0,99	1,15	0,98	0,96	1,03	0,98	2,46	2,38	1,00	2,37	2,39	0,95
0,06	0,96	1,12	0,96	0,94	1,00	0,96	2,21	2,14	0,97	1,89	1,90	0,93
0,07	0,93	1,09	0,94	0,92	0,98	0,94	2,00	1,94	0,95	1,55	1,57	0,91
0,08	0,91	1,06	0,92	0,90	0,95	0,92	1,82	1,77	0,92	1,31	1,32	0,89
0,09	0,89	1,03	0,90	0,89	0,93	0,90	1,67	1,62	0,90	1,14	1,14	0,87
0,10	0,87	1,00	0,88	0,87	0,91	0,88	1,53	1,49	0,88	1,00	1,00	0,86
0,20	0,71	0,78	0,72	0,72	0,74	0,73	0,82	0,81	0,71	0,46	0,46	0,71
0,30	0,59	0,63	0,61	0,60	0,61	0,61	0,55	0,55	0,60	0,32	0,32	0,60
0,40	0,51	0,52	0,52	0,51	0,52	0,52	0,41	0,41	0,51	0,25	0,25	0,51
0,50	0,44	0,44	0,45	0,44	0,44	0,45	0,32	0,33	0,44	0,21	0,21	0,44
0,60	0,38	0,38	0,39	0,39	0,39	0,39	0,27	0,27	0,38	0,18	0,18	0,39
0,70	0,34	0,33	0,34	0,34	0,34	0,34	0,23	0,23	0,34	0,16	0,16	0,34
0,80	0,30	0,29	0,30	0,30	0,30	0,30	0,20	0,20	0,30	0,14	0,14	0,31
0,90	0,27	0,26	0,27	0,27	0,27	0,27	0,17	0,18	0,27	0,13	0,13	0,27
1,00	0,24	0,23	0,24	0,25	0,24	0,24	0,16	0,16	0,24	0,12	0,12	0,25

Tablo 3-8’de $k = 0$ ’ dan başlayarak bazı k ’lar için VIF değerleri verilmektedir. Bu tablo VIF grafiğinde gösterilen değerlerin tablosudur. Tüm VIF değerlerinin 10’den küçük olduğu ilk değer k değeri olarak alınabilir. Buradan tüm değişkenlerin VIF değerlerinin 10 değerinin altında olduğu ilk değer $k = 0,02$ ’dir.

Tablo 3-9. k Analiz tablosu

k	R^2	Sigma	B'B	Ortalama VIF	Enbüyük VIF
0,00	0,60	13,34	0,66	4,42	17,51
0,01	0,59	13,41	0,63	3,32	11,41
0,02	0,59	13,51	0,61	2,41	6,61
0,03	0,58	13,61	0,59	1,94	4,37
0,04	0,58	13,70	0,57	1,66	3,14
0,05	0,57	13,79	0,56	1,47	2,46
0,06	0,56	13,88	0,54	1,33	2,21
0,07	0,56	13,97	0,53	1,23	2,00
0,08	0,55	14,05	0,52	1,14	1,82
0,09	0,55	14,13	0,51	1,07	1,67
0,10	0,54	14,21	0,50	1,01	1,53
0,20	0,50	14,89	0,41	0,70	0,82
0,30	0,46	15,44	0,35	0,55	0,63
0,40	0,43	15,90	0,30	0,45	0,52
0,50	0,40	16,28	0,26	0,38	0,45
0,60	0,38	16,61	0,23	0,33	0,39
0,70	0,36	16,89	0,20	0,29	0,34
0,80	0,34	17,14	0,18	0,26	0,31
0,90	0,32	17,36	0,16	0,23	0,27
1,00	0,30	17,56	0,15	0,21	0,25

Tablo3-9, k 'nın seçimine bağlı olarak bazı istatistiklerde meydana gelen değişimleri göstermektedir. R^2 değerini maksimize eden çözüm EKK'dır. Bu nedenle $k = 0$ için R^2 en büyüktür. Seçilen k değeri için bu değerden çok sapma olmaması gerekmektedir. Sigma, hata kareler ortalamasının kareköküdür. Bu değeri minimize eden çözüm EKK'dır. Bu nedenle seçilen k değeri için bu değerden çok sapma olmaması gerekmektedir. $B'B$, standartlaştırılmış regresyon katsayılarının kareler toplamıdır. k 'nın seçimine göre bu değerlerin durağanlaşması gerekmektedir. Ortalama VIF, her k değerine karşılık gelen VIF değerlerinin ortalamasını verir. Enbüyük VIF, her k değerine karşılık gelen VIF değerlerinin en büyüğünü verir. Uygun k değerinde tüm VIF değerlerinin 10'dan küçük olması gerektiği için, k değeri bu kısımdan rahatlıkla belirlenebilir. $k=0,02$ 'den itibaren Enbüyük VIF 6.61 olduğu için $k \geq 0,02$ olmalıdır. $k = 0,02$ kuramsal beklentilere de uygun olduğu için bu değer seçilmiştir.

Tablo 3-10. $k=0.02$ için Ridge ve EKK karşılaştırması

Bağımsız Değişkenler	Ridge Kat.	EKK Kat.	Std'mış Ridge	Std'mış EKK	Ridge Std.Hata	EKK Std.Hata
Sabit	21,31	17,91				
Yaş	0,14	0,14	0,08	0,08	0,11	0,12
Boy	-7,36	-6,90	-0,04	-0,03	13,33	13,60
Karın çevresi	29,69	29,83	0,16	0,16	11,30	11,44
Glikoz	0,00	0,00	0,01	0,00	0,03	0,03
HDL	-0,10	-0,11	-0,06	-0,06	0,11	0,12
LDL	0,02	0,02	0,04	0,04	0,03	0,03
Üre	-0,05	-0,06	-0,08	-0,10	0,07	0,08
Kreatinin	2,65	3,00	0,18	0,21	1,61	1,85
Potasyum	0,23	0,38	0,01	0,01	2,87	2,96
HGB	-0,99	-1,94	-0,08	-0,15	1,95	3,72
HCT	0,43	0,77	0,10	0,17	0,70	1,33
Diastolik	1,02	1,04	0,72	0,73	0,09	0,09
R^2	0,59	0,60				
Sigma	13,51	13,30				

$k=0,02$ değerine göre ridge katsayıları, EKK katsayıları, standartlaştırılmış ridge katsayıları ve EKK katsayıları, standart hataları, R^2 ve sigma değerleri Tablo3-10'da görülmektedir. Ridge regresyon çoklu bağlantı problemini ortadan kaldırdığı için bu yöntemle elde edilen kestirimlerin standart hatalarının EKK regresyonu ile elde edilen kestirimlerden daha düşük çıkması gerekmektedir. Tablo3-10'dan da görüleceği üzere özellikle güçlü çoklu bağlantıya neden olan HGB ve HCT değişkenlerinin kestirimlerinde standart hatalar yaklaşık olarak %50 oranında azalmıştır. Diğer tüm değişkenler için de standart hatalarda azalmalar meydana gelmiştir. Ayrıca analiz sonucunda R^2 ; EKK için 0.60, Ridge regresyon için 0.59 olarak hesaplanmıştır. Sonuç olarak Ridge regresyon, R^2 'de önemli bir değişime yol açmadan daha güvenilir, daha düşük standart hatalı kestirimler vermiştir.

Tablo 3.11. Ridge Regresyon Katsayıları

Bağımsız Değişkenler	Regresyon Katsayıları	Standart Hata	Std'mış Reg. Katsayıları	VIF	p değeri
Sabit	21,31				
Yaş	0,14	0,11	0,08	1,09	0,114
Boy	-7,36	13,33	-0,04	1,26	0,291
Karın çevresi	29,69	11,30	0,16	1,06	0,005
Glikoz	0,00	0,03	0,01	1,03	0,442
HDL	-0,10	0,11	-0,06	1,11	0,189
LDL	0,02	0,03	0,04	1,05	0,283
Üre	-0,05	0,07	-0,08	3,59	0,242
Kreatinin	2,65	1,61	0,18	3,46	0,049
Potasyum	0,23	2,87	0,01	1,10	0,468
HGB	-0,99	1,95	-0,08	6,52	0,307
HCT	0,43	0,70	0,10	6,61	0,268
Diastolik	1,02	0,09	0,72	1,01	<0.0001

Tablo 3-11, $k=0.02$ değeri için Ridge regresyonun detaylı çözümünü vermektedir. Ridge regresyon modelinde incelenen değişkenlerden karın çevresi, kreatinin ve diastolik kan basıncının istatistiksel olarak anlamlı olduğu belirlenmiştir. Yukarıdaki sonuca göre; söz konusu veriler için sistolik kan basıncını etkileyebilecek faktörlere ilişkin regresyon denklemimiz;

$$Y = 21.31741 + 29.68736 * X_1 + 2.64729 * X_2 + 1.022062 * X_3$$

Y = Sistolik kan basıncı

X_1 = Karın çevresi

X_2 = Kreatinin

X_3 = Diastolik kan basıncı

olarak elde edilmiştir.

Tablo 3.12. $k=0.02$ için ANOVA Tablosu

Değ.Kay.	sd	HKT	HKO	F-oranı	p
Sabit	1	2522820	2522820		
Model	12,00	29938,81	2494,90	13,67	<0.0001
Hata	115,00	20991,66	182,54		
Toplam	127,00	50930,47	401,03		
Bağ. Değ. Ort.	140,39				
Sigma	13,51				
R ²	0,59				
Değişim kats.	0,10				

Tablo 3.12’de $k=0.02$ için Ridge regresyon modelinin önemliliğini test eden varyans analizine ilişkin bilgiler sunulmuştur. Tablo 3.12’den de görüldüğü gibi $p<0.05$ olduğu için Ridge regresyon modeli anlamlıdır. Bir başka ifadeyle; sistolik kan basıncının istatistiksel olarak önemli değişkenlerle açıklanabileceği söylenebilir. O halde incelenen veriler için sistolik kan basıncı; karın çevresi, kreatinin ve diastolik kan basıncı değişkenleri ile %59 oranında tahmin edilebilir.

4. SONUÇLAR VE TARTIŞMA

Bağımsız değişkenlerin çoklu doğrusal bağlantı ile ilgili istatistikleri incelendiğinde, söz konusu verilerde güçlü çoklu doğrusal bağlantı probleminin var olduğu anlaşılmaktadır. Bu nedenle EKK yöntemiyle elde edilen sonuçlar geçerliliğini ve tutarlılığını kaybetmektedir. Diğer bir ifadeyle, çoklu doğrusal bağlantılı verilerde, regresyon katsayılarının standart hataları, büyüklükleri ve işaretleri doğru bir biçimde tahmin edilememektedir. Verilerde çoklu doğrusal bağlantı olması durumunda Ridge regresyon yöntemi EKK yöntemine göre daha durağan, kuramsal ve beklentilere uygun sonuçlar vermektedir. Her ne kadar EKK veya yanlı tahmin tekniği Ridge regresyon'dan birisinin seçimi, yanlı veya yansız tahmincilerden birisinin seçimi anlamına gelse de, gerçekte durum böyle değildir. Bilindiği gibi, pratik anlamda, EKK tahmincileri sadece modelin hatasız tanımlanması durumunda yansızdırlar. Bu nedenle pratikte EKK 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. Bu çalışmada çoklu doğrusal bağlantı sorununun bir sonucu olarak EKK regresyonuna ilişkin katsayılarının standart hatalarının, Ridge regresyonundan yüksek olduğu görülmektedir. Ridge regresyonun, hesaplanan daha küçük standart hata değerleri ile sistolik kan basıncının tahmininde daha tutarlı sonuçlar verdiğini belirlenmiştir. Ayrıca regresyon katsayılarının büyüklükleri de çoklu doğrusal bağlantıdan olumsuz bir şekilde etkilenmektedir. Bu problemi yine Ridge regresyonun yöntemi ortadan kaldırabilmektedir.

Optimum yanlılık sabitini araştırmak amacıyla VIF ve Ridge grafiklerinden yararlanarak, yanlı regresyon katsayılarının durağanlaştığı ve bu katsayılara ait VIF değerlerinin birlikte 10'dan küçük olduğu bölgede yaklaşık bir k değeri seçilerek iterasyonlara başlanmaktadır. İterasyonlar sonucunda Ridge Regresyon için seçilen optimum 0,02 yanlılık sabiti ($k = \%2$), standartlaştırılmış katsayılara ait VIF değerlerinin tümünün 10'dan küçük olduğu bölgede, kuramsal beklentilere uygun olarak tüm kriterlerin birlikte değerlendirilmesi ile seçilmiştir [5].

Tıpta klinik, patolojik, radyolojik, laboratuvar vb. bulguları ile bir fenomenin çok sayıda değişkeni ele alınarak veri toplanmaktadır. Bu değişkenlerin birçoğu birbirleri ile yüksek derecede ilişkili (correlated) olmakta, bazıları ise fenomen ile çelişir durumda bilgi

vermektedirler. Bu durum tanı ve tedavide tereddütlere, yanılgılara yol açabilmektedir. Hangi değişkenlerin problemi incelemede etkin, yardımcı rol oynadıklarını ve hangilerinin bu problemi incelemede etkin rolünün bulunmadığını belirlemede ileri istatistik yöntemlerden yararlanmak faydalı sonuçlar vermektedir [26]. Bu kapsamda, birbirleri ile yüksek derecede ilişkili olan değişkenler arasındaki çoklu bağlantı sorununu ortadan kaldırabilen değişik yöntemlerin kullanılması daha etkin sonuçlar verecektir. Bu çalışmada, sistolik kan basıncını etkileyebilecek değişkenler arasında güçlü çoklu doğrusal bağlantı belirlendiğinden, bu problemi ortadan kaldırmak amacıyla Ridge regresyonun yöntemi kullanılmıştır.

Bağımsız değişkenler birbiriyle bağıntılı ise, bunlardan bazılarının modelden çıkartılması gerekebilir. Ancak hangi değişkenler modelden çıkarılacağı özellikle klinik çalışmalarda oldukça önemlidir. Modelden yanlışlıkla bir değişkenin çıkartılması, modelin hatalı tanımlanmasına neden olabilir. Bu nedenlerden dolayı, çoklu doğrusal bağlantı söz konusu olduğunda, değişken çıkarmak yerine birbirinden bağımsız bileşenler türeten “Temel Bileşenler Regresyon Analizi” veya EKK tekniğinin düzeltilmiş şekli olan ve yanlış standartlaştırılmış regresyon katsayılarını tahmin eden “Ridge Regresyon” teknikleri kullanılabilmektedir [5]. Ayrıca bu iki yönteme ek olarak; modeldeki değişkenleri çıkarmadan çoklu bağıntıyı ortadan kaldırabilmek için EKK kestiricisine alternatif olarak önerilen ve EKK kestiricilerine göre daha küçük hata kareler ortalaması veren kestiriciler; Stein ve Liu kestiricileri de kullanılabilmektedir [4].

Bir olayın ortaya çıkmasında; etkisi bilinen, etkisi olduğundan şüphelenilen ve etkisinin olup olmadığı henüz bilinmeyen çok sayıda etken (faktör, belirleyici değişkenler) bulunabilir. Bu gibi durumlarda sebep-sonuç ilişkilerini bilimsel olarak ortaya çıkarabilmek için çeşitli istatistik yöntemlerden yararlanmak gerekir. Özellikle sebep-sonuç ilişkisi bilinen durumlarda bile etkenlerin bazılarının majör etken, bazılarının ise minör etken olarak tanımlanması, problemlerin çözülmesinde önceliklerin ortaya konması bakımından büyük önem taşır. Bu çalışmada, sistolik kan basıncını etkileyebilecek bazı değişkenler arasındaki matematiksel model, EKK regresyonu ve Ridge regresyon ile tahmin edilmiştir. Modelde yer alan önemli değişkenler, bireyleri daha hasta olmadan korumak, aydınlatmak ve koruyucu tedbirler alınabilmeleri açısından önemli olabilmektedir. MRFIT çalışmasında, 316.000 erkek hastanın 12 yıllık takibinde sistolik kan basıncının diastolik kan basıncından daha fazla hipertansiyon ve kardiovasküler risk belirleyicisi olduğu bildirilmiştir [22].

Bu nedenle çalışmada yanıt değişkeni olarak sistolik kan basıncı seçilmiştir. Ridge regresyonu analizi sonucunda; karın çevresi, kreatinin, diastolik kan basıncı değişkenleri, sistolik kan basıncını tahmin etmek için istatistiksel açıdan önemli faktörler olarak bulunmuştur.

Hipertansiyon, dünyada önlenebilir ölüm nedenleri içerisinde önde gelen risk faktörlerindendir. 2000 yılı itibariyle 972 milyon insanda hipertansiyon vardır ve bu sayı dünyadaki erişkin nüfusunun % 26,4'üne denk gelmektedir. Türkiye'de yaklaşık 15-16 milyon hipertansiyon hastasının olduğu öngörülmektedir. Türkiye genelinde yaygınlığı, yandaş hastalık ve risk faktörleri ile birlikteliği göz önüne alındığında hipertansiyon ülkemiz için önemli bir halk sağlığı sorunudur [9].

Hipertansiyon erken tanı ile kontrol altına alınabilen bir hastalık olduğu için değişik risk faktörlerinin belirlenmesi ile gelişimi ve ilerlemesine karşı önlemler alınabilir. Türkiye'de hipertansiyonun tahmini ve ilişkili faktörlerin belirlenmesine yönelik olarak ileri istatistik yöntemleri ele alan ve karşılaştıran çok fazla çalışma bulunmamaktadır. Türe ve ark. çalışmalarında, primer hipertansiyonun tahminlenmesinde lojistik regresyon analizi (LR), esnek diskriminant analizi (EAA), ve yapay sinir ağları (YSA), yöntemlerini karşılaştırmışlardır. Sonuç olarak, aile hikâyesi, lipoprotein A, trigliserid, sigara kullanımı ve vücut kitle indeksi değişkenlerinin, kontrol ve hipertansiyonlu hasta gruplarını tahmin etmede kullanılabileceğine ve YSA'nın performansının LR ile EAA'dan istatistiksel olarak daha iyi olduğuna karar vermişlerdir [30]. Ayrıca Onay çalışmasında, ortalama arteriyel kan basıncını etkileyen değişkenleri temel bileşenler regresyonu ile belirlenmeye çalışmıştır. Temel bileşen regresyonunu uyguladığında, yaş, nabız, boy, HGB, trigliserit, HDL kolesterol, LDL kolesterol ve HCT değişkenlerinin, ortalama arteriyel kan basıncını etkileyen önemli faktörler olduğunu belirlemiştir [23]. Bu çalışmada, Ridge regresyon analizi sonucunda, sistolik kan basıncını tahmin etmek için istatistiksel olarak önemli faktörler; karın çevresi, kreatinin ve diastolik kan basıncı olarak bulunmuştur. Bu sonuç, var olan önemli değişkenlere ek olarak hipertansiyonun kontrol altına alınabilmesi açısından dikkate alınmalıdır.

Sonuç olarak, sistolik kan basıncını etkileyebilecek değişkenler arasında güçlü çoklu doğrusal bağlantı olduğu için, Ridge Regresyon yönteminin EKK yöntemine göre daha tutarlı, geçerli, durağan ve kuramsal beklentilere uygun tahminler sağladığı görülmektedir.

5. ÖNERİLER

1. İncelenen problemlerde çoklu bağlantılı bağımsız değişkenlerin modelden çıkartılması istenmiyor ve çoklu bağlantılı sorunun giderilmesi de gerekli ise, Ridge Regresyon yönteminin kullanılması önerilmektedir. Ancak yanlılık katsayısının seçiminde, tüm kriterlerin birlikte değerlendirilmesi tavsiye edilebilir.
2. Çalışmada hipertansiyonla birincil derecede ilişkili olabileceği düşünülen idrarda sodyum miktarı ve tiroid hormonu gibi bazı değişkenler Fırat Tıp Merkezi'ne ek maliyet getirdiği için ve maddi olanaklar kısıtlı olduğu için modele alınamamıştır. Ayrıca zaman ve insan gücündeki kısıtlardan dolayı aile öyküsü, sigara içme, kardiyovasküler hastalık hikayesi gibi değişkenler de modele alınamamıştır. Sonraki çalışmalarda bu değişkenlerin de ilave edilmesi modelin gücünü artırabilecektir.
3. Hipertansiyonu etkileyebilecek birçok geleneksel ve yeni risk faktörleri vardır. Gelişen teknolojinin, özellikle 3G ve cep telefonunun radyasyon yayıcı özelliği, kan basıncının yükselmesi ile çeşitli kalp hastalıklarına neden olabilir. Bu nedenle geleneksel ve yeni risk faktörleri ile hipertansiyon ilişkisi, çok merkezli ve daha büyük örneklemelerde araştırılmalıdır.
4. Hipertansiyonun tahminlenmesine yönelik olarak; sağlam olanların ilgilenilen etken(ler)e maruz kalma durumları sonrasında, tüm grup (kohort) hastalığın ortaya çıkması açısından yeterli bir süre boyunca izlenerek, hastalığın gelişme durumlarının saptandığı kohort araştırmaları gerçekleştirilebilir. Kohort araştırmalarda, etken(ler)e maruz kalan ve kalmayan gruptaki hastalık risklerinin oranlanmasına ile hesaplanan göreceli riskler hipertansiyon için de hesaplanabilir.
5. Çalışmanın sonraki aşamalarında, sistolik kan basıncını etkileyebilecek faktörlerin belirlenmesinde; yapay sinir ağları, temel bileşenler regresyonu, Stein ve Liu kestiricileri ile bunların olası uyarlamaları en etkin istatistik yöntemin bulunabilmesi için karşılaştırılabilir.

KAYNAKLAR

- [1] **Akdeniz, F. ve Çabuk, A.** Ridge regresyon teorisinde 1970-2001 arası gelişmeler. <http://idari.cu.edu.tr/sempozyum/bil24.htm>.
- [2] **Akgül, A.**, 1997. Tıbbi Araştırmalarda İstatistiksel Analiz Teknikleri. Ankara, 419.
- [3] **Aktaş, C. ve Yılmaz, V.** 2003. Çoklu bağlantılı modellerde Liu ve Ridge regresyon kestiricilerinin karşılaştırılması. *Anadolu Üniversitesi Bilim Ve Teknoloji Dergisi*, **4**, 2, 189-194
- [4] **Aktaş, C.** 2007. Çoklu bağıntı ve Liu kestiricisiyle enflasyon modeli için bir uygulama. *ZKÜ Sosyal Bilimler Dergisi*, **3**, 6, 67-79
- [5] **Albayrak, A.,S.**, 2005. Çoklu doğrusal bağlantı halinde en küçük kareler tekniğinin alternatifi yanlı tahmin teknikleri ve bir uygulama. *ZKÜ Sosyal Bilimler Dergisi*, **1**, 1.
- [6] **Alpar, R.**, 1997 . Uygulamalı Çok Değişkenli İstatistiksel Yöntemlere Giriş, Ankara.
- [7] **Freund, J.,E., Çev: Şenesen,Ü.** 2001. Matematiksel İstatistik. Literatür Yayıncılık, İstanbul.
- [8] **Girginer, N.**, 1996 . Ridge regresyon yöntemiyle Türk tekstil ve konfeksiyon sanayii dışsatımının analizi, *Yüksek Lisans Tezi*, Anadolu Üniversitesi Sosyal Bilimler Enstitüsü ,Eskişehir.
- [9] **Göçgeldi,E.,Babayiğit,M.,A.,Hassoy,H.,Açıkel,C.,Taşçı, İ.,Ceylan,S.** 2008. Hipertansiyon tanısı almış hastaların algıladıkları yaşam kalitesi düzeyinin ve etki eden faktörlerin değerlendirilmesi. *Gülhane Tıp Dergisi*. **50**,172-179
- [10] **Gujarati,D.,N. , Çev: Şenesen,Ü, Şenesen,G.,G.** 1999. Temel Ekonometri. Literatür Yayıncılık. İstanbul.
- [11] **Hintze,J.** 2007. NCSS and GESS. NCSS, LLC, Kaysville, Utah, USA. www.ncss.com
- [12] **Hoerl, A.,E. and Kennard, R.,W.** 1970a. Ridge regression: biased estimation for non-orthogonal problems. *Technometrics*, **12**,1,55-67.
- [13] **Hoerl, A.,E. and Kennard, R.,W.** 1970b. Ridge Regression: Applications to Non-Orthogonal Problems. *Technometrics*, **12**,1,69-82.
- [14] **İmir, E.**, 1986. Çoklu bağlantılı doğrusal modellerde ridge regresyon yöntemiyle parametre kestirimi , *Yüksek Lisans Tezi*, Anadolu Üniversitesi Sosyal Bilimler Enstitüsü ,Eskişehir.
- [15] **İpek,O.**, Ridge Regresyon Üzerine Bir Çalışma <http://idari.cu.edu.tr/sempozyum/bil28.htm>.

- [16] **Kalaycı, Ş.**,2008. SPSS Uygulamalı Çok Değişkenli İstatistik Teknikleri, Asil Yayın Dağıtım, Ankara.
- [17] **Karagöz,M.**, 2006. İstatistik Yöntemleri, Ekin Kitabevi Yayınları. Bursa.
- [18] **Marquardt,D.,W.** 1970. Generalized inverses, ridge regression, biased linear estimation,and nonlinear estimation , *Technometrics*, **12**, 3,591-612.
- [19] **Marquardt,D.,W. and Snee, R.,D.** 1975. Ridge Regression in Practice. *JASA Journal Of The American Statistical Assocation*, **29**, 1, 3-20.
- [20] **Mcdonald, G.,C. and Galarneau, D.,I.** 1975. A monte carlo evaluation of some ridge type estimators. *JASA Journal Of The American Statistical Assocation*, **70**, 350, 407-416.
- [21] **Montgomery,D.,Peck,E.** 1991. *Introduction To Linear Regression Analysis*, NewYork.
- [22] **MRFIT.** 1982. Multiple risk factor intervention trial. Risk factor changes and mortality results, *JAMA*, 248,1465-1477.
- [23] **Onay,İ.** 2003. Ortalama arteriyel kan basıncını etkileyen faktörlerin temel bileşen regresyonu ile belirlenmesi.*Yüksek Lisans Tezi*.Gazi Üniversitesi, Fen Bilimleri Enstitüsü.
- [24] **Orhunbilge,A.,N.** 2002. Uygulamalı Regresyon ve Korelasyon Analizi ,İ.Ü Basım ve Yayınevi ,İstanbul.
- [25] **Özdamar, K.** 2003. SPSS ile Biyoistatistik, Eskişehir.
- [26] **Özdamar, K.** 2005. Hipertansiyon ve tahmin modelleri.
<http://www.anakarder.com/yazilar.asp?yaziid=508&sayid=20>
- [27] **Özkale,M.,R.** 2007. Çoklu iç ilişki ile ilgili problemler, *Doktora tezi*, Çukurova Üniversitesi Fen Bilimleri Enstitüsü, Adana.
- [28] **Sümbüloğlu,K.,Sümbüloğlu,V.** 2007. Biyoistatistik . Hatipoğlu Basım Yayın, Ankara.
- [29] **Sümbüloğlu,K.,Sümbüloğlu,V.** 2005. Klinik ve Saha Araştırmalarında Örneklem Yöntemleri Ve Örneklem Büyüklüğü, Alp Ofset Matbaacılık ,Ankara.
- [30] **Türe,M., Kurt,İ., Yavuz,E.,Kürüm,T.** 2005. Hipertansiyonun tahmini için çoklu tahmin modellerinin karşılaştırılması.(Sinir ağları, lojistik regresyon ve esnek ayırma analizleri). *Anadolu Kardiyoloji Dergisi*. **5**, 24-8
- [31] **Türkay,G.,S.** 1996. Ridge regresyon yöntemiyle tofaş firmasının (1975-1994) yılları arası otomobil talep miktarı analizi, *Yüksek Lisans Tezi* .Anadolu Üniversitesi Sosyal Bilimler Enstitüsü, Eskişehir.
- [32] **Uslu,V.,R.** 1991. Ridge regresyon ve öğrenci başarısı üzerine bir uygulama, *Yüksek Lisans Tezi*, Ondokuz Mayıs Üniversitesi Fen Bilimleri Enstitüsü, Samsun.
- [33] **Ünver,Ö.,Gamgam,H.** 1986 . Uygulamalı İstatistik Yöntemler ,Ankara.

- [34] **Wetherill, G., B.** 1986. *Regression Analysis with Applications* , Chapman and Hall, New York.

EKLER

EK Tablo 1: Veri tablosu

SIRA	YAŞ	BOY	KAR, ÇEV	GLUKOZ	HDL	LDL	ÜRE	KREATİNİN	POTASYUM	HGB	HCT	DİASTOLİK	SİSTOLİK
1	52	1,55	1,13	133	38	156	20	0,7	4,3	15	42,7	90	180
2	53	1,52	0,86	86	64	135	28	0,9	4,6	14	39,2	90	140
3	61	1,64	1,08	89	29	114	41	1,4	5	14	39,9	80	130
4	71	1,65	1,1	88	53	186	39	1	4,5	13	38,8	100	160
5	54	1,76	1,12	92	38	96	27	1,2	4,6	17	48,1	100	140
6	58	1,83	1	97	45	106	27	1,3	4,5	15	44,1	90	130
7	80	1,47	0,75	97	35	105	85	1,9	4,3	13	36,4	90	100
8	77	1,5	0,91	100	58	182	43	1,1	4,3	16	46,8	80	140
9	47	1,68	1,07	56	48	150	69	1,6	3,2	15	43,7	110	170
10	61	1,55	0,84	196	55	125	15	0,8	4,6	14	39,1	90	130
11	47	1,5	0,92	90	82	109	29	0,8	4,1	12	35	80	120
12	54	1,6	1,25	120	71	219	44	1,1	4	14	40,1	90	150
13	63	1,74	0,86	89	39	120	57	2,7	2,9	13	37,3	80	160
14	60	1,8	0,97	171	44	126	32	1	4,2	16	45,4	80	120
15	35	1,63	1,07	88	47	144	18	0,7	4,6	13	38,8	90	130
16	52	1,68	0,87	79	84	102	30	1,6	4,9	15	43,2	110	170
17	67	1,64	1,1	91	75	113	37	1,1	4,1	16	46,5	80	120
18	54	1,76	1,03	80	42	139	31	1,1	4,1	14	40,2	100	140
19	64	1,62	1,04	91	40	99	31	0,7	4,2	12	35,4	70	130
20	48	1,6	0,98	124	45	141	27	0,8	4,1	13	36,6	120	190
21	40	1,58	1,02	86	54	153	19	0,7	4,6	13	38,7	70	110
22	80	1,47	1,02	273	56	146	39	1	5	15	47,1	100	150
23	64	1,76	1,03	93	51	167	31	1	4,9	13	39	70	140
24	59	1,62	1,1	172	44	173	37	0,9	4,9	14	40,9	120	170
25	67	1,64	0,94	94	52	134	60	1,3	5	14	41,6	100	150
26	62	1,68	0,99	94	62	164	30	0,9	4,2	14	41,2	70	110
27	49	1,65	1,03	119	34	143	14	0,8	4,7	12	34,8	110	150
28	63	1,7	1,14	88	45	123	52	1,8	5,4	16	46,5	80	135

SIRA	YAŞ	BOY	KAR,ÇEV	GLUKOZ	HDL	LDL	ÜRE	KREATİNİN	POTASYUM	HGB	HCT	DİASTOLİK	SİSTOLİK
29	55	1,9	0,91	88	70	105	73	2,4	4,9	13	36,3	70	110
30	48	1,5	0,76	71	45	122	101	2,8	5,3	9,8	27,5	90	140
31	77	1,55	0,78	93	58	111	101	2,4	5,6	12	33,9	50	100
32	25	1,72	0,9	93	48	156	77	1,3	5,2	14	40,9	80	120
33	65	1,54	1,05	97	48	164	56	0,8	3,9	13	36,2	80	120
34	60	1,69	1,01	95	44	173	31	1,4	4,8	16	46,2	80	120
35	55	1,68	1,11	108	52	129	28	1,2	4,9	15	41,9	80	140
36	60	1,58	1,16	123	60	279	37	1,1	4,7	14	41	90	160
37	52	1,55	0,97	93	46	110	34	0,9	4,9	14	39,9	80	120
38	55	1,56	1,22	88	58	208	29	0,7	4	15	41,5	80	140
39	25	1,63	0,93	90	48	134	12	0,7	4,5	14	41,2	70	110
40	50	1,68	1,01	89	46	198	29	1,1	4,9	17	48,3	80	170
41	52	1,75	1,03	85	32	130	28	1,1	4,9	14	40,4	70	130
42	70	1,52	1,17	112	78	116	29	0,6	3,8	14	41,9	80	130
43	51	1,75	1,07	127	50	163	27	1,1	4,3	15	43,6	70	130
44	53	1,55	1,12	107	57	112	34	0,8	5,1	14	39,7	90	150
45	56	1,82	1,07	221	45	148	50	1,4	5,6	17	47,4	70	110
46	60	1,5	1,15	95	50	116	28	0,8	3,8	15	42,3	100	180
47	48	1,78	0,9	94	48	221	34	1,1	4,5	15	42	80	100
48	48	1,52	1,03	90	53	113	30	0,7	4,3	16	45,3	100	150
49	47	1,77	1,05	85	47	142	34	1,1	4,6	16	45,2	80	120
50	30	1,75	1	164	48	58	278	13,9	3,9	9	25,4	80	130
51	60	1,45	0,87	90	56	106	23	0,8	4,3	13	37,3	80	140
52	71	1,8	0,89	158	40	122	223	3,1	5,7	12	34,9	80	140
53	62	1,66	1,05	15	48	146	41	1,3	4,4	14	40	80	120
54	48	1,68	0,97	77	43	148	34	1	4,7	17	47,4	95	145
55	71	1,6	1,05	127	42	211	48	1	4,5	12	35,5	100	160
56	60	1,62	1,18	222	44	146	30	0,8	4	13	37,3	80	150
57	73	1,8	1,06	86	37	155	60	1,1	5,1	13	36,5	90	140
58	75	1,87	1,1	82	42	96	28	0,8	4,6	15	42,5	70	130

SIRA	YAŞ	BOY	KAR,ÇEV	GLUKOZ	HDL	LDL	ÜRE	KREATİNİN	POTASYUM	HGB	HCT	DİASTOLİK	SİSTOLİK
59	65	1,6	0,95	87	46	163	60	1,3	4,7	17	49,5	80	120
60	49	1,6	0,97	80	40	147	41	1,1	4,8	15	43,3	90	160
61	52	1,53	0,98	82	49	189	32	0,9	4,2	13	39	100	150
62	52	1,6	1,04	268	50	106	28	0,9	4,3	12	34,1	100	150
63	69	1,65	1,12	118	53	164	45	1,4	4,9	17	49,1	110	190
64	63	1,6	0,8	107	47	157	31	1,2	4,5	15	42,6	100	150
65	66	1,65	0,86	84	40	184	34	1,2	4,7	13	37,3	90	150
66	57	1,57	1,07	113	44	141	58	2	4,3	14	40,7	100	150
67	46	1,5	1,18	192	74	109	19	0,7	4,8	12	36,7	100	150
68	57	1,62	0,95	95	49	140	27	0,8	4,5	13,5	39,2	130	180
69	71	1,65	1	91	53	159	47	1,3	4,1	15,1	42,6	80	140
70	73	1,72	1,03	78	51	175	41	1,4	4,3	16,6	48,7	60	120
71	57	1,67	1,09	91	55	123	46	1,2	4,5	15,2	43,6	70	130
72	50	1,8	1,02	103	63	119	33	1,6	4,8	16,2	47,1	80	140
73	65	1,48	1,2	92	66	203	37	1,2	4,9	13,7	40	80	140
74	59	1,45	0,96	110	49	140	44	1,3	4,5	14,1	40,9	90	160
75	50	1,54	1,09	106	36	150	177	8,9	4,1	11,7	32,5	90	180
76	49	1,43	0,96	94	53	153	36	0,9	4,5	14,8	41,6	80	150
77	69	1,64	1,04	95	50	207	41	1,3	4,3	14,5	41,6	80	120
78	55	1,7	1,13	219	50	186	26	1,8	4,4	15,5	45,1	90	160
79	62	1,73	1,07	106	46	173	35	1,3	4,3	16,6	47,9	110	170
80	45	1,6	1,29	107	53	127	28	0,8	4,3	14,8	43,1	60	130
81	42	1,57	1,16	110	49	140	32	1,1	4,5	12,5	37	80	140
82	50	1,56	1,12	91	45	177	26	0,9	4,6	15,6	46	70	110
83	50	1,7	1,01	110	49	140	44	1,3	4,5	14,1	40,9	100	170
84	63	1,6	1,04	98	45	126	16	0,9	4,4	14,3	41,9	110	190
85	59	1,6	1,2	117	46	163	32	0,9	4,6	14,4	41,8	100	150
86	60	1,58	1,05	96	47	102	38	1	4,3	14,7	41,8	70	110
87	64	1,8	1,1	147	42	154	50	1,2	5,3	12,0	34,5	70	130
88	62	1,55	1,07	105	47	133	28	1	4,3	13,1	36,5	80	140

SIRA	YAŞ	BOY	KAR,ÇEV	GLUKOZ	HDL	LDL	ÜRE	KREATİNİN	POTASYUM	HGB	HCT	DİASTOLİK	SİSTOLİK
89	72	1,73	1,07	196	41	66	28	1	4,3	16,3	46,2	80	120
90	68	1,55	1,04	97	54	84	49	1,1	5,1	11,9	37,7	90	140
91	51	1,75	1,07	90	43	75	26	1,1	4,9	15,5	44,5	70	120
92	63	1,68	1,11	107	50	77	41	1,3	4,7	14	41	90	180
93	78	1,55	0,75	94	39	119	27	0,8	4,4	12,4	36,1	80	130
94	72	1,85	1,05	128	41	73	45	1,4	5,4	14,9	42,9	80	140
95	58	1,56	0,93	98	58	101	41	1	4	13,6	43,4	80	140
96	63	1,63	1,05	102	59	142	43	0,8	4,3	13,6	39,1	90	150
97	48	1,6	1,1	235	55	134	69	1,7	5,4	11,7	33,8	80	140
98	50	1,8	1,04	93	47	111	25	1,1	4,8	15,8	45,5	80	130
99	53	1,75	0,97	256	50	167	31	0,9	4,4	15,3	44,2	70	140
100	57	1,65	0,95	77	37	90	36	1,2	4,2	16,4	47,1	70	130
101	55	1,72	0,89	97	54	83	35	1,0	4,3	14,3	41,3	90	140
102	70	1,55	0,85	89	73	254	39	0,9	4,5	12,9	37,2	80	120
103	68	1,51	1,06	140	54	132	38	0,8	4,6	10,9	33,4	80	140
104	55	1,82	0,9	311	34	68	59	1,4	4,7	12,8	37,5	90	140
105	51	1,65	1,03	129	39	122	30	1	4,4	15,5	44,6	120	170
106	65	1,65	1,09	105	45	103	53	1,2	4,4	14,6	41,5	80	120
107	71	1,6	1,03	76	43	124	29	0,9	4,4	14,6	41,9	100	150
108	63	1,65	1,09	94	43	108	31	1	4,8	17,5	50,1	110	170
109	20	1,58	0,75	79	80	122	40	0,9	4,6	13,4	38,6	90	120
110	44	1,78	1,15	100	41	117	23	0,7	4,2	14,6	41,8	120	170
111	61	1,61	0,96	173	34	115	32	0,9	3,5	14,7	41,5	100	140
112	58	1,62	1,26	95	58	152	36	0,7	4,4	14,3	40,5	80	120
113	65	1,58	1,01	93	79	185	28	0,9	4,6	13	39,9	80	140
114	61	1,75	1	85	42	175	37	1,6	4,3	14,8	43	80	130
115	79	1,7	1,06	111	40	126	64	1,1	4,2	15,9	46	110	160
116	63	1,6	1,12	77	54	105	28	0,8	4,3	13	37,2	100	150
117	63	1,64	0,93	93	52	160	34	1,2	3,9	15,6	44,5	80	120
118	58	1,76	1,05	89	44	76	32	1	3,7	15,9	45,0	70	130

SIRA	YAŞ	BOY	KAR,ÇEV	GLUKOZ	HDL	LDL	ÜRE	KREATİNİN	POTASYUM	HGB	HCT	DİASTOLİK	SİSTOLİK
119	59	1,81	0,93	89	45	122	37	1	4,4	15,3	44,6	90	160
120	75	1,6	0,92	98	51	65	49	1	4,4	12,3	35,9	80	130
121	42	1,54	1,3	130	45	130	24	0,8	4,1	12,3	37,1	80	120
122	67	1,67	1,05	112	43	115	38	1,1	4,3	13,9	40,7	70	130
123	70	1,65	1,04	112	46	130	68	1,7	4,3	15,6	45,3	70	130
124	55	1,48	1,02	105	76	162	28	0,9	4,4	16,7	48,8	90	130
125	52	1,57	1,08	95	58	110	25	1	4,6	13,2	38,7	80	140
126	55	1,73	0,99	70	54	113	31	1	4,4	15,2	44	90	150
127	55	1,65	1,12	90	81	145	33	0,9	4,4	14,7	41,9	110	160
128	54	1,58	1,07	112	43	175	25	0,7	4,2	16,4	44,4	90	130

ÖZGEÇMİŞ

1982 yılında Elazığ'da doğdum. İlk, orta ve lise öğrenimimi Elazığ'da tamamladıktan sonra 2000 yılında Fırat Üniversitesi Matematik bölümünü kazandım. 2004 yılında bölüm derecesiyle mezun olduktan sonra dört yıl çeşitli özel kuruluşlarda matematik öğretmenliği yaptım. 2007-2008 bahar döneminde Fırat Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilimdalı, Uygulamalı İstatistik bölümünde yüksek lisansa başladım. 2009 Kasım Fırat Üniversitesi Fen-Edebiyat Fakültesi İstatistik Bölümü, Uygulamalı İstatistik anabilim dalına araştırma görevlisi olarak atandım. Halen bu göreve devam etmekteyim. Evli ve bir çocuk annesiyim.