LOJİSTİK REGRESYON MODELİ VE GERİYE DOĞRU ELİMİNASYON YÖNTEMİYLE DEĞİŞKEN SEÇİMİNİN HİPERTANSİYON RİSKİ ÜZERİNE UYGULAMASINDA BOOTSTRAP YÖNTEMİ

Özgür ATABEY

YÜKSEK LİSANS TEZİ
İSTATİSTİK

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

EKİM 2010
ANKARA

Özgür ATABEY tarafından hazırlanan "LOJİSTİK REGRESYON MODELİ VE GERİYE DOĞRU ELİMİNASYON YÖNTEMİYLE DEĞİŞKEN SEÇİMİNİN HİPERTANSİYON RİSKİ ÜZERİNE UYGULAMASINDA BOOTSTRAP YÖNTEMİ" adlı bu tezin Yüksek Lisans olarak uygun olduğunu onaylarım.

Yrd.Doç.Dr. Meltem EKİZ …….
Tez Danışmanı, İstatistik Anabilim Dalı

Bu çalışma, jürimiz tarafından oy birliği ile İstatistik Anabilim Dalında Yüksek Lisans tezi olarak kabul edilmiştir.

Yrd.Doç.Dr. Jale BALİBEYOĞLU …….
İstatistik Anabilim Dalı, G.Ü.

Yrd.Doç.Dr. Meltem EKİZ …….
İstatistik Anabilim Dalı, G.Ü.

Yrd.Doç.Dr. Sibel ATAN …….
Ekonometri Anabilim Dalı, G.Ü.

Tarih: 08/10/2010

Bu tez ile G.Ü. Fen Bilimleri Enstitüsü Yönetim Kurulu Yüksek Lisans derecesini onamıştır.

Prof. Dr. Bilal TOKLU …….
Fen Bilimleri Enstitüsü Müdürü

TEZ BİLDİRİMİ

Tez içindeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, ayrıca tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

Özgür ATABEY

**LOJİSTİK REGRESYON MODELİ
VE GERİYE DOĞRU ELİMİNASYON YÖNTEMİYLE
DEĞİŞKEN SEÇİMİNİN HİPERTANSİYON RİSKİ ÜZERİNE
UYGULAMASINDA BOOTSTRAP YÖNTEMİ
(Yüksek Lisans Tezi)**

Özgür ATABEY

**GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

Ekim 2010

ÖZET

Bu çalışma beş bölümden oluşmaktadır. İkinci bölümde lojistik regresyon modeli hakkında genel bilgiler verildikten sonra lineer regresyon modeliyle arasındaki ilişki ve lojistik regresyonun tercih edilme nedenleri üzerinde durulmuştur. Üçüncü bölümde tek değişkenli ve çok değişkenli lojistik regresyon modellerinde parametre tahmin yöntemleri, katsayıların önem testleri, model katsayılarının yorumlanması, model yapılandırma stratejileri ve model uyumunun belirlenmesi detaylı bir şekilde anlatılmıştır. Dördüncü bölümde yeniden örnekleme tekniklerinden biri olan bootstrap örnekleme yöntemi hakkında genel bilgiler verilmiştir. Uygulama bölümünde ise hastaneye hipertansiyon şikayeti ile başvuran hastalar üzerinde çalışılmış olup geriye doğru eleme yöntemi ile parametre tahminleri yapılarak lojistik regresyon modeli kurulmuştur. Daha sonra aynı veri seti üzerine bootstrap yöntemi uygulanmış, bulunan parametre tahminleri ve standart hatalar geriye doğru eleme yöntemi sonucunda bulunan değerlerle karşılaştırılmıştır.

Bilim Kodu : 205.1.066
Anahtar Kelimeler : Lojistik Regresyon, Bootstrap Yöntemi, Geriye Doğru Eleme Yöntemi
Sayfa Adedi : 122
Tez Yöneticisi : Yrd.Doç.Dr. Meltem EKİZ

THE BOOTSTRAP METHOD IN THE IMPLEMENTATION OF VARIABLE SELECTION ON HYPERTENSION RISK THROUGH LOGISTIC REGRESSION MODEL AND BACKWARD ELIMINATION
(M.Sc.Thesis)

Özgür ATABEY

GAZI UNIVERSITY
INSTITUTE OF SCIENCE AND TECNOLOGY
October 2010

ABSTRACT

This study contains five chapters. In the second chapter after general information is given on the logistic regression model, its relation with the linear regression model and the reasons of prefering logistic regression are summarized. Parameter estimating methods for single and multi-variable logistic regression models, significance tests of coefficients, interpretations of model coefficients, model structuring strategies and the determination of fitting the model are explained in detail in the third chapter. The fourth chapter is focused on the general information of the bootstrap sampling method, which is one of the re-sampling techniques. Furthermore, the patients making applications on hypertension to the hospital are studied and by using the backward elimination method the logistic regression model is built. Then the bootstrap method is applied on the same data set and parameter estimates and standart errors are compared with the results obtained from the backward elimination method.

Science Code : 205.1.066
Key Words : Logistic Regression, Bootstrap Method, Backward Elimination Method
Page Number : 122
Adviser : Asisst.Prof. Meltem EKİZ

TEŞEKKÜR

Bu tez konusunda bana yön veren, çalışmam süresince değerli öneri ve eleştirileri ile benden desteğini esirgemeyen değerli hocam Sayın Yrd. Doç. Dr. Meltem EKİZ' e, çalışmam boyunca gösterdikleri sabır, anlayış ve desteklerinden dolayı annem Necla ATABEY, babam Kemal ATABEY, ablam Demet ATABEY' e ve tüm arkadaşlarıma teşekkür etmeyi borç bilirim.

İÇİNDEKİLER

Sayfa

Sayfa

Sayfa

ÇİZELGELERİN LİSTESİ

Çizelge **Sayfa**

Çizelge **Sayfa**

ŞEKİLLERİN LİSTESİ

Şekil **Sayfa**

Şekil **Sayfa**

1. GİRİŞ

Regresyon analizi, bağımlı(cevap, sonuç, yanıt) değişken ile bir veya daha fazla bağımsız değişken arasındaki sebep-sonuç ilişkisini ortaya koyma amacına yönelik yapılan en önemli veri analiz yöntemlerinden birisidir[3].

Bağımsız değişken ile sonuç (bağımlı) değişkeni arasında ilişki kurmak için istatistiksel uygulamalarda birçok regresyon yöntemi geliştirilmiştir. Geliştirilen yöntemlerden sadece bir tanesinin kullanımı her zaman hedeflenen noktaya ulaşmamız için yeterli olmamaktadır. Bu nedenle alternatif yöntemler geliştirilmiştir. Lojistik Regresyon da bu sayede alternatif olarak üretilen yöntemlerden birisidir.

Çok değişkenli istatistiksel verilerin sınıflandırılması, bu verilere uygulanabilecek çeşitli istatistiksel yöntemler için gerekli ve yararlı bilgiler verecektir. Gözlemleri verilerin yapısında bulunan olası gruplara atamak için kullanılan yöntemlerden üç tanesi, kümeleme(clustering), diskriminant(discriminant) ve lojistik regresyon(logistic regression) analizidir.

Kümeleme analizinde verilerin yapısındaki grup(küme) sayısı bilinmemekte, gözlemler uzaklık ya da benzerlik ölçütlerine göre kümelenmektedir. Burada amaç yalnızca gözlemlerin oluşturduğu küme yapısını bulmaktır. Discriminant ve lojistik regresyon analizinde ise verilerin yapısındaki grup sayısı bilinmekte ve bu verilerden yararlanarak bir ayrımsama modeli elde edilmektedir. Kurulan bu model yardımı ile veri kümesine yeni alınan gözlemlerin gruplara atamaları yapılmaktadır.

Bu üç yöntemden kümeleme ve diskriminant analizi şimdiye dek çok geniş olarak incelenmiş yöntemlerdir. Lojistik regresyon ise daha çok son yıllarda yoğun bir şekilde kullanılmaya başlanmıştır. Yöntem genelde çeşitli varsayım bozulumları durumunda diskriminant analizi ve çapraz tablolara bir alternatif olarak uygulanmaktadır. Kullanım nedeni olarak lineer regresyon analizinin uymadığı bazı durumlar olması gösterilebilir. Bağımlı değişken 0, 1 gibi iki ya da ikiden çok düzey içeren kesikli değişken olduğunda normallik varsayımı bozulmakta ve lineer

regresyon analizi uygulanamamaktadır. Böyle durumlarda lojistik regresyon analizi önerilmektedir[5]. Bu üç yöntemin karşılaştırması aşağıda verildiği gibidir.

Diskriminant Analizi:

1. Küme sayısı önceden bilinmekte ve analiz boyunca değişmemektedir.

2. Gelecekte kullanılabilir fonksiyonlar verir. Bu özelliği ile kümelemeden ayrılır; ana bileşenler, kanonik korelasyon ve çok değişkenli regresyon analizine benzer.

3. Değişkenlerin bazıları sürekli bazıları ise kesikli olamaz.

4. Sık kullanılan, en çok bilinen yöntemdir.

5. Veriler normal dağılımlı olmalıdır.

Kümeleme Analizi:

1. Küme sayısı önceden tam olarak bilinmemektedir.

2. Gelecekte kullanılabilir özelliği yoktur ve az kullanılan bir yöntemdir.

3. Kovaryans matrisine ilişkin varsayım yoktur.

4. Çok değişkenli varyans analizi, lojistik regresyon analizi, çok boyutlu ölçekleme gibi çok değişkenli analizlerle yakından ilgilidir.

5. Veriler veya uzaklık değerleri normal dağılımlı olmalıdır

Lojistik Regresyon:

1. Küme sayısı önceden bilinmektedir.

2. Değişkenlerden bazıları kesikli bazıları sürekli olabilir.

3. Az kullanılan ancak son yıllarda yaygınlaşan bir yöntemdir.

4. Model esnek ve kolay yorumlanabilir.

5. Varsayım(normallik, ortak varyans) bozulmalarında diskriminant ve çapraz tablolara alternatif oluşturur. Çünkü normallik varsayımı yoktur.

Lojistik regresyon modelinin lineer regresyon modelinden ayrıldığı en önemli nokta, lineer regresyon modelinde yer alan bağımlı değişkenin sürekli, lojistik regresyon modelinde bulunan bağımlı değişkenin ise 0 ve 1 gibi ikili(binary) ya da ikiden çok düzey içeren(polychotomous) kesikli bir değişken olmasıdır. Lojistik regresyon, bağımsız değişken veya değişkenlerin bağımlı değişken üzerinde yapmış olduğu etkileri, bağımlı değişkenin iki kategorisinden birinin gerçekleşme olasılığının diğer kategorinin gerçekleşme olasılılığıyla karşılaştırılmasından faydalanarak analiz eder[16].

Lojistik Regresyon Analizinin kullanım amacı, istatistikte kullanılan diğer model yapılandırma teknikleri ile aynıdır. En az değişkeni kullanarak en iyi uyuma sahip olacak şekilde bağımlı ile bağımsız değişkenler arasındaki ilişkiyi tanımlayabilmek ve amaca yönelik kabul edilebilir bir model kurmaktır[6].

Lojistik Regresyon, kullanım kolaylığı ve sayısal verilerle basit bir şekilde yorumlanabilir olması nedeniyle ön plana çıkmış ve son zamanlarda sıklıkla kullanılan yöntem durumuna gelmiştir.

Bu çalışmanın ikinci bölümünde lojistik regresyon modeli hakkında genel bilgiler verildikten sonra lineer regresyon modeliyle arasındaki ilişki ve lojistik regresyonun tercih edilme nedenleri üzerinde durulacaktır. Çalışmanın üçüncü bölümünde tek değişkenli ve çok değişkenli lojistik regresyon modellerinde parametre tahmin yöntemleri, katsayıların önem testleri, model katsayılarının yorumlanması, model yapılandırma stratejileri ve model uyumunun belirlenmesinden sonra modelin değerlendirilmesi detaylı bir şekilde anlatılacaktır. Çalışmanın dördüncü bölümünde modelde yer alması gereken değişkenlerin, değişken seçim yöntemlerine dayalı olarak tespit edilmesi ve bu amaç için çalışmalarda kullanılabilen bootstrap örnekleme yönteminin temeli üzerinde durulacaktır. Çalışmanın son bölümünde ise hipertansiyon şikayeti olan bireyler üzerinde çalışılmış ve yeniden örnekleme tekniklerinden olan bootstrap yöntemi, lojistik regresyon analizi ile kullanılmıştır. İlgili tahminler yapılarak en iyi model kurulmuştur.

2. GENEL BİLGİLER

2.1. Lojistik Regresyonun Tarihsel Gelişimi Ve Kullanım Alanları

Lojistik regresyon modelleri, son yıllarda biyoloji, tıp, ziraat, ekonomi, veterinerlik ve taşıma sahalarında yaygın olarak kullanılmaktadır.

Lojistik modelin biyolojik deneylerin analizi için kullanımı ilk olarak Berkson (1944) tarafından önerilmiş, Cox (1970) bu modeli gözden geçirerek çeşitli uygulamalarını yapmıştır. Özet gelişmeler ise ilk kez Anderson (1979, 1983) tarafından verilmiştir[6]. Lojistik regresyon modellerinin yaygın bir şekilde kullanılır hale gelmesi, katsayı tahmin yöntemlerinin geliştirilmesi ve lojistik regresyon modellerinin daha ayrıntılı incelenmesine sebep olmuştur. Carnfield (1962), lojistik regresyondaki katsayı tahmin işlemlerinde diskriminant fonksiyonu yaklaşımını ilk kez kullanarak popüler hale getirmiştir[8]. Ayrıca lojistik modelin uyumu ile ilgili birçok çalışma yapılmıştır. Bunlar arasında Aranda-Ordaz (1981) ve Johnson (1985) tarafından yapılan çalışmalar en önemlileridir. Pregibon (1981) iki grup lojistik modelde etkin (influential), aykırı (outlier) gözlemleri ve belirleme ölçülerini (diagnostic), Lesaffre (1986), Lesaffre ve Albert (1989) ise çoklu grup lojistik modellerde etkin ve aykırı gözlemlerle belirleme ölçütlerini incelemişlerdir[18]. Lee (1984) basit dönüşümlü (cross-over) deneme planları için lineer lojistik modeller üzerinde durmuştur. Bonney (1987) lojistik regresyon modelinin kullanımı ve geliştirilmesi üzerinde çalışmıştır. Roberts ve ark. (1987) lojistik regresyonda standart ki-kare, olabilirlik oranı (G^2), "pseudo" en çok olabilirlik tahminleri, uyum mükemmelliği ve hipotez testleri üzerine çalışmalar yapmışlardır[8]. Houck(1988) simetrik olmayan verilere lojistik modelin uydurulmasında yetersizlik olduğunu ve bağımsız değişkenlere ait ölçeğin değiştirilmesi ile lojistik modelin simetrik olmayan verilere uyabileceğini göstermiştir. Ruiz-Velasco(1989) açıklayıcı değişkenlerin normal dağılıma sahip olduğu varsayımı altında, parametreler hakkındaki hipotezleri test etmek amacıyla lineer diskriminant analizinde lojistik regresyonun asimtotik etkisini hesaplamıştır. Ali ve Khan(1989) tek dereceli istatistiklerin fonksiyonel

momentleri ve log-lojistik dağılımdan ikinci derece istatistiklerin bölünmüş momentlerini elde etmişlerdir. Hosmer ve arkadaşları (1989) lineer olmayan modellerde en iyi alt grubun seçim yöntemleri üzerine çalışmışlar ve çok fazla zaman alan bir modelleme işlemi olduğundan dolayı paket programlarının kullanımını önermişlerdir. Duffy (1990) lojistik regresyonda hata terimlerinin dağılışı ve parametre değerlerinin gerçek değerlere yaklaşımını incelemiştir[23]. Alho (1990) balıkçılıkta lojistik regresyonun uygulanıp uygulanamayacağı üzerinde çalışmış, koşullu en çok olabilirlik yöntemini kullanarak olasılıkları tahmin etmiştir. Birey sayısının çok olduğu durumlarda daha tutarlı sonuçlar elde edilebileceğini göstermiştir. Başarır (1990) klinik veriler üzerinde ayrımsama sorunu ve çok değişkenli lojistik regresyon analizi üzerinde çalışmıştır. Corrol ve Wand(1990) lojistik regresyon parametrelerinin yarı parametrik tahminleri üzerinde çalışmışlar ve parametre tahminlerinin bulunmasında Kernel regresyon tekniğini kullanmışlardır. Morris ve Silk (1992) mısır bitkisinde kök büyümesinin hücresel büyümesini incelemişler ve zamana bağlı lojistik regresyon modelini geliştirmişlerdir. Hsu ve Leonard (1995) lojistik regresyon fonksiyonlarında Bayes tahminlerinin elde edilmesi işlemleri üzerine çalışmışlar ve lojistik regresyonda Monte Carlo dönüşümünün kullanılabileceğini göstermişlerdir[23]. Gardside ve Glueck (1995) insanlarda beslenme şekli, sigara ve alkol kullanımı, fiziksel aktivite gibi risk faktörlerinin kalp hastalığı üzerindeki etkilerini incelemiştir[18]. Heise ve Myers(1996) tek değişkenli lojistik regresyon için optimal deneme planları üzerinde çalışmışlar ve iki tip deneme planı sunmuşlardır. Bunlar Q-optimal ve D-optimal deneme planlarıdır. Q-optimal planın D-verimliliği ve D-optimal planın Q-verimliliği üzerinde durulmuştur. Elhan (1997) lojistik regresyon yöntemini kullanarak kroner arter kalp hastalığına etki eden risk faktörlerini incelemiştir. Akkaya ve Pazarlıoğlu (1998) lojistik regresyon modellerinin ekonomi alanında kullanımını örneklerle incelemişlerdir[23]. Cox ve ark. (1998) kardiovasküler hastalıklar ve hipertansiyon arasındaki ilişkiyi incelemişlerdir[19]. Poples ve ark. (1991), Buescher ve ark. (1993), Kloiber ve ark. (1996), kadınlarda düşük doğum ağırlığını etkileyen risk faktörlerini; Santos ve ark. (1998) kafein tüketimi ve düşük doğum ağırlığı arasındaki ilişkiyi, Sable ve Herman (1997) erken doğum ve düşük doğum ağırlığı arasındaki ilişkiyi incelemişlerdir[4].

Çizelge 2.1' de lojistik modelin çeşitli uygulamalarının kullanıldığı 801 tane çalışmanın konulara göre dağılımı verilmiştir[23].

Çizelge 2.1. Lojistik Modelin Kullanıldığı Çalışmaların Konulara Göre Dağılımı

ÇALIŞMA ALANLARI		**ÇALIŞMA SAYISI**	**TOPLAM**	
TIP	Genel	214 %26,7	262 %32,7	
	Halk Sağlığı	17 %2,1		
	Besleme-Diyet(İnsanlar)	31 %3,9		
TARIM (Hayvansal ve Bitkisel) ve VETERİNERLİK	Entomoloji	37 %4,6	73 %9,1	381 %47,6
	Fitopatoloji	36 %4,5		
	Sütçü Sığırlar	112 %14	225 %28,1	
	Etçi Sığırlar	8 %1		
	Mastitis	17 %2,1		
	Domuzlar	15 %1,9		
	Atlar	10 %1,2		
	Koyunlar	10 %1,2		
	Tavuklar	8 %1		
	Maymunlar	1 %0,1		
	Hindiler	1 %0,1		
	Arıcılık	3 %0,4		
	Geyikler	1 %0,1		
	Keçiler	3 %0,4		
	Kediler	5 %0,6		
	Köpekler	18 %2,2		
	Fareler	3 %0,4		
	Sindirim-Beslenme	5 %0,6		
	Su Ürünleri	5 %0,6		
	Tarla Bitkileri	61 %7,6	76 %9,5	
	Bahçe Bitkileri	4 %0,4		
	Bitki Besleme	11 %1,4		
	Genetik	7 %0,9	7 %0,9	
ÇEVRE-DOĞA-EKOLOJİ		51 %6,4	109 %13,5	
ORMANCILIK		42 %5,2		
GIDA		8 %0,1		
MİKROBİYOLOJİ		8 %0,1		
DİĞERLERİ	Eğitim-Sosyal Ekonomi	39 %4,9	49 %6,2	
	Turizm	3 %0,4		
	Jeoloji-Uzaktan Algılama	7 %0,9		
TOPLAM		801	%100	

Çizelgeden de görüldüğü gibi, biyolojik alanlarda lojistik modelin kullanımı oldukça yaygındır. İncelenen çalışmaların %32,7' si Tıp, %47,6' sı Tarım ve Veterinerlik,

%13,5’ i Çevre, Doğa, Ekoloji, Ormancılık, Gıda ve Mikrobiyoloji alanındadır. Diğer alanlarda lojistik modelin kullanımı ise %6,2’ dır[23].

2.2. Lojistik Regresyon Modeli

Regresyon yöntemleri bir ya da birden fazla açıklayıcı değişken ile sonuç değişkeni arasındaki ilişkiyi inceler. Genellikle sonuç değişkeni kesikli bir değer olup, iki ya da daha fazla olası değere bağlıdır.

Lojistik regresyon analizi sonuç değişkeninin ikili, üçlü ve çoklu değerler aldığı, açıklayıcı değişkenlerle sebep-sonuç ilişkisini inceleyen bir yöntemdir.

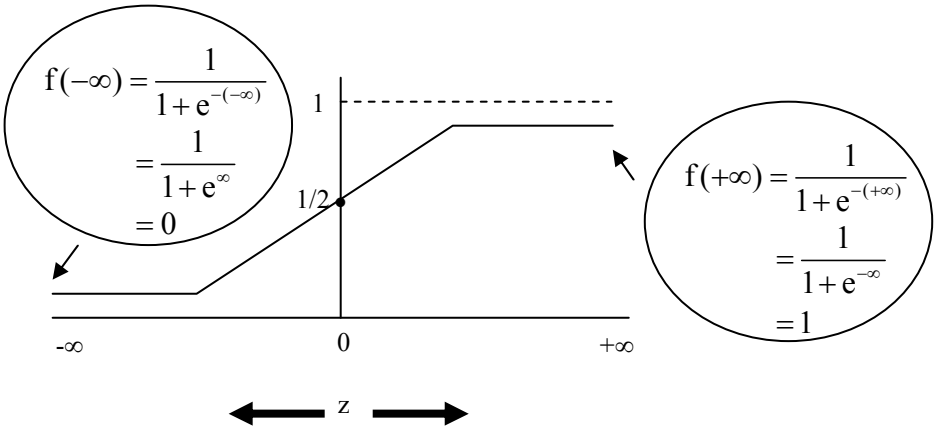
Bu yöntemde, açıklayıcı değişkenlerin bağımlı değişkenler üzerindeki etkileri olasılık olarak hesaplanarak risk faktörlerinin olasılık olarak belirlenmesi sağlanır[22].

Doğrusal regresyon analizinde bağımlı değişkenin değeri tahmin edilirken, lojistik regresyon analizinde bağımlı değişkenin alacağı değerlerden birinin gerçekleşme olasılığı tahmin edilir. Lojistik modelin kurulduğu, matematiksel formu oluşturan lojistik fonksiyonun f(z) olduğu varsayılırsa;

$$f(z) = \frac{1}{1 + e^{-z}}$$

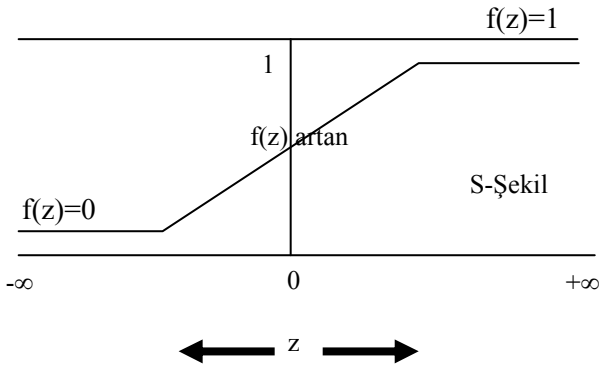
şeklinde tanımlanır.

Şekil 2.1 de z' nin tanım aralığının (-∞, +∞) arasında olduğu açıkça görülmektedir. z' nin değeri ne olursa olsun f(z) fonksiyonunun değişim aralığı 0 ile 1 arasındadır. Lojistik fonksiyonun tercih edilmesindeki önemli sebeplerden biri, f(z) lojistik fonksiyonun 0 ile 1 arasında bir değişim aralığına sahip olmasıdır. Çünkü model 0 ile 1 arasında yer alan herhangi bir olasılık üzerine kurulmuştur.



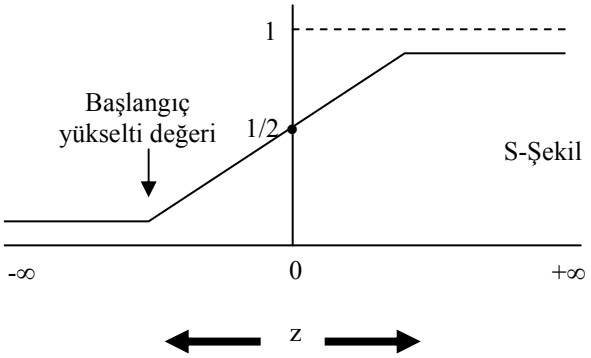
Şekil 2.1. Lojistik fonksiyonun tanım aralığı

Lojistik modelin tercih edilmesinin diğer bir nedeni de lojistik fonksiyonun biçimidir.



Şekil 2.2. Lojistik fonksiyonun şekli

Şekil 2.2' de görüldüğü gibi lojistik fonksiyon S şeklinde olup z değerinin artmasıyla f(z) sıfıra yaklaşır. Daha sonra 1'e doğru artmaya başlayarak z +∞'a yaklaştığında 1'e asimptot olur. Böylece z=0'da simetri elde edilir. Sonuç itibariyle S şeklinde bir eğri meydana gelir ve sürekli olan bu fonksiyon 0 ile 1 arasında değerler alır.



Şekil 2.3. Lojistik fonksiyonun başlangıç yükselti değeri

S şeklindeki lojistik fonksiyonda z, çeşitli risk faktörlerinin katılımını gösteren bir indeks olarak kabul edilirse f(z) de z değerindeki riski gösterir. Şekil 2.3' den de görüldüğü üzere, yükselti değerine kadar bireyin riski minimumdur. Sonra risk ortadaki z değerlerinde hızla artmakta ve z yeteri kadar arttığında 1 civarında kalmaktadır.

Bu yükselti değeri epidemiologlar tarafından, hastalık koşullarının değişikliğini belirtmek ve gerekli uygulamaları yapmak için ortaya atılmıştır. Başka bir ifadeyle, S şeklindeki lojistik model bir epidemiolojik araştırma sorusunun çok değişkenli doğal bir uygulaması olmaktadır[16].

Lojistik modelde sonuç değişkeninin kesikli olması nedeniyle açıklayıcı değişkenle olan ilişkisi saçılım grafiğinde açıkça görülmez. Bunun için sonuç değişkeni yerine f(z) olasılık değeri kullanılarak çizim yapılır. f(z) fonksiyonu sürekli olup Şekil 2.3' de görüldüğü gibidir.

Lojistik regresyon fonksiyonu, diğer regresyon fonksiyonları gibi bağımlı değişken ile bir veya daha fazla bağımsız değişkenler arasındaki ilişkiyi en iyi şekilde tanımlamak için kullanılan bir yöntemdir. Aynı zamanda bağımsız değişkenler olan açıklayıcı değişkenlerle, bağımlı değişkenler arasında kurulan modele ilişkin çıkarımlar, öngörülerin yapılmasında yardımcı olur. Böylece riskten etkilenme olasılığı tahmin edilir. $k = 1,..., p$ olmak üzere, p değişken sayısı iken model,

$z = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$

ile ifade edilir. α ve β_k bilinmeyen parametreleri temsil etmektedir. Böylece z ; α , β ve X ' lerin lineer toplamı şeklinde yazılabilir.

$$f(z) = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-\left(\alpha+\sum_{k=1}^{p}\beta_k X_k\right)}}$$

şeklindeki ifadeyle lojistik model oluşturmuş olmaktadır[16].

Regresyon problemlerindeki amaç, verilen bir bağımsız değişken değerine bağlı olarak bağımlı (sonuç) değişkeninin ortalama değerini bulmaktır. Bu değer koşullu ortalama olarak adlandırılır ve E(Y/x) ile gösterilir. Burada x bağımsız değişkeni, Y' de bağımlı değişkeni gösterdiğine göre E(Y/x) ifadesi "x değeri verildiğinde, Y' in beklenen değeri" şeklinde okunur. Lineer regresyon analizinde, koşullu ortalamanın, x' in doğrusal bir denklemi olduğu varsayıldığında,

$E(Y/x) = \beta_0 + \beta_1 X$

şeklinde yazılır. Doğrusal regresyon analizinde herhangi bir x değeri için, E(Y/x), -∞ ile +∞ arasında değişen değerler almaktadır. Buna karşılık bağımlı değişken iki düzey içeren kesikli bir değişkenden oluşuyorsa, yani lojistik regresyon fonksiyonu ise x' deki her birim değişme sonucunda E(Y/x)'de oluşan değişiklik, koşullu ortalama 0'a ya da 1'e yaklaştıkça azalır.

Lojistik dağılım kullanıldığında gösterimi kolaylaştırmak için, x bilindiğinde Y' in koşullu ortalamasını göstermek için $\pi(x)=E(Y/x)$ ifadesi kullanılır ve lojistik regresyon modeli,

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \tag{2.1}$$

ile ifade edilir. Koşullu ortalamanın 0 ile 1 arasında bir olasılık değerini alması için bağımlı değişken π(x) 'i $(-\infty,+\infty)$ aralığında tanımlı hale getirebilecek bir dönüşüm uygulamak gerekir. Eş. 2.1' de verilen lojistik regresyon modeli üzerinde yapılacak $\pi(x)/[1-\pi(x)]$ dönüşümü bağımlı değişkenin sınırlarını $(0,\infty)$ yapar. Bağımlı değişkenin sınırlarını $(-\infty,+\infty)$ yapmak için ise $\pi(x)/[1-\pi(x)]$ oranının logaritması alınır. Bu sayede elde edilmiş olan yeni bağımlı değişken, bağımsız değişkenin lineer bir fonksiyonu olur. π(x) 'i $(-\infty,+\infty)$ aralığında tanımlı hale getiren bu dönüşüm "lojit dönüşüm" olarak adlandırılır ve π(x) cinsinden,

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x \tag{2.2}$$

şeklinde gösterilir. $\pi(x)/[1-\pi(x)]$ oranı odds olarak adlandırılır.

$g(x)$, lineer regresyon modelinde istenen çoğu özelliği taşır. Lojit $g(x)$ parametreleri (β_0,β_1) bakımından lineer ve x' in aldığı değerlere bağlı olarak $(-\infty,+\infty)$ aralığındadır.

Lojistik regresyon modelinde sonuç değişkeninin bir gözlemi $y = E(Y/x)+\varepsilon$ şeklinde gösterilir. ε değeri hata terimi olarak adlandırılır ve gözlemin koşullu ortalamadan sapma miktarını ifade eder. ε , 0 ortalama ve bağımsız değişkenin her bir düzeyi için sabit bir varyansla normal dağılır. $(\varepsilon \sim N(0,\sigma^2))$ [11].

Verilen x için sonuç değişkeninin koşullu dağılımı E(Y/x) ortalamasına ve sabit bir varyansa sahip bir normal dağılımdır. Fakat iki sonuçlu bağımlı değişken için durum farklıdır. Bu durumda verilen x için sonuç değişkenini Y = π(x)+ε diye ifade ederiz.

ε ' un mümkün olan iki değerden başka değer alamayacağı varsayılırsa;

$Y=1$ ise $\pi(x)$ olasılıkla $\varepsilon=1-\pi(x)$,

$Y=0$ ise $\pi(x)$ olasılıkla $\varepsilon=-\pi(x)$ değerini alır.

Böylece ε , 0 ortalamalı ve π(x)[1-π(x)] varyanslı bir dağılıma sahip olur. Yani;

$$E(\varepsilon)=[(1-\pi(x))\times\pi(x)]+[-\pi(x)\times(1-\pi(x))]=0$$

$$\begin{aligned}V(\varepsilon)&=E(\varepsilon^2)-[E(\varepsilon)]^2=E(\varepsilon^2)\\&=[(1-\pi(x))^2\times\pi(x)]+[(-\pi(x))^2\times(1-\pi(x))]\\&=[(1-2\times\pi(x)+\pi(x)^2)\times\pi(x)]+[\pi(x)^2\times(1-\pi(x))]\\&=[\pi(x)-\pi(x)^2]\\&=\pi(x)\times[1-\pi(x)]\end{aligned}$$

şeklindedir.

Sonuç değişkeni Y ' in koşullu dağılımı, $\pi(x)=E(Y/x)$ koşullu dağılımına göre bir binom dağılımıdır. Özet olarak, sonuç değişkeninin iki düzeyli olması halinde regresyon analizinde:

1) Regresyon eşitliğindeki koşullu ortalama 0 ve 1 arasında bir değer olmalıdır.

2) Normal dağılım değil de binom dağılımı hatanın dağılımını tanımlar ve analiz bunun üzerine kuruludur.

3) Doğrusal regresyonda kullanılan ilkeler, lojistik regresyon analizinde de yol göstericidir[11].

2.3. Lojistik Regresyonun Lineer Regresyon İle İlişkisi

Model kurulumunda en sık kullanılan yöntem, sonuç değişkeni sürekli olan doğrusal regresyon modelidir. Lojistik regresyonda, doğrusal regresyon analizinde olduğu gibi bazı değişken değerleri göz önüne alınarak tahmin yapılmaya çalışılır. Fakat bu iki yöntem arasında üç önemli fark vardır:

a) Doğrusal regresyon analizinde tahmin edilecek bağımlı değişken sürekli iken, lojistik regresyon analizinde bağımlı değişkenler kesikli bir değer almaktadır.

b) Doğrusal regresyon analizinde bağımlı değişkenin değeri tahmin edilirken, lojistik regresyon analizinde ise bağımlı değişkenin alabileceği değerlerden birinin gerçekleşme olasılığı tahmin edilir.

c) Doğrusal regresyon analizinde bağımsız değişkenin çok değişkenli normal dağılım göstermesi şartı aranırken lojistik regresyon analizinde böyle bir şart yoktur[6].

Bu sebeplerden dolayı, lojistik regresyon modellemesinde lineer(doğrusal) regresyon analizinde kullanılan yöntemlerden yararlanılacaktır.

2.4. Lojistik Regresyonun Tercih Edilme Nedenleri

Lojistik regresyon binary (ikili) verilerin gösterimi için daha basit ve uygun olduğu için kullanılmaktadır. İkili veriler için doğrusal regresyon kullanıldığı zaman, üç problem ortaya çıkar:

1. problem, hata teriminin varyansı sabit değildir.

2. problem, hata terimi normal dağılmamıştır.

3. problem, tahminin 0 ve 1 arasında olma zorunluluğu yoktur.

Yukarıda bahsedilen problemlerden 1. problem ağırlıklı en küçük kareler yöntemi kullanılarak çözülebilir. 2. problem; örneklem kümesi çok büyük olduğunda EKK yöntemi kullanılarak hata teriminin normal dağılması sağlanabilir.
Ancak 3. problemin üstesinden gelinememektedir[21].

Diğer yöntemlere alternatif olarak kullanılan lojistik regresyon analizinin güncel olmasının nedenleri şöyle özetlenebilmektedir.

a) Sonuç değişkeni kesikli iken açıklayıcı değişkenlerin hem sürekli hem de kesikli olma durumlarında uygulanabilmektedir.

b) Lojistik modelin parametreleri epidemioloji de yapılan ölçümlere benzediği için yorumları kolay olmaktadır. Epidemiolojide katsayıların exponansiyeli hastalık riski olarak yorumlanır.

c) Lojistik modelin parametre sayısı, doğrusal regresyon modeli ve diskriminant fonksiyonu ile aynı olmaktadır.

d) Lojistik modele dayalı analizler için standart paket programlar vardır.

e) Açıklayıcı değişkenlerin olasılık fonksiyonlarının dağılımı üzerinde kısıt olmaması (yarı parametrik) nedeni ile çeşitli testler uygulanabilmektedir.

Epidemioloji ve diğer medikal uygulamaların yanı sıra deneysel verilerin analizinde, askeri konularda, meteorolojide, ziraatte, taşımacılıkta, ekonomi v.b. alanlarda sıkça kullanılan lojistik regresyon analizi farklı varsayımlar durumunda aynı lojistik formülasyona götürdüğü için varsayım bozulmalarına karşı daha güçlü bir yöntemdir[5].

Lojistik regresyon, lineer regresyona göre karışık görülebilir. Ancak, çoğu istatistik yazılımı lojistik regresyonu, lineer regresyondan daha basit bir şekilde kullanım olanağı vermektedir.

Sonuç olarak, yüzde modellerinde lojistik regresyonun lineer regresyondan daha iyi bir yaklaşım olduğu söylenebilir. Lojistik regresyonda her zaman istatistiksel olarak anlamlı tahminlerde bulunmanın doğal bir avantajı vardır. Ve çoğu durumda gözleme yakın bir tahmin elde edilir.

Bazı özel koşullar altında lineer modellerin sonuç değişkenini tanımlamak için yeterli olmasına rağmen, gözlemlerin yüzde şeklinde gösterildiği durumlarda lojistik regresyonun kullanılması gerektiği düşünülmektedir[26].

3. LOJİSTİK REGRESYON MODELİNİN KURULMASI VE ANALİZİ

Lojistik regresyon analizinde değişkenler arasında çoklu bağlantı olmamalıdır. Bunun için herhangi bir değişkenin diğer değişkenlerin lineer bileşimi şeklinde yazılmaması gerekir. Böylece analizde bazı değişkenlerin toplamı ya da ortalamaları orijinalleriyle aynı anda yeni bir değişken olarak kullanılmamalıdır. Daha ayrıntılı düşünülürse, bu tür yeni bir değişkenin fonksiyona ilave bir bilgi katmayacağı görülebilir. Bu problem bağlantı ya da çoklu bağlantı adını alır. Gözlem sayısının azlığı bu sorunun ortaya çıkması olasılığını arttırır.

Çoklu bağlantı regresyon analizinde regresyon katsayılarının yanlış tahmin edilmesine, katsayıların standart hatalarının artmasına, t-testinin geçersiz olmasına ve modelin tahmin gücünün azalmasına sebebiyet verebilir. Lojistik regresyon analizinde de benzer sorunlara yol açabilir. Bu yüzden eğer varsa, çoklu bağlantı durumunun tespit edilmesi ve gerekli düzeltme işlemlerinin yapılması gerekmektedir[10].

3.1. Parametre Tahmin Yöntemleri

$i = 1,...,n$ olmak üzere, $\left(x_i, y_i\right)$gibi n tane bağımsız gözlem çiftinin olduğunu varsayılsın. y_i iki düzeyli sonuç değişkeni, x_i i' inci birim için bağımsız değişkenin aldığı değerdir. Sonuç değişkeni için 0 ve 1 kodlarının sırasıyla belirli bir karakteristik yokluğu ya da varlığı temsil ettiği varsayılsın. Eş. 1.1' de verilen lojistik regresyon modelini tahmin edebilmek için bilinmeyen β_0 ve β_1 parametreleri tahmin edilmelidir. Parametreler tahmin edildikten sonra, parametrelerin modele katkısı ve modelin anlamlılığı test edilir. Lojistik modelde parametrelerin tahmin edilmesi için çeşitli yöntemler ortaya atılmıştır. Bu yöntemlere aşağıda kısaca değinilmiştir.

3.1.1. En küçük kareler yöntemi(EKK)

Doğrusal regresyonda bilinmeyen parametreleri bulmak için sıklıkla kullanılan yöntem EKK yöntemidir. Bu yöntemle modele göre tahmin edilen Y değerlerinin gözlemlenen değerlerden sapmalarının karesini minimize edecek β_0 ve β_1 değerleri elde edilir. Bağımlı değişkenin kesikli olması durumunda EKK yöntemi söz konusu varsayımları sağlamaz[20].

$y = E(y/x) + \varepsilon$ şeklindeki lineer regresyon modelinde EKK yöntemi istenilen istatistiksel özelliklere sahip tahmin edicileri sağlamaktadır. Bu özellikler: hataların normal dağılması, ortalamanın sıfır olması ve varyansın bağımsız değişkenin her bir seviyesi için sabit kalmasıdır. Fakat lojistik regresyon analizinde bu varsayımlar geçerli değildir. Bu sebeplerden dolayı lojistik regresyon analizinde EKK yöntemi uygulanmamaktadır. Lojistik regresyon analizinde EKK yönteminin yerine EÇOB yöntemini kullanılmaktadır.

3.1.2. En çok olabilirlik (EÇOB) yöntemi

Lojistik modelin parametrelerinin tahmini için en sık kullanılan yöntem olan EÇOB yöntemini(maximum likelihood method) ilk kullanan Krog (1916)' dur. Bu yöntem gözlenen veri kümesini elde etmenin olasılığını en büyük yapan bilinmeyen parametrelerin değerlerinin tahminlerini verir. Bu metodu uygulamaya geçmeden önce EÇOB fonksiyonu oluşturulmalıdır. β_0 ve β_1 gibi parametrelerin EÇOB tahmin edicileri, fonksiyonu en büyük yapan değerleri bulacak şekilde seçilir.

EÇOB fonksiyonu lojistik regresyon modelinde şöyle bulunur; lojistik regresyon modeli için EÇOB fonksiyonunun elde edilmesinde Y bağımlı değişkeninin 0 ve 1 değerlerini aldığı varsayılırsa, $\pi(x)$ ifadesi bağımsız değişkenin değeri verildiğinde Y' nin 1'e eşit olma koşullu olasılığını verir. $1-\pi(x)$ ifadesi ise bağımsız değişkenin değeri verildiğinde Y' nin 0'a eşit olma koşullu olasılığını gösterir. Yani (x_i, y_i)' nin $y_i = 1$ olduğunda olabilirlik fonksiyonuna katkısı $\pi(x_i)$ kadar, $y_i = 0$

olduğunda olabilirlik fonksiyonuna katkısı ise $1-\pi(x_i)$ kadardır. (x_i, y_i)' nin olabilirlik fonksiyonuna katkısı aşağıda gösterildiği gibidir.

$$\zeta(x_i) = \pi(x)^{y_i}\left[1-\pi(x)\right]^{1-y_i} \tag{3.1}$$

Gözlemlerin birbirinden bağımsız olduğu varsayılırsa, Eş. 3.1'de verilen terimlerin çarpımı olabilirlik fonksiyonunu verir.

$$l(\beta) = \prod_{i=1}^{n} \zeta(x_i) \tag{3.2}$$

EÇOB' in temel ilkesi, β tahminlerinin Eş. 3.2'yi maksimum yapmasıdır. Eş. 3.2'nin logaritmasıyla çalışmak matematiksel olarak daha kolay olacağından log-olabilirlik fonksiyonu şu şekilde tanımlanır.

$$L(\beta) = \ln\left(l(\beta)\right) = \sum_{i=1}^{n}\left\{y_i \ln\left[\pi(x_i)\right] + (1-y_i)\ln\left[1-\pi(x_i)\right]\right\} \tag{3.3}$$

Eş. 3.3 ile verilmiş olan log-olabilirlik fonksiyonunu maksimum yapan β değerlerini bulabilmek için β_0 ve β_1' e göre türevi alınarak 0' a eşitlenmek suretiyle en çok olabilirlik eşitlikleri,

$$\sum_{i=1}^{n}\left[y_i - \pi(x_i)\right] = 0 \tag{3.4}$$

ve

$$\sum_{i=1}^{n} x_i\left(y_i - \pi(x_i)\right) = 0 \tag{3.5}$$

elde edilir. Bir lojistik modelde, parametrelerin tahmininde kullanılabilecek iki alternatif EÇOB tahmin tekniği bulunmaktadır. Bunlar, koşulsuz yöntem ve koşullu yöntemdir. Modelde tahmin edilecek parametrelerin sayısı, gözlem sayısından küçükse koşulsuz en çok olabilirlik tahmin yöntemi, büyükse koşullu en çok olabilirlik tahmin yöntemi kullanılır. Uygulamada, genellikle gözlem sayısı tahmin edilmek istenen parametre sayısından büyük olmaktadır, bu sebeple koşulsuz en çok olabilirlik tahmin metodu kullanılmaktadır. Koşullu en çok olabilirlik tahmin yönteminin kullanılması gereken bir yerde koşulsuz en çok olabilirlik tahmin yönteminin kullanılması yanlı sonuçlar vermektedir. Koşullu en çok olabilirlik ise her zaman uygundur ancak çok fazla matematiksel işlem gerektirir ve istatistik paket programlarında çalıştırılması uzun sürdüğünden pek kullanılmaz. Bu sebeple en geniş kullanım sahası koşulsuz en çok olabilirlik tahmin yöntemidir ve en çok olabilirlik denilince bu yöntem akla gelmektedir.

Doğrusal regresyonda olabilirlik eşitlikleri kolay çözülebilen doğrusal denklemlerdir, fakat lojistik regresyon analizinde bu ifadeler β_0 ve β_1'e göre doğrusal olmayan, üstel denklemler olduklarından bu denklemlerin çözümü için özel metotlar gerekmektedir. Bu problem iteratif olup istatistik paket programlar ile çözümlenebilir. İki denklemin çözümünden elde edilen β değerlerine EÇOB tahmin edicileri denir ve $\hat{\beta}$ ile gösterilir[2]. Genellikle "^" sembolü EÇOB tahminini göstermektedir. $\hat{\pi}(x_i)$ ifadesi $\pi(x_i)$ in EÇOB tahmin edicisidir. Bu nicelik, verilen $x = x_i$ değeri için Y' in 1'e eşit olma koşullu olasılığının tahminini vermektedir. Buradan Eş. 3.4,

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\pi}(x_i)$$

şeklinde yazabilir[11].

3.1.3. Yeniden ağırlıklandırılmış iteratif en küçük kareler yöntemi

Lojistik regresyonda parametre tahmininde kullanılan diğer bir yöntem de iteratif ağırlıklandırılmış en küçük kareler yöntemidir. Bu yöntem, lojistik regresyon analizinde olduğu gibi hata terimlerine ilişkin varyansların eşit olmadığı durumda doğru tahmin sağlar.

$j = 1,2,...,J$ olmak üzere gruplandırılmış verilerde J grubun her birinde n_j denemeden r_j başarı elde edildiğinde başarı oranı $P_j = \frac{r_j}{n_j}$ şeklinde gösterilir. Varyansı ise; $\sigma_j^2 = \mathrm{Var}(r_j / p_j) = \mathrm{Var}(P_j) = P_j.\left(\frac{1-P_j}{n_j}\right)$ olduğundan her binom dağılımlı gözlem için varyans değişmektedir. Bu durumda $lojit\left(r_j / n_j\right)$' nin açıklayıcı değişkenler üzerinde $w_j = \frac{n_j}{P_j.\left(1-P_j\right)}$ ağırlığı ile ağırlıklandırılmış regresyonu uygulanır[24]. Fakat w_j ağırlık değerleri de P_j'nin bir fonksiyonu olduğu için en küçük kareler yöntemi iteratif olarak uygulanarak ağırlık değerleri her adımda yeniden elde edilerek çözüme ulaşılır.

3.1.4. Minimum logit ki-kare yöntemi

Ağırlıklı en küçük kareler tahmin yönteminin özel bir biçimidir. Berkson' un (1955) geliştirdiği bu yöntemde, $2 \times J$ çapraz tablolarındaki beklenen ve gözlenen lojit değerler arasındaki farktan yararlanılmaktadır. Bu yöntem tekrarlı veriler olması durumunda kullanılmaktadır.

Veriler j grupta tekrar edildiğinde ve her grupta tekrar sayısı çok olduğunda, katsayıların tahmin edicileri ağırlıklı en küçük kareler yöntemi ile elde edilebilmektedir.

Bir önceki yöntemde (yeniden ağırlıklandırılmış iteratif en küçük kareler yöntemi) değinilen P_j başarı olasılığı, lojistik fonksiyon eşitliğinde tanımlandığı gibidir. P_j olasılığı üzerinde yapılan lojit dönüşüm sonuç değişkenini oluşturmaktadır. Tahmin edicide kullanılan ağırlık değerleri $n_j P_j \left(1 - P_j\right)$ ile elde edilmektedir. Yöntem, sonuç değişkeninin açıklayıcı değişkenler üzerindeki ağırlık değeri olarak tanımlanan lojit değer ile ağırlıklandırılmış regresyondan en küçük kareler tahminlerini elde etmeye dayanmaktadır. Buradan tek adımda bulunan ağırlıklı en küçük kareler tahminleri minimum lojit ki-kare tahminleri adını almaktadır.

Olasılık değerinin 0 ya da 1 olduğu durumda lojit değeri tanımlı olmayacağı için P_j yerine $P_j + 1/2n_j$ değerinin alındığı ayarlanmış lojit ki-kare yöntemi kullanılmaktadır[5].

Kısaca değinilen dört tahmin yöntemi dışında kullanılan bazı tahmin yöntemleri vardır. Bunlardan en çok bilinenleri, iteratif olmayan en küçük kareler yöntemi ile diskriminant fonksiyonuna dayalı tahmin yöntemidir. Ancak bu yöntemlere çok özel durumlarda kullanılmaları nedeniyle bu çalışmada ele alınmamıştır.

3.1.5. Tahmin yöntemlerinin karşılaştırılması

EÇOB yöntemi her zaman tutarlı, etkin ve yeterli tahminler vermekte, ancak bu tahminler her zaman yansız olmamaktadır[25]. Yansızlık ve normal dağılımlılık asimptotik bir özelliktir. Doğrusal olasılık modeli parametrelerinin ağırlıklı en küçük kareler tahmini ile lojistik modelin EÇOB tahmini, varsayımlar sağlandığı sürece benzer istatistiksel özelliklere sahiptir. Tek farklılık, EÇOB yönteminde fonksiyonun doğrusal olmaması nedeniyle iteratif çözümün gerekli olmasıdır. Öte yandan minimum lojit ki-kare yönteminden de asimptotik olarak etkin ve yeterli tahmin ediciler elde edilmektedir.

Sonuç olarak, nokta tahmini için minimum lojit ki-kare yönteminin, çıkarsama için ise EÇOB yönteminin kullanılması önerilmektedir[25]. Bu arada (bazı

sağlam(robust) tahmin yöntemleri) Bayes ve Kernel tahmin yöntemleri de bu amaçla kullanılmaktadır[25].

3.2. Tek Değişkenli Lojistik Regresyon Modelinde Katsayıların Önem Testinin Yapılması

Bir veri kümesinin modellenmesi, uyum ve test işlemlerinden daha zor ve zahmetlidir. Parametre tahminleri yapıldıktan sonra, modeldeki değişkenin önemliliği araştırılır. Tek değişkenli lojistik regresyon modelinde parametre tahminlerinin önem testi yapılırken, genel olarak istatistiksel hipotezlerin yardımıyla modeldeki bağımsız değişkenlerin sonuç değişkeni ile arasındaki ilişkinin önemliliği test edilir.

Modeldeki bir bağımsız değişkenin parametre tahmininin önem testi için " bağımsız değişkeni içeren model, bağımsız değişkeni içermeyen modeldekinden daha detaylı bilgi veriyor mu?" sorusunun araştırılması gerekmektedir. Bu soruya cevap bulabilmek için sonuç değişkeninin gözlenen değerlerini, her iki modelden elde edilmiş tahmin edilen değerlerle karşılaştırmak gerekmektedir. Eğer değişkenli modelin tahmin edilen değerleri değişkeni kapsamayan modelden daha iyi ise inceleme yapılan değişkenin model için önemli(anlamlı) olduğu sonucuna varılır. Burada gözlenen ve tahmin edilen değerlerini karşılaştırmak için kullanılan matematiksel fonksiyon problemden probleme değişiklik gösterir.

Doğrusal regresyon modelinde değişkenlerin önem kontrolü için geçerli olan genel yöntem lojistik regresyon için de temel bir yaklaşım oluşturulur. Bu iki yaklaşımın karşılaştırılması, sürekli ve iki sonuçlu yanıt değişkenlerinin modellenmesi arasındaki farklılıkları belirler.

Lineer regresyonda eğim katsayısının önemine karar vermek için ilk olarak “Varyans Tablosunun Analizi” yapılır. Bu tablo, genel kareler toplamını kendi içinde iki parçaya ayrılır:

1) SSE (residual sum-of-squares) regresyon doğrusunun etrafındaki sapmaların kareleri toplamı veya artık kareler toplamı

2) SSR (regression sum-of-squares) bağımlı değişkenin ortalaması etrafındaki sapmaların kareleri toplamı veya regresyon kareler toplamı.

Bu sadece iki modeldeki gözlenen değerlerin tahmin edilen değerlerle karşılaştırılmasının bir yoludur. Lineer regresyonda gözlenen ve tahmin edilen değerlerin karşılaştırılması, ikisi arasındaki uzaklığın karesine dayanır. Eğer i'ninci birey için y_i' nin gözlenen değeri ve $\hat{y}_i$' in de tahmin edilen değeri gösterdiği varsayılırsa, bu karşılaştırma için kullanılan istatistik,

$$SSE = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2$$

şeklindedir.

Modelde bağımsız değişken olmadığı zaman tek parametre β_0 olur ve $\hat{\beta}_0$ yanıt değişkeninin ortalaması olan $\bar{y}$'ya eşittir. Yani $\hat{y}_i = \bar{y}$ şeklinde yazılabilir ve SSE toplam varyansa eşit olur. Modele bağımsız değişkeni eklediğimizde, SSE' de meydana gelen her düşüş, bağımsız değişken için eğim katsayısının sıfır olmamasından kaynaklanacak ve bu da SSR ile gösterilecektir. Buna göre kullanılan istatistik aşağıdaki gibi olacaktır:

$$SSR = \left[\sum_{i=1}^{n} \left(y_i - \bar{y}_i \right)^2 \right] - \left[\sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 \right]$$

Doğrusal regresyonda ilgilenilen durum SSR' nin büyüklüğü ile ilgilidir. SSR (regresyon kareler toplamı) değerinin büyük olması bağımsız değişkenin modele katkısının önemli olduğunu, küçük olması ise bu bağımsız değişkenin yanıt değişkenini tahmin etmemizde yararlı olmadığını gösterir.

Lojistik regresyondaki temel prensip:"değişken içeren ve içermeyen modellerden elde edilen tahmin değerleri ile yanıt değişkeninin gözlenen değerlerinin karşılaştırılmasıdır"[12]. Katsayıların önem testi üç farklı yöntemle yapılabilmektedir. Bunlar; Olabilirlik oran(likelihood ratio) testi, Wald testi ve Skor (score) testidir.

3.2.1. Olabilirlik oran testi

Lojistik regresyon analizinde bağımsız değişkenler ile sonuç değişkeni arasındaki ilişkiyi karşılaştırma işlemi logaritmik olabilirlik fonksiyonu ile yapılır. Bu karşılaştırmayı anlayabilmenin bir yolu, doymuş modelden elde edilen tahmin değerleri, sonuç değişkeninin gözlenen değerleri olarak kabul etmektir. Doymuş model veri sayısı kadar parametre içeren modeldir.

Olabilirlik fonksiyonunu kullanarak gözlenen değerlerle tahmin edilen değerlerin karşılaştırılması aşağıdaki şekilde tanımlanır:

$$\mathrm{D} = -2\ln\left[\frac{\textit{Modelin olabilirliği}}{\textit{Doymuş modelin olabilirliği}}\right] \tag{3.6}$$

Eş. 3.6' da $\left[\frac{\textit{Modelin olabilirliği}}{\textit{Doymuş modelin olabilirliği}}\right]$ ifadesi olabilirlik oranı(likelihood ratio) olarak adlandırılır. Elde edilen bu değer hipotez testi için kullanılır ve böyle bir teste olabilirlik oran testi adı verilir. Eş. 3.6 ile verilen olabilirlik oranı Eş. 3.3' deki log olabilirlik fonksiyonu cinsinden yazılacak olursa,

$$D = -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\hat{\pi}_i}{y_i}\right) + \left(1-y_i\right)\ln\left(\frac{1-\hat{\pi}_i}{1-y_i}\right)\right] \tag{3.7}$$

elde edilir. Burada $\hat{\pi}_i = \hat{\pi}\left(x_i\right)$ şeklinde de yazılabilmektedir.

Eş. 3.7' deki D istatistiği bazı yazarlar tarafından sapma (deviance) istatistiği olarak adlandırılır ve uyum iyiliğine karar verirken bazı yaklaşımlar için önemli bir rol oynar[19]. Lojistik regresyon için sapma (deviance), doğrusal regresyondaki artık kareler toplamı ile aynıdır. Eş. 3.7' de verilen sapma, lineer regresyon için hesaplanırsa SSE' ye eşit olur.

Bağımsız bir değişkenin önemini araştırırken, denklemde bağımsız değişkenin olduğu ve olmadığı iki D değeri hesaplanır ve karşılaştırılır. D' deki değişim aşağıdaki gibidir.

G = D (değişkensiz model için) – D (değişkenli model için)

G' yi hesaplamak için farkı alınacak olan D değerlerinin yukarıda belirtilmiş her iki durumu için de doymuş modelin olabilirlikleri aynı olduğundan, G istatistiği;

$$\mathrm{G} = -2\ln\left[\frac{\text{Değişkensiz modelin olabilirliği}}{\text{Değişkenli modelin olabilirliği}}\right]$$

olur. Tek bağımsız değişkenli özel durumda, değişkenin modelde olmadığında β_0'ın EÇOB tahmini $\ln\left(\frac{n_1 = \sum y_i}{n_0 = \sum\left(1 - y_i\right)}\right)$' dir. Buradan G istatistiği de;

$$G = -2\ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n}\hat{\pi}_i^{\,y_i}\left(1-\hat{\pi}_i\right)^{1-y_i}}\right]$$

ya da

$$G = 2\left\{\sum_{i=1}^{n}\left[y_i \ln\left(\hat{\pi}_i\right) + \left(1 - y_i\right)\ln\left(1 - \hat{\pi}_i\right)\right] - \left[n_1 \ln\left(n_1\right) + n_0 \ln\left(n_0\right) - n \ln(n)\right]\right\} \tag{3.8}$$

şeklinde ifade edilir.

Kurulan modelin sınanmasında gerçek modelin olabilirlik oranı ile tahmin edilmiş modelin olabilirlik oranı arasındaki farkın ki-kare dağılıp dağılmamasına bakılır. $\beta_1 = 0$ hipotezinin doğru olduğu varsayımı altında, G istatistiği 1 serbestlik derecesiyle ki-kare dağılımına sahip olacaktır[19].

Eş. 3.8' daki ilk terim değişken modeldeyken elde edilen log-olabilirlik değeridir ve denklemin kalan kısmı kolayca n_1 ve n_0 değerlerini denklemde yerine koyarak elde edilir.

Log-olabilirlik ve olabilirlik oran testi ile modele en son dahil edilen değişkenlerin önemlilik testleri yapılır. Tek bağımsız değişkenli durumlarda, ilk olarak yalnızca sabit terimi kapsayan model kurulur. Sonra sabit terimle birlikte bağımsız değişkeni kapsayan model kurulur. Bu durum yeni log-olabilirlikte artış sağlar. Olabilirlik oran testi, bu farkın -2 ile çarpımıyla elde edilir.

3.2.2. Wald testi

Log-olabilirlik oran testindeki varsayımlar Wald testi için de geçerliliğini korumaktadır. Eğim parametresinin en çok olabilirlik tahmini olan $\hat{\beta}_1$' nın kendi standart hatasına bölünmesi sonucu Wald testi elde edilir. Elde edilen oran standart normal dağılır(z). Lojistik regresyon modeli için Wald test istatistiği aşağıdaki gibidir.

$$W = \frac{\hat{\beta}_1}{SE\left(\hat{\beta}_1\right)} \sim Z_{(\alpha \text{ veya } \alpha/2)}$$

Ancak Hauck ve Donner (1977), Wald testinin performansını incelemişler ve bazı durumlarda bağımsız değişkenin katsayısı önemli olduğu halde anlaşılamayan bir nedenle bu testin ele alınan katsayıyı sıklıkla önemsiz olarak değerlendirildiğini bulmuşlardır. Bu nedenle Hauck ve Donner, katsayıların önem testinin yapılması için olabilirlik oran testinin kullanılmasını önermişlerdir[7]. Dikkatli bir şekilde incelenirse hem olabilirlik oran testi(G) hem de Wald testi(W) için, katsayıların en çok olabilirlik tahminlerinin bilinmesi gerekmektedir. Tek değişkenli durum için bu değerlerin hesaplanması zor değildir fakat değişken sayısının çok olduğu durumda hesaplamalar daha zor bir hal almaktadır.

3.2.3. Skor testi

Skor testi, Wald testinde değişken sayısı çok olduğunda artan hesap yükünü büyük ölçüde azaltan bir yöntemdir. En büyük avantajı bu olsa da, çoğu paket programda bulunmaması ise en büyük dezavantajıdır. Genel hatlarıyla matris hesaplamalarının kullanıldığı çok değişkenli bir test olan skor testi için gerekli varsayımlar log-olabilirlik oran testindeki varsayımlarla aynıdır. Skor testi için test istatistiği aşağıda verildiği gibidir.

$$ST = \frac{\sum_{i=1}^{n} x_i \left(y_i - \bar{y}\right)}{\sqrt{\bar{y}\left(1-\bar{y}\right)\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}}$$

Kısaca özetlemek gerekirse, tek değişkenli modeller için bir değişkenin katsayısının önem testini yapmak için geçerli yöntem lineer regresyon analizindeki yaklaşımlara benzemektedir. Fakat ikili sonuç değişkeninin olduğu durumlarda olabilirlik fonksiyonu kullanılmalıdır[11].

3.3. Çoklu(Çok Değişkenli) Lojistik Regresyon Modeli

Lojistik regresyon modelinin tek bağımsız değişkenden daha fazla değişken olması durumu için genelleştirilmiş haline "çok değişkenli durum" denir. Çoklu lojistik regresyon modeli için temel düşünce modeldeki katsayıların tahmini ve bu katsayıların anlamlılıklarının test edilmesidir. Bu da tek bağımsız değişkenli modele benzer şekilde yapılır.

$x' = (x_1, x_2, ..., x_p)$ vektörü ile gösterilen p tane bağımsız değişken kümesi ele alınsın. Şimdilik bu değişkenlerin her birinin sürekli olduğu varsayılsın. Sonuç değişkeninin $Y = 1$ şeklinde olduğu durumda koşullu olasılık $P(Y = 1/x) = \pi(x)$ şeklinde olur. Çoklu lojistik regresyon modelinin lojiti ise aşağıdaki denklemle ifade edildiği gibidir.

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p \tag{3.9}$$

Eş. 3.9' den sonuç değişkeninin koşullu olasılığı,

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}} \tag{3.10}$$

ile ifade edilir[11].

3.3.1. Çoklu Lojistik Regresyon Modelinin Kurulması

Birbirinden bağımsız n tane (x_i, y_i), $i = 1,2,...,n$ gözlem çiftinin var olduğu varsayılsın. Çoklu lojistik regresyon modeli kurulurken, tek değişkenli modelde olduğu gibi $\beta' = \left(\beta_1, \beta_2, \ldots \beta_p\right)$ parametre vektörü elde edilmelidir. β' vektörünü elde etmek için Eş. 3.3 de verilmiş olan EÇOB metodu kullanılır[11].

Olabilirlik fonksiyonu Eş. 3.2'de verilen tek değişkenli durum ile hemen hemen aynıdır. Tek fark $\pi(x)$' in Eş. 3.10'da verildiği gibi çok değişkenli durum için tanımlanmasıdır. Log-olabilirlik fonksiyonunun $p+1$ katsayıya göre türevi alınarak p+1 tane olabilirlik denklemi,

$$\sum_{i=1}^{n}\left[y_i-\pi\left(x_i\right)\right]=0$$

$$\sum_{i=1}^{n}x_{ij}\left[y_i-\pi\left(x_i\right)\right]=0,\qquad j=1,2,3..,p$$

şeklinde elde edilir.

$\hat{\beta}$,bu denklemlerin çözümünü gösteren tahmin vektörü olmak üzere, çoklu lojistik regresyon modeli için tahmin edilen değerler $\hat{\pi}(x_i)$ olup, Eş. 3.10'deki ifadede katsayı tahminleri olan $\hat{\beta}$ ve bağımsız değişkenlerin aldığı değerler olan x_i yerine konarak elde edilir. Tahmin edilen $\hat{\beta}$' ların varyans ve kovaryanslarının tahmin edililmesi için EÇOB yöntemi kullanılır. Kısacası elde edilen tahminler log-olabilirlik fonksiyonunun ikinci dereceden kısmî türevlerinden oluşmuş olan matris yardımıyla elde edilir. Kısmi türev j,u=0,1,2,...,p, $\pi=\pi(x_i)$ olmak üzere aşağıdaki biçimdedir:

$$\frac{\partial^2 L(\beta)}{\partial\beta_j^2}=-\sum_{i=1}^{n}x_{ij}^2\pi_i\left(1-\pi_i\right) \tag{3.11}$$

$$\frac{\partial^2 L(\beta)}{\partial\beta_j\partial\beta_u}=-\sum_{i=1}^{n}x_{ij}x_{iu}\pi_i\left(1-\pi_i\right) \tag{3.12}$$

Eş. 3.11 ve Eş. 3.12'da verilen terimlerin negatiflerini kapsayan $p+1\times p+1$ boyutunda elde edilen matris $\mathrm{I}(\beta)$ ile gösterilir ve "bilgi (information) matrisi" olarak adlandırılır. $\hat{\beta}$' ların varyans ve kovaryansları bilgi matrisinin tersinden elde edilir($\sigma^2(\hat{\beta})=\mathrm{I}^{-1}(\hat{\beta})$). Çok özel durumların dışında bu matrisi açık bir şekilde yazmak mümkün değildir. $\hat{\beta}_j$' nin varyansı $\sigma^2(\hat{\beta}_j)$ olmak üzere j. diagonal elemanından $\hat{\beta}_j$ ve $\hat{\beta}_u$' nun kovaryansı $\sigma(\hat{\beta}_j,\hat{\beta}_u)$ ie ifade edilir. $\hat{\beta}$' ların standart hataları,

$$\hat{SE}\left(\hat{\beta}_j\right)=\sqrt{\sigma^2\left(\hat{\beta}_j\right)} \quad , \quad j=0,1,2\ldots p$$

ile elde edilir[11].

3.3.2. Modelin önemlilik testi

Model kurulduktan sonra modelin önemliliğinin test edilmesi gerekmektedir. Burada yine tek değişkenli lojistik regresyon modelinde olduğu gibi olabilirlik oran testi kullanılmaktadır. EÇOB oran testi *G* istatistiğine bağlıdır. Tek fark $p+1$ parametreyi kapsayan $\hat{\beta}$ vektörüne dayanan modelden bulunan $\hat{\pi}$ değeridir. "Modelde bulunan p bağımsız değişene ait eğim katsayısının sıfıra eşit olduğu" hipotezi altında, *G* istatistiği p serbestlik derecesi ile ki-kare dağılır. H_0 hipotezinin reddedilmesi durumunda en az bir katsayının sıfırdan farklı olduğu sonucuna varmadan önce modelde bulunması gerektiği düşünülen tüm katsayıların Wald test istatistiği ile test edilmesi gerekmektedir. Wald istatistiği modeldeki herhangi bir değişkenin önemli mi yoksa önemsiz mi olduğunu göstermektedir.

Burada önemli olan en iyi modeli en az parametre ile oluşturmaktır. Yani sadece önemli olan değişkenleri modele dahil edip yeni model kurmaktır. Bunun için de tam modelle, azaltılmış modelin olabilirlik oranları karşılaştırılır. Bu karşılaştırmada *G*

istatistiği, $(v_2 - v_1)$ serbestlik dereceli ki-kare dağılır. v_2=(tam modeldeki değişken sayısı + 1), v_1=(azaltılmış modeldeki değişken sayısı+1)' dir.

$G = -2$ [(azaltılmış model için log-olabilirlik)–(tam model için log-olabilirlik)]

eşitliği ile bulunur. G istatistiği için $(v_2 - v_1)$ serbestlik dereceli p değeri bulunabilir. Eğer p değeri 0,05 den büyük ise değişken sayısı azaltılmış modelin, tam model kadar iyi olduğu söylenebilir. Ayrıca modele girecek değişkenleri belirlemede istatistiksel katsayıların önemlilik testi yeterli olmamaktadır. Bunun yanında birçok etkene bakmak gerekmektedir. Örneğin, kategorik olarak ölçeklenmiş bir bağımsız değişkenin modelden çıkarılacağı veya modele dahil edileceği kabul edilsin. Bu bağımsız değişkenin modelden çıkarılacağı veya modele dahil edileceği zaman, onun bütün dizayn değişkenleri de modelden çıkarmalı veya modele dahil edilmelidir. Genelleme yapıldığında kategorik bir değişkenin k düzeyi varsa, bu değişkenin modelden çıkarılmasının olabilirlik oran testi için serbestlik derecesine katkısı k-1 olmaktadır.

Bağımsız değişkenlerden bazıları kesikli(ırk, cinsiyet, tedavi grubu vb.) ise bunları sürekli değişkenlermiş gibi kabul ederek modele dahil etmek uygun olmamaktadır. Bundan dolayı çeşitli ve farklı düzeyleri göstermek için sayısal değeri olmayan keyfi tanımlayıcılar veya kodlar kullanılmaktadır. Bu kodlarla tanımlanan değişkenlere dummy (kukla) veya (dizayn) değişkenleri denir. Örneğin bağımsız değişkenlerden birinin ırk değişkeni olduğu ve "siyah", "beyaz" ve "diğerleri" olarak kodlandığı varsayılsın. Irk değişkeni için kategori sayısı 3 olduğundan 2 tane dizayn değişkeni (D_1 ve D_2) kullanılmaktadır[11].

3.4. Lojistik Regresyon Modelinin Katsayılarının Yorumlanması

Model kurulduktan, katsayıların hesaplanması ve öneminin değerlendirilmesi işlemlerinden sonra katsayıların yorumlanması işlemine geçilir. Kurulan herhangi bir modelin yorumlanması modeldeki tahmin edilen katsayılardan bir anlam

çıkarabilmeyi gerektirir. Bağımsız değişkendeki bir birimlik değişim, bağımsız değişkenin fonksiyonundaki değişim oranını gösterir. Bununla beraber yorumlama işleminin iki koşulu vardır.

1) Bağımlı ve bağımsız değişkenler arasındaki fonksiyonel ilişki saptanmalıdır.

2) Bağımsız değişken için bir birimlik değişime uygun olarak tanımlanmalıdır.

İlk aşamada bağımlı değişkenin fonksiyonunun, bağımsız değişkenin doğrusal fonksiyonu olup olmadığı test edilmelidir. Bu fonksiyona link fonksiyonu denir[19].

Doğrusal fonksiyonun link fonksiyonu birim matristir. Bunun nedeni bağımlı değişkenin tanımı gereği parametreleriyle doğrusal olmasıdır. Lojistik regresyon modelinde ise link fonksiyonu lojit formdadır ve

$$g(x) = \ln\left\{\frac{\pi(x)}{1-\pi(x)}\right\} = \beta_0 + \beta_1 x$$

şeklinde gösterilir.

Lojistik regresyonda $\beta_1 = g(x+1) - g(x)$ bağımsız değişkendeki x' in bir birimlik değişimiyle lojitte meydana gelen değişimi gösterir.

3.4.1. Modelde yalnız iki düzeyli(Dichotomous) bağımsız değişkenin olduğu durum

x' in 0 ve 1 değerlerini aldığı varsayılsın. $\pi(x)$ ve $1-\pi(x)$'in ikişer değişkeni vardır. $\pi(x)$ için $\pi(0)$ ve $\pi(1)$, $1-\pi(x)$ için $1-\pi(0)$ ve $1-\pi(1)$ dır. Bağımsız değişkenin ikili olduğu durumda lojistik regresyon modelinin değerleri;

$$x=1, y=1 \Rightarrow \ \pi(1)=\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0\beta_1}} \qquad x=1, y=0 \Rightarrow 1-\pi(1)=\frac{1}{1+e^{\beta_0+\beta_1}}$$

$$x=0, y=1 \Rightarrow \pi(0)=\frac{e^{\beta_0}}{1+e^{\beta_0}} \qquad x=0, y=0 \Rightarrow 1-\pi(0)=\frac{1}{1+e^{\beta_0}}$$

x=1 olan bireyler arasında sonuç değişkenin görülme($y=1$) odds oranı $\frac{\pi(1)}{1-\pi(1)}$,

x=0 olan bireyler arasında sonuç değişkenin görülme($y=1$) odds oranı $\frac{\pi(0)}{1-\pi(0)}$ dir.

Olasılık değerlerinin logaritması lojit olarak adlandırılır ve aşağıdaki gibi ifade edilir.

$$g(1)=\ln\left\{\frac{\pi(1)}{1-\pi(1)}\right\}$$

$$g(0)=\ln\left\{\frac{\pi(0)}{1-\pi(0)}\right\}$$

x = 1 için olasılığın x = 0 için olan olasılığa oranı odds oranı olarak ψ sembolü ile gösterilir ve,

$$\psi=\frac{\frac{\pi(1)}{1-\pi(1)}}{\frac{\pi(0)}{1-\pi(0)}} \tag{3.13}$$

şeklinde ifade edilir. Odds oranının ln' i log-odds oranı olarak adlandırılır ve

$$\ln(\psi) = \ln\left[\frac{\frac{\pi(1)}{1-\pi(1)}}{\frac{\pi(0)}{1-\pi(0)}}\right]$$

$$= g(1) - g(0)$$

ile ifade edilir ve bu lojit farkı olarak tanımlanır. Lojistik regresyon modeli için bulunan $\pi(1)$, $\pi(0)$, $1-\pi(1)$ ve $1-\pi(0)$ değerleri yukarıdaki eşitlikte yerine konursa;

$$\psi = \frac{\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\frac{1}{\beta_0}}{\frac{e^{\beta_0}}{1+e^{\beta_0}}\frac{1}{e^{\beta_0+\beta_1}}}$$

$$= \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}}$$

$$= e^{\beta_1}$$

olur. Burada da açıkça görüldüğü gibi iki düzeyli bağımsız değişkenin lojistik regresyonu için odds oranı $\psi = e^{\beta_1}$, lojit farkı ise $\ln(\psi) = \beta_1$' dir.

Teorik olarak örneklem genişliği yeteri kadar büyük olduğu zaman, $\hat{\psi}$'nin dağılımı normal dağılım olur. Odds oranı için α anlamlılık düzeyinde güven aralığının tahmini, β_1 katsayısı için güven aralığının alt ve üst noktalarının belirlenmesinden sonra bu değerlerin e üssünün alınmasıyla aşağıdaki şekilde elde edilir.

$$\exp\left[\hat{\beta}_1 \pm z_{1-\alpha/2}\ SE\left(\hat{\beta}_1\right)\right]$$

Bağımsız değişken iki düzeyli olduğunda lojistik regresyon programları dizayn değişkenlerini oluşturmak için iki farklı metot sunmaktadır. Bunlardan birincisi

marjinal metot diğeri ise kısmi metottur. Marjinal metot ortalamalardan sapma, kısmi metot ise referans hücre metodu diye adlandırılmaktadır.

Kısmi metot da *x*' in en küçük değerine 0, en büyük değerine 1 değeri atanmaktadır. Örneğin; cinsiyet değişkeninin *Erkek=1* ve *Kadın=2* olarak kodlandığını varsayalım. Kısmi metot kullanılarak elde edilen D dizayn değişkeni *Erkek =0*, *Kadın=1* şeklinde kodlanacak ve D için tahmin edilen $\hat{\beta}_1$ katsayısının *e* üssü, kadınların erkeklere göre odds oranının tahminini verecektir.

Marjinal metotta ise kısmi metottaki *x* 'in en küçük değerine -1, en büyük değerine ise 1 değeri atanarak dizayn değişkeni oluşturulmaktadır. *Erkek=1* ve *Kadın=3* olarak kodlanmışsa dizayn değişkeni (D) *Erkek =-1* ve *Kadın=1* olarak kodlanacaktır. Marjinal metotta kadınların erkeklere göre odds oranının tahmini;

$$\begin{aligned}\ln\left[\psi\left(kadin\,,\,erkek\right)\right] &= \hat{g}\left(kadin\right)-\hat{g}\left(erkek\right)\\ &= \hat{g}\left(D=1\right)-\hat{g}\left(D=-1\right)\\ &= \left[\hat{\beta}_0+\hat{\beta}_1\ \ (D=1)\right]-\left[\hat{\beta}_0+\hat{\beta}_1\ \ (D=-1)\right]\\ &= 2\hat{\beta}_1\end{aligned}$$

ile elde edilir. Buradan $\hat{\psi}=\exp\left(2\hat{\beta}_1\right)$ olarak bulunur ve $\hat{\beta}_1$ için güven aralıklarının alt ve üst sınırları,

$$\exp\left[2\hat{\beta}_1 \pm z_{1-\alpha/2}\ \ 2S\hat{E}\left(\hat{\beta}_1\right)\right]$$

şeklinde bulunur.

Özet olarak iki düzeyli değişken için önemli olan parametre odds oranıdır. Lojistik regresyon katsayısı ve odds oranı arasındaki ilişki lojistik regresyon sonuçlarının yorumlanması için temel oluşturmaktadır[12].

3.4.2. Modelde ikiden fazla düzeyli bağımsız değişkenin olduğu durum

Bağımsız değişkenin ikiden fazla düzey içerdiği durumlar da olabilir. Bu durumlar için de dizayn değişkenleri kullanılır.

Dizayn edilmiş değişkenlerin seçimi, referans hücre metodu(kısmi metot) ile yapılabilir. Bu yönteme göre, referans grup olarak seçilen düzey 0, diğer bütün gruplar için dizayn edilmiş değişken 1 yapılarak seçim yapılır. Çizelge 3.1' de bu durum gösterilmiştir.

Çizelge 3.1. Beyazların referans grup olarak kullanıldığı ırk değişkeni için dizayn değişkenlerinin referans hücre metoduyla belirlemesi

IRK (Kod)	Dizayn Değişkenleri		
	D_1	D_2	D_3
Beyaz (1)	0	0	0
Siyah (2)	1	0	0
İspanyol (3)	0	1	0
Diğerleri (4)	0	0	1

Örnek olarak siyahların beyazlara göre karşılaştırılması,

$$\begin{aligned}\ln\left[\hat{\psi}\left(siyah, beyaz\right)\right] &= \hat{g}\left(siyah\right) - \hat{g}\left(beyaz\right) \\ &= \left[\hat{\beta}_0 + \hat{\beta}_{11}\ \left(D_1 = 1\right) + \hat{\beta}_{12}\ \left(D_2 = 0\right) + \hat{\beta}_{13}\ \left(D_3 = 0\right)\right] \\ &\quad - \left[\hat{\beta}_0 + \hat{\beta}_{11}\ \left(D_1 = 0\right) + \hat{\beta}_{12}\ \left(D_2 = 0\right) + \hat{\beta}_3\ \left(D_3 = 0\right)\right] \\ &= \hat{\beta}_{11}\end{aligned}$$

şeklindedir.

Genel olarak, herhangi bir lojistik regresyon katsayısı için α anlamlılık düzeyinde güven aralığı,

$$\hat{\beta}_{ij} \pm z_{1-\frac{\alpha}{2}} S\hat{E}\left(\hat{\beta}_{ij}\right)$$

ile bulunur. Odds oranı için α anlamlılık düzeyinde güven aralığı ise bu limitlerin e üssü alınarak,

$$\exp\left[\hat{\beta}_{ij} \pm z_{1-\alpha/2} \;\; \hat{SE}\left(\hat{\beta}_{ij}\right)\right]$$

şeklinde bulunur[12].

Dizayn değişkenlerinin kodlanmasının ikinci bir yolu ise ortalamadan sapma(marjinal) metodudur. Bu kodlama yöntemi genel ortalamadan grup ortalamasının sapmasının etkisini açıklar. Lojistik regresyonda, "grup ortalaması" grubun lojitidir ve "genel ortalama" ise tüm grupların ortalama lojitidir. Bu yönteme göre, dizayn değişkeninin tüm değerleri -1 alınıp, geri kalan diğer değişkenler için 0 ve 1 kodlaması kullanılır. Bu yöntem Çizelge 3.2' de gösterilmiştir.

Çizelge 3.2. Beyazların referans grup olarak kullanıldığı ırk değişkeni için dizayn değişkenlerinin ortalamadan sapma metoduyla belirlemesi

IRK (Kod)	Dizayn Değişkenleri		
	D_1	D_2	D_3
Beyaz (1)	-1	-1	-1
Siyah (2)	1	0	0
İspanyol (3)	0	1	0
Diğerleri (4)	0	0	1

Ortalamadan sapma kodlaması kullanılarak bulunan katsayıların yorumu referans hücre metodu kadar kolay ve açık değildir. Tahmin edilen katsayıların e üssünün alınması, belirli bir grup için odds değerlerinin, oddsların geometrik ortalamasına oranını vermektedir. Örnek olarak birinci dizayn değişkeni için hesaplama yapıldığında, $\hat{g}_j$ bağımsız değişkenin j' inci kategorisi için lojit olmak üzere,

$$\exp\left(\hat{\beta}_{11}\right)=\exp\left(\hat{g}_2-\bar{g}\right)=\exp\left(\hat{g}_2\right)/\exp\left[\Sigma g_j/4\right]$$

şeklinde ifade edilir.

Bu hesaplanmış olan odds oranı gerçek odds oranı değildir. Çünkü pay ve paydada bulunan nicelikler iki farklı kategori için odds değerlerini temsil etmemektedir. Tahmin edilen katsayının e üssünün alınması, ortalama oddsa göre odds değerini ifade etmektedir.

Bu yöntemle bulunan parametre tahminleri bir kategorinin, referans bir kategoriye göre odds oranını tahmin etmek için de kullanılabilir. Örnek olarak Çizelge 3.2.' de verilmiş olan dizayn değişkenlerinin belirlenerek siyahların beyazlara göre log-odds değerinin tahmini,

$$\begin{aligned}\ln\left[\hat{\psi}\left(siyah,beyaz\right)\right]&=\hat{g}\left(siyah\right)-\hat{g}\left(beyaz\right)\\&=\left[\hat{\beta}_0+\hat{\beta}_{11}\left(D_1=1\right)+\hat{\beta}_{12}\left(D_2=0\right)+\hat{\beta}_{13}\left(D_3=0\right)\right]\\&\quad-\left[\hat{\beta}_0+\hat{\beta}_{11}\left(D_1=-1\right)+\hat{\beta}_{12}\left(D_2=-1\right)+\hat{\beta}_{13}\left(D_3=-1\right)\right]\\&=2\hat{\beta}_{11}+\hat{\beta}_{12}+\hat{\beta}_{13}\end{aligned}\qquad(3.14)$$

şeklinde hesaplanır ve Eş. 3.14' ün varyansı,

$$\begin{aligned}\hat{\text{var}}\left\{\ln\left[\hat{\psi}\left(siyah,beyaz\right)\right]\right\}&=4\,\text{var}\left(\hat{\beta}_{11}\right)+\text{var}\left(\hat{\beta}_{12}\right)+\text{var}\left(\hat{\beta}_{13}\right)+\\&\quad+4\,\text{cov}\left(\hat{\beta}_{11},\hat{\beta}_{12}\right)+4\,\text{cov}\left(\hat{\beta}_{11},\hat{\beta}_{13}\right)+2\,\text{cov}\left(\hat{\beta}_{12},\hat{\beta}_{13}\right)\end{aligned}$$

ile tahmin edilir[12]. Bu ifadede her bir terim lojistik regresyon hesaplaması için kullanılan paket programlardan bulunabilir.

Sonuç olarak ortalamadan sapma metoduyla odds oranlarının tahmini, referans hücre metoduna göre daha karmaşık hesaplamalar gerektirmektedir.

3.4.3. Modelde sürekli bir bağımsız değişkenin olduğu durum

Lojistik regresyon modeli sürekli bir bağımsız değişkeni içerdiğinde, tahmin edilen katsayıların yorumlanması değişkenin modele nasıl girdiğine bağlı olur. Modelde sürekli bir değişken olması durumunda, bu değişkenin katsayısının yorumlanması amacıyla geliştirilen metotta lojitin değişkenle doğrusal olduğu varsayılır[12].

Lojitin sürekli değişken x ile doğrusal olduğu varsayımı altında lojit $g(x)=\beta_0+\beta_1 x$ olur. Eğim katsayısı (β_1), x' deki "1" birimlik artışın log odds değerinde meydana getireceği değişimi verir. x'in herhangi bir değeri için $\beta_1=g(x+1)-g(x)$ olur. Bağımsız değişken değerinde gözlemlenmiş olan"1" birimlik değişim genellikle istatistiksel olarak önemli olmamaktadır. Örnek olarak, yaş değişkenindeki "1" birim artışın ya da sistolik kan basıncındaki 1 mm Hg artışın önemli sayılmayacak kadar küçük olduğu bir gerçektir. Yaştaki 10 yıllık artışın ya da kan basıncındaki 10 mm Hg artışın log-odds değerinde meydana getireceği değişimi ele almanın daha yararlı olacağı kabul edilir. Diğer yandan x' in tanım aralığı (0,1) ise, bağımsız değişken değerinde meydana gelecek "1" birimlik değişim log odds değerinde çok büyük bir etki yapacaktır. O halde 0,01 birimlik artış da daha gerçekçi olacaktır. Bu nedenden dolayı sürekli ölçekli bağımsız değişken için geçerli yorumlar yapabilmek amacı ile bağımsız değişkende gözlemlenen "c" birimlik bir değişim için nokta ve aralık tahmin metotları geliştirilmiştir[12].

x'deki c birimlik bir değişim için log-odds oranı, $g(x+c)-g(x)=c\beta_1$ olmak üzere iki lojit farkından elde edilir ve karşılık gelen odds oranı bu lojit farkın e'nin üssüne yazılmasıyla odds oranı(c)=odds oranı(x+c,x)=$\exp(c\beta_1)$ oluşur.

3.4.4. Çok değişkenli durumda katsayıların yorumlanması

Tek değişkenli modeller kurularak yapılan veri analizleri çoğunlukla uygun olmamaktadır. Çünkü bağımsız değişkenler genellikle birbirleriyle ilişkili değildir ve sonuç değişkeninin farklı düzeyleri için dağılımları farklı olabilir. Bu sebepten dolayı

çok değişkenli analizle verilerin modellenmesi daha anlamlı sonuçlar verir. Bu tarz bir analizin amacı, her bir değişkenin tahmin edilen etkisini ve modeldeki diğer bağımsız değişkenler ile arasındaki ilişkiyi istatistiksel olarak ayarlamaktır. Bu amaç çok değişkenli lojistik regresyon modeline uygulanarak, tahmin edilen her bir katsayının modeldeki diğer değişkenler için ayarlama yapan log-odds'un tahmin edilmesini sağlar.

Çok değişkenli lojistik regresyon modelinden tahmin edilen katsayıları tamamen yorumlayabilmek için "diğer değişkenler için istatistiksel olarak ayarlama" teriminin ne anlama geldiğini açıklamak gerekmektedir. Öncelikle ayarlama işlemini doğrusal regresyon modelinde uygulamak ve sonra da bu konuyu lojistik regresyon modeline uyarlamak daha yararlı olacaktır.

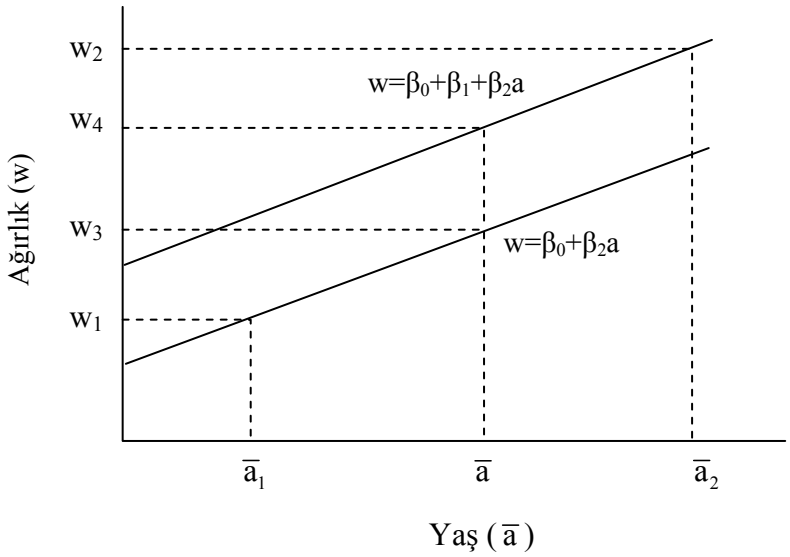
Çok değişkenli durumda modelde yalnızca iki bağımsız değişken olduğu (biri iki düzeyli, diğeri sürekli) varsayılsın. Öncelikli olarak, risk faktörünün iki düzeyli bağımsız değişken olduğu model ele alınsın. Risk faktörüne maruz olma "var ya da yok" olarak kodlandığında, sürekli bir değişken (örneğin yaş) için ayarlama yapılması epidemiyolojik araştırmalarda oldukça sık karşılaşılan bir durumdur. Doğrusal regresyondaki benzer durum kovaryans analizi olarak adlandırılmaktadır[11]. Örneğin birbirinden farklı iki gruptaki çocukların ortalama ağırlıkları karşılaştırılmak istensin. Ağırlıklarla ilişkili olan birçok faktör vardır ve bunlardan birisi de yaştır. Yaş haricinde diğer bütün faktörlerin her iki gruba olan etkileri hemen hemen aynıysa, o zaman iki grubun ağırlıklarının karşılaştırılması için tek değişkenli analiz yeterli olmaktadır. Bu karşılaştırma iki grup arasındaki farkın doğru tahminini sağlamaktadır. Bununla beraber, eğer bir grup diğerinden daha gençse, o zaman iki grubun karşılaştırılması anlamsız olmaktadır. Çünkü gözlenen farklılığın küçük bir kısmı yaştaki farklılıktan kaynaklanmaktadır. Gruplar arasındaki yaş farklılığını elemine etmeden grupların farklılığını belirlemek mümkün değildir. Bu şekilde yaş ve ağırlık arasındaki ilişkinin doğrusal olduğu ve her bir grup için sıfırdan farklı olmak üzere aynı eğime sahip olduğu varsayılmaktadır. Grup farklılıkları hakkında çıkarsama yapmadan önce genellikle bu varsayımların her ikisi de kovaryans analizi ile test edilir.

Şekil 3.1'deki durumu açıklayan istatistiksel modele göre ağırlık (w), yaş ise (a) ile gösterilmek üzere, ağırlık $w = \beta_0 + \beta_1 x + \beta_2 a$ ile ifade edilsin[11]. Modelde grup 1 için x=0 ve grup 2 için x=1 kodlanmıştır. Bu modelde β_1 parametresi gruplar arasındaki gerçek ağırlık farkını gösterirken, β_2 parametresi yaştaki her bir yıl için ağırlıktaki değişimin oranını verir. Grup 1 için yaş ortalaması $\overline{a}_1$ ve grup 2 için yaş ortalaması $\overline{a}_2$ olsun. Bu değerler Şekil 3.1' de gösterilmiştir. Grup 1'in ortalama ağırlığının, grup 2'nin ortalama ağırlığıyla karşılaştırılması, w_1'in w_2 ile karşılaştırılmasına denktir. Bu farkın modelle ifade edilmesi aşağıdaki gibidir.

$$\left(w_2 - w_1\right) = \beta_1 + \beta_2\left(\overline{a}_2 - \overline{a}_1\right)$$

Böylece karşılaştırma işlemi yalnızca gruplar arasındaki ağırlığın gerçek farkını $\left(\beta_1\right)$ değil aynı zamanda grupların yaşları arasındaki farkı yansıtan $\beta_2\left(\overline{a}_2 - \overline{a}_1\right)$ terimini de kapsamaktadır.

Yaş için istatistiksel ayarlama (düzeltme), yaşın herhangi bir ortak değerinde iki grubu karşılaştırmayı gerektirir. Bu değer iki grubun ortalamaları için kullanılan $\overline{a}$ ile gösterilsin (Şekil 3.1). Buna göre w_4' ün w_3 ile karşılaştırılması, $\left(w_4 - w_3\right) = \beta_1 + \beta_2\left(\overline{a} - \overline{a}\right) = \beta_1$ şeklinde olup iki grup arasındaki gerçek farka eşittir. Uygulamada yaş için herhangi bir ortak değer seçilmesi gruplar arasındaki gerçek farkı değiştirmez, fakat bu değeri genel ortalama olarak seçmenin iki tane önemli nedeni vardır. Bunlar biyolojik olarak kabul edilebilir olması ve yaş ile kilo arasındaki ilişkinin doğrusal ve her bir grup içinde sabit olduğu kabul edilen sınırlar arasında olmasıdır.



Şekil 3.1. Değişik yaş dağılımlarına sahip iki farklı grubun ağırlıklarının karşılaştırılması

Şekil 3.1'de gösterilen durumun aynısı göz önünde bulundurulsun fakat bağımlı değişken olan kilo yerine iki düzeyli bir değişken ve dikey ekseninde lojit olduğu varsayılsın. Bu durumda lojit $g(x,a)=\beta_0+\beta_1 x+\beta_2 a$ olarak alınır. Sonuç değişkeni ve lojit $2x2$ tabloda çapraz sınıflandıktan sonra tek değişkenli karşılaştırmadan elde edilen log odds oranının yaklaşık değeri $\beta_1+\beta_2(\bar{a}_2-\bar{a}_1)$ olarak bulunur. Bu karşılaştırmayla yaş dağılımındaki farklılıktan dolayı grup etkisi yanlış olarak tahmin edilir. Bu farkı göz önünde bulundurmak ya da ayarlama yapmak için modele yaş değişkeni dahil edilir ve yaşın ortak bir değerinde (örnek olarak genel ortalama $\bar{a}$) lojit fark hesaplanır. Bu lojit fark $g(x=1,\bar{a})-g(x=0,\bar{a})=\beta_1$ olur[11]. Sonuç olarak β_1 katsayısı, iki grup aynı yaş dağılımına sahip olduğunda tek değişkenli karşılaştırmadan elde etmeyi beklediğimiz log-odds oranına eşittir.

Örnek olarak yaş değişkeninin sürekli değişken değil de, 45 yaşı kesim noktası kabul ederek ikili bir değişken olarak modele dahil edildiği varsayılsın. Yaş düzeltmeli grup etkisini elde edebilmek için iki ikili değişkenden oluşan model kurulduktan sonra (yaş için ikili değişkenin ortak bir değerinde) grubun iki düzeyi için lojit fark hesaplanır. Değişken türü ve sayısı ne olursa olsun uygulanacak işlem benzerdir. Düzeltilmiş odds oranları, bireylerin yalnızca farklılık gösteren karakterlerini

karşılaştırıp diğer bütün değişkenlerini sabit tutarak elde edilir. Yapılmış olan düzeltme, sonuç değişkeninin iki düzeyi için diğer bütün değişkenlerin etkisi sabit tutulduğunda, bireylerin yalnızca test edilmek istenen özel karakteristiğe göre farklılık gösterip göstermeyeceğinin belirlenmesi açısından önem teşkil etmektedir.

İstatistiksel olarak düzeltilmiş log odds oranları ve odds oranları yorumlanacağı zaman bir nokta göz önünde bulundurulmalıdır. Düzeltmenin etkisi tamamıyla modelin varsayımlarının sağlanmasıyla bağlantılıdır (doğrusal ve sabit eğime sahip olma gibi). Bu varsayımlardan sapmalar ayarlamayı yararsız kılmaktadır. Varsayımlardan sapmalara bir örnek ilişkinin doğrusal fakat eğimin farklı olduğu durumdur ve bu etkileşim olarak adlandırılır[11]. (Bkz. Şekil 3.2)

3.4.5. Etkileşim ve etki karışımı

Bu bölümde etkileşim kavramı ele alınıp, lojistik regresyon modeli üzerinde yapmış olduğu etkilerinin nasıl kontrol edileceği üzerinde durulmuştur. Ayrıca modeldeki tahmin edilen katsayıların, etkileşim ve etki karışımından nasıl etkilendikleri gösterilecektir.

Etki karışımı için düzeltme işlemi, bir önceki bölümde belirtildiği gibi herhangi bir etkileşim olmadığı durumda uygundur.

Etki karışımı hem sonuç değişkeni hem de risk faktörü (önemli bağımsız değişken) ile ilişkili olan bir birlikte değişeni(kovaryant) tanımlamakta kullanılır. Her iki ilişki mevcut olduğu zaman risk faktörü ve sonuç değişkeni arasındaki ilişki "etki karışmış" şeklinde ifade edilir.

İki sonuçlu bir risk faktörü ve sürekli bir değişken içeren bir model olduğu varsayılsın. Bağımlı ve bağımsız değişken arasındaki ilişki risk faktörünün her bir seviyesi için farklılık göstermiyorsa incelenen risk faktörü ve bağımsız değişken arasında etkileşim yoktur denir. Bu anlatılanlar grafikle gösterildiğinde, birbirine paralel iki doğru çizmek gerekir(Bkz. Şekil 3.2). Bu çizgilerin her biri risk

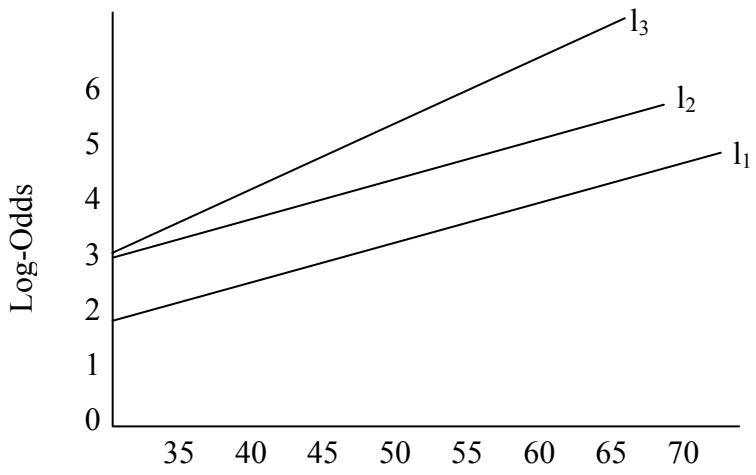
faktörünün bir düzeyini temsil etmektedir. Genel bir ifadeyle etkileşimin olmaması, model tarafından iki veya daha fazla değişkenden oluşan, ikinci dereceden ya da daha yüksek dereceden terim olmamasıyla kendini göstermektedir.

Etkileşimin olduğu durumlarda, risk faktörü ve bağımlı değişken arasındaki ilişki bağımsız değişkenin değerine bağlı olur. Yani, bağımsız değişkenin değeri risk faktörünün etkisini değiştirir. Epidemiyolojistler, etki değiştirici (effect modifier) terimini, bir değişkenin risk faktörünü etkilemesi durumunu tanımlamak için kullanmışlardır.

Genel olarak etkileşimin olmadığı model iki veya daha fazla değişkenle ilişkili 2. dereceden veya daha yüksek dereceden terimi olmayan bir model olarak nitelendirilir.

Bağımsız değişken ve risk faktörü arasındaki etkileşim şekil üzerinde gösterildiğinde, risk faktörünün her bir düzeyini ifade eden doğrular birbirine paralel olmaz(Bkz. Şekil 3.2). Başka bir ifadeyle eğimleri farklı doğrular bağımsız değişken ve risk faktörü arasında bir etkileşim olduğunu gösterir.

Etkileşim terimini açıklamak için, bağımlı değişkenin koroner kalp hastalığı durumunun (var ya da yok), risk faktörünün cinsiyeti ve bağımsız değişkenin yaş ile ifade edildiği bir örnek ele alınsın. l_1' in kadınlar için, l_2' nin ise erkekler için yaşın bir fonksiyonunun lojitini gösterdiği varsayılsın. Şekil 3.2 de görüldüğü üzere l_1 ve l_2 ' nin aynı eğime sahip ve birbirine paralel olması, yaşın etkisinin hem kadınlarda hem de erkeklerde aynı olduğunu ifade eder. Bu durumda, kadın ve erkeklerde kalp hastalığının görülmesi yaşa bağlı olarak değişmiyor yani etkileşim yoktur denilebilir. Dolayısıyla erkeklerin kadınlara göre log-odds oranı(l_2'nin l_1'e karşı log-odds oranı) "l_2-l_1" ile elde edilir. Bu fark, iki doğru arasındaki dik uzaklığa eşit olup bağımsız değişkenin(yaş) her değeri için aynıdır. Çünkü her değer için doğruların eğimleri aynıdır.



Şekil 3.2. Etkileşimin olup olmadığını gösteren üç farklı modelin lojitlerinin grafiği

Etkileşimin olduğu durumu incelemek için l_3' ün yaşın bir fonksiyonunun lojitini gösterdiği varsayılsın. l_3 doğrusunun eğiminin l_1 doğrusunun eğiminden daha fazla eğime sahip olduğu Şekil 3.2 de açık olarak görülmektedir. Bu durum risk faktörü olan cinsiyet ile bağımsız değişkeni ifade eden yaş arasında bir etkileşimin olduğunu gösterir. Erkeklerin kadınlara göre log-odds oranı(l_3' ün l_1'e karşı log-odds oranı) "l_3-l_1" ile bulunur. Bu fark iki doğru arasındaki dik uzaklığa eşit olup, bağımsız değişkenin(yaş) her değeri için değişmektedir. Bundan dolayı risk faktörünü gösteren cinsiyet için odds oranı karşılaştırılmasının yapılması, bağımsız değişkeni ifade eden yaşın değeri belirtilmeden tahmin edilemez. Bu da bize göstermektedir ki, yaş bir etki değiştiricisidir.

Bundan dolayı, bir (x) bağımsız değişkeninin etki değiştirici olup olmadığına karar vermek için lojitlerin çizimlerinden yararlanılarak iki koşulun sağlanıp sağlanmadığına bakılır. Birinci koşul bağımlı değişkenin bağımsız değişkenle kesinlikle ilişkili olması, ikinci koşul ise bağımsız değişkenin risk faktörleriyle kesinlikle ilişkili olmasıdır[11].

Kısaca, bir bağımsız değişkenin etki karıştırıcı olup olmadığı, bağımsız değişkeni kapsayan ve kapsamayan modellerden elde edilen risk faktörü değişkeninin tahmin edilen katsayılarının karşılaştırılmasıyla anlaşılır. Karşılaştırma sonucunda, bu

modellerden elde edilen risk faktörü için tahmin edilen katsayılarda "istatistiksel olarak önemli" herhangi bir değişim olup olmadığı bu bağımsız değişkeni etki karıştırıcı olarak belirler ve modele dahil edilmesi için yeterlidir. Etki karıştırıcı ve etkileşim değerlerinin modele dahil edilmesi ve tanımlanması, değişken sayısından ve bu değişkenlerin ölçme düzeylerinden bağımsızdır.

Ayrıca etki karıştırıcı etkisi, uygun temel etkilerin ve çarpımsal terimlerin lojistik regresyon modeline dahil edilmesi ile giderilebilir[17].

3.4.6. Etkileşim olduğu durumlarda odds-oranlarının tahmini

Modelde bulunan bir risk faktörü ile başka bir değişken arasında etkileşim olması halinde risk faktörü için tahmin edilen odds oranı, risk faktörünün etkileşim içinde olduğu değişkenin değerine bağlı olarak tanımlanır. Bu durumda odds oranlarını tahmin etmek için kullanılan formülde bir değişiklik yapılarak etkileşim içinde olan değişkenler arasındaki lojit fark da dikkate alınır[12].

Risk faktörünü (F), bağımsız değişkeni (X) ve onların etkileşimini de (FxX)' in ifade ettiği bir modelin kurulmuş olduğu varsayılsın. F=f ve X=x değerlerini aldığında bu modelin lojiti

$$g(f,x)=\beta_0+\beta_1 f+\beta_2 x+\beta_3 fx \tag{3.15}$$

olur.

X=x' de sabit tutulduğunda F=f_0'' a karşı F=f_1 düzeyleri için log-odds oranı,

$$g(f_1,x)=\beta_0+\beta_1 f_1+\beta_2 x+\beta_3 f_1 x$$

ve

$$g(f_0,x)=\beta_0+\beta_1 f_0+\beta_2 x+\beta_3 f_0 x$$

ifadelerinin farkının lojitinin alınmasıyla

$$\begin{aligned}\ln[\psi(F=f_1,F=f_0,X=x)] &= g(f_1,x)-g(f_0,x)\\ &= (\beta_0+\beta_1 f_1+\beta_2 x+\beta_3 f_1 x)-(\beta_0+\beta_1 f_0+\beta_2 x+\beta_3 f_0 x)\\ &= \beta_1(f_1-f_0)+\beta_3 x(f_1-f_0) \qquad (3.16)\end{aligned}$$

elde edilir. Eş. 3.16' de parametrelerin yerine tahmin edilmiş değerler kullanıldığında tahmin edilmiş log-odds oranı elde edilir ve varyansın tahmini,

$$\begin{aligned}\hat{Var}\left\{\ln\left[\hat{\Psi}(F=f_1,F=f_0,X=x)\right]\right\} &= \hat{Var}(\hat{\beta}_1)(f_1-f_0)^2+x[(f_1-f_0)]^2\hat{Var}(\hat{\beta}_3)\\ &\quad +2x(f_1-f_0)^2\,\hat{\text{cov}}(\hat{\beta}_1,\hat{\beta}_3) \qquad (3.17)\end{aligned}$$

şeklinde olur.

Lojistik regresyon programlarının hemen hepsi modeldeki tahmin edilen parametrelerin varyans ve kovaryansının tahmin değerlerini verir. Tahmin değerleri elde edildikten sonra, Eş. 3.17' de bulunan değerlerin yerine konulmasıyla odds-oranının varyansı tahmin edilir. $\psi(F=f_1,F=f_0,X=x)$ için α anlamlılık düzeyinde güven aralığının alt ve üst limitleri,

$$\exp\left(\left[\hat{\beta}_1(f_1-f_0)+\hat{\beta}_3 x(f_1-f_0)\right]\pm z_{1-\alpha/2}\hat{SE}\left\{\ln\left[\hat{\Psi}(F=f_1,F=f_0,X=x)\right]\right\}\right) \qquad (3.18)$$

ile bulunur.

F iki düzeyli bir risk faktörü iken Eş. 3.17' in varyansı ve log-odds tahmin edicileri daha basit şekil alır. Eğer f_1=1, f_0=0 olarak alınırsa log-odds oranının tahmini,

$$\ln\left[\hat{\Psi}(F=1,F=0,X=x)\right]=\hat{\beta}_1+\hat{\beta}_3 x \qquad (3.19)$$

olarak bulunur ve buradan varyansın tahmini,

$$\hat{Var}\left\{ \ln\left[\hat{\Psi}(F=1,F=0,X=x)\right]\right\} = \hat{Var}\left(\hat{\beta}_1\right) + \hat{Var}\left(\hat{\beta}_3\right)x^2 + 2\,\hat{\text{cov}}\left(\hat{\beta}_1,\hat{\beta}_3\right)x \tag{3.20}$$

şeklinde olur. Odds oranı için tahmin edilen güven aralığının alt ve üst sınırları,

$$\exp\left(\left[\hat{\beta}_1 + \hat{\beta}_3 x\right] \pm z_{1-\alpha/2}\hat{SE}\left\{ \ln\left[\hat{\Psi}(F=1,F=0,X=x)\right]\right\}\right) \tag{3.21}$$

ile elde edilir[12].

3.5. Lojistik Regresyon İçin Model Yapılandırma Stratejisi

Bundan önceki bölümlerde lojistik regresyon modelindeki katsayıların tahmin edilmesi, test edilmesi ve yorumlanması üzerinde durulmuştur. Az sayıda bağımsız değişkene sahip modeli kurmak kolaydır. Bağımsız değişken sayısının çok olduğu durumda ise modeli kurmak daha karmaşık bir hal alır. Bu karışık durumun üstesinden gelmek ve veri kümesini iyi bir şekilde modelleyebilmek için bazı metotların geliştirilmesine ihtiyaç duyulmuştur.

Herhangi bir metodun amacı, en iyi modeli verecek değişkenleri seçmektir. Bu amacı gerçekleştirmek için: (a) modele dahil edilecek değişkenleri seçmek için temel bir plana, (b)hem modeldeki değişkenlerin incelenmesi açısından hem de modelin temel uyumunun belirlemesi açısından, model yeterliliğini tayin edebilecek çeşitli metotlara ihtiyaç vardır[15].

3.5.1. Değişken seçimi

Bir değişkenin modele dahil edilmesi için kriter, bir problemden diğerine ve bir bilimsel yaklaşımdan diğerine değişmektedir. İstatistiksel model oluşturmada genel yaklaşım mümkün olan en az değişkeni kullanarak modeli açıklamaktır. Modele

eklenen değişken sayısı ne kadar çok olursa, tahmin edilen standart hata o kadar büyük olur. Miettinen(1976) yapmış olduğu çalışmalarında modele katkısı olup olmadığına bakılmaksızın bağımlı değişkenle ilişkisi olduğu düşünülen her bir bağımsız değişkeni modele dahil ederek, bu değişkenlerin veri üzerinde tek başlarına etkili olmasalar bile birlikte alındıklarında önemli derecede etki gösterebildikleri sorununu gündeme getirmiştir. Bu yaklaşımla ilgili büyük bir problem ise; modelin yeterince uygun olmaması ve sayısal olarak olduğundan daha büyük ve doğru olmayan tahminler üretmesidir. Bu durum genellikle modele dahil edilen değişken sayısının, üzerinde çalışılan birey sayısına oranla daha büyük olduğunda görülmektedir[11].

Lesaffre (1986), lojistik regresyon analizinde değişken seçim yöntemleri olarak doğrusal regresyon analizinde temel olan ileriye doğru seçim(forward selection), geriye doğru eleme(backward elimination), adımsal seçim(stepwise selection) ve tüm olası alt kümeler seçim(all subsests selection) yöntemlerini incelemiştir. Hosmer ve arkadaşları(1989), yeni bir seçim algoritması geliştirmişler, Miller(1984) ise doğrusal regresyon modeli için geliştirdiği değişken seçim algoritmasının lojistik modele uygulanabileceğini vurgulamıştır.

Bunların yanı sıra bazı araştırmacılar orijinal değişkenler yerine çıkarsanan(hipotenik) değişkenleri kullanarak değişken seçimine gidilmesini önermişlerdir. Bunlardan D'Agostino ve Pozen(1982) , değişkenlerin kümelenmesini takiben adımsal seçim yönteminin uygulanmasını önermektedirler. İlk aşama olan kümeleme işlemi istatistiksel yöntemlere ve mantıksal nedenlere dayanmaktadır. Bu da aynı özelliğe sahip değişkenlerin bir küme oluşturması demektir. İkinci aşamada her kümedeki birinci temel bileşenler üzerinde standart seçim yöntemi uygulanır.

Örneklem genişliğinin çok büyük olduğu durumlarda yukarıda değinilen yöntemlerden hiçbiri kullanılamamaktadır. Bu durumda değişken seçimi için bir yol, veri kümesini rastgele alt gruplara bölmek ve bunların her birine standart seçim yöntemlerinden birini uygulamaktır. Değişken sayısı az olduğunda yöntem seçilen kümede tekrar edilebilmektedir.

Öte yandan temel seçim yöntemlerini karşılaştırmak amacı ile çeşitli çalışmalar yapılmıştır. Bunlardan birinde Berk(1978), yalnızca ileriye doğru seçim ve geriye doğru eleme yöntemlerinin, tüm olası alt kümeler seçim yönteminden çok farklı olmaları durumunda aynı sonuca ulaşılacağını göstermiştir. Ayrıca büyük örneklem durumunda ileriye doğru seçim yönteminin, tüm olası alt kümeler seçim yöntemine yakın sonuçlar verdiğini belirtmiştir.

Yukarıda değinilen yöntemlerden biri ile açıklayıcı değişkenlerin bir alt kümesini seçmek ve katsayı tahminlerini aynı verileri kullanarak yapmak tahminlerin yanlı olmasına neden olur. Miller(1984), bu yan'ı, ihmal(ommission bias), rekabet(competition or selection bias) ve durdurma kuralı(stopping rule bias) olmak üzere üç grupta incelemiştir. Yan sorunu için getirilen çözüm, örneklemi deney ve test gruplarına bölerek değişken seçimini deneyde, katsayı tahminlerini de test kümesinde yapmaktır. Ancak bu durumda bilgi kaybı büyük olacağı için, örneklem sayısının fazla olduğunda uygulanması önerilen bir yaklaşımdır[5].

Bir lojistik regresyon modeli için değişken seçiminde izlenmesi gereken adımlar şöyledir;

(1) Değişken seçme işlemi her bir değişkenin ayrı ayrı tek değişkenli analizlerinin yapılması ile başlar. Sınıflandırılmış veya sıralanmış ölçme düzeyinde ölçülen sürekli bağımsız değişkenlerin k düzeyine karşı sonuç değişkeninin(y=0,1) çapraz tablosu oluşturulur. Ki-kare testinin k-1 serbestlik derecesiyle olabilirlik oranı, tek bağımsız değişkeni kapsayan lojistik regresyon modelindeki k-1 dizayn değişkeninin katsayılarının önemi için olabilirlik oran testinin değeriyle tamamen birbirine eşittir. Ek olarak, en azından orta derecede ilişki gösteren değişkenler için düzeylerden biri referans grup olarak kullanılarak her bir odds oranı, güven aralıklarıyla birlikte tahmin edilebilir.

Oluşturulan çapraz tabloda gözlenen frekans değeri sıfır olan bir yada birden fazla durum varsa odds oranlarının bazılarının nokta tahmini sıfır yada sonsuz olabilir.

Herhangi bir lojistik regresyon modeline böyle bir değişkenin dahil edilmesi istenmeyen sayısal sonuçları ortaya çıkarır. Bu durumu önlemek için birkaç yöntem mevcuttur. Bu yöntemler;

(a) Bağımsız değişkenin kategorilerini sıfırlı hücreyi ortadan kaldıracak şekilde birleştirmek,

(b) O kategoriyi tamamen iptal etmek,

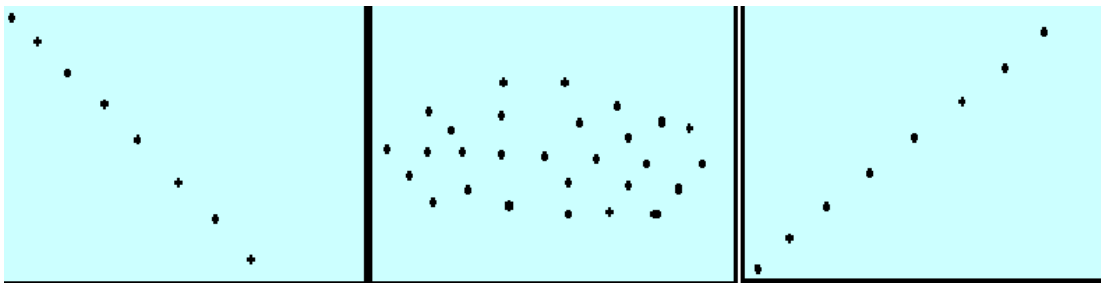
(c) Eğer değişken sıralama ölçme düzeyinde ölçülmüşse, o değişkeni sürekli değişken olarak modellemektir[11].

Sürekli değişkenler için tek değişkenli lojistik regresyon modelleri kurularak, bu değişkenler için tahmin edilen katsayılar $\hat{\beta}$, katsayıların standart hataları $S\hat{E}\left(\hat{\beta}\right)$ ve katsayıların öneminin belirlenmesi için olabilirlik oran testleri ve Wald test istatistiklerinin değerleri hesaplanır. Olabilirlik oran testi yardımıyla sadece sabit terimi kapsayan modele ilişkin log-olabilirlik değeri "L_0" belirlenir. Daha sonra çok değişkenli modelde yer alması düşünülen her bir değişkene ilişkin tek değişkenli analiz sonucunda modelde sadece bu değişkenlerin yer almasıyla elde edilen yeni log-olabilirlik değerleri "$L_j, \quad j = 1,2,...,p$" bulunur. Bulunan bu değerlerden yararlanarak her bir değişkeni içeren tek değişkenli model için G istatistikleri hesaplanır. Hesaplanan G istatistiklerinden, olabilirlik oran testi yardımıyla çok değişkenli modele girmeye aday değişkenlerin tek değişkenli analizleri sonucunda önemliliklerinin belirlenmesinde yararlanılır[15].

Alternatif bir analiz de bağımsız iki örnek ortalamasının karşılaştırıldığı tek değişkenli t-testidir. t-testi analizi genellikle grup ortalaması, standart sapmaları ve p-değerini içerir. Bu test bağımsız değişkenin modelde kalıp kalmayacağına karar vermek için kullanılabilir. Çünkü bu testten elde edilen p-değeri Wald istatistiği ya da olabilirlik oran testinden elde edilen p-değerleriyle aynı anlamda ve büyüklüktedir[11].

Sürekli değişkenler için tek değişkenli lojistik uyuma karar verilmesi saçılım grafiği(scatterplot) yardımıyla yapılabilmektedir. Bu grafik lojit ölçekle hazırlanmış olup, sadece değişkenin potansiyel önemini belirtmekle kalmayıp, değişkenin uygun ölçekte olup olmadığını belirler[11].

Saçılım(scatterplot) grafikleri iki değişken arasındaki ilişki hakkında genel bir bilgi edinmemizi sağlar. Ancak ilişkinin miktarı hakkında yorum yapabilmek için korelasyon katsayısının hesaplanması gerekir. Korelasyon katsayısı(r) iki değişken arasındaki ilişkinin ölçüsüdür ve -1 ile 1 arasında değişim gösterir.



(a) $r = -1$	(b) $r = 0$	(c) $r = 1$
Mükemmel negatif ilişki	İlişki yok	Mükemmel pozitif ilişki

Şekil 3.3. Saçılım grafiği(sctterplot) çeşitleri ve yorumları

Yukarıdaki saçılım grafikleri;

(a) Değişkenlerin birinin artışına bağlı olarak diğerinde azalma olan doğrusal ilişki olduğu

(b) İki değişken arasında ilişki olmadığı

(c) Değişkenlerden birisindeki artışa bağlı olarak diğerinde de artış olan doğrusal ilişki olduğu

şeklinde açıklanır[13].

(2) Tek değişkenli analizlerin ardından çok değişkenli analiz için değişken seçimi işlemine geçilir. Herhangi bir değişken için yapılan tek değişkenli test sonucunda elde edilen p değeri, 0,25' ten küçük($p<0,25$) ise o değişken istatistiksel olarak anlamlıdır denilir ve çok değişkenli modele girmeye aday olarak seçilir.

Değişken seçiminde p değerinin 0,25 olarak belirlenmesi, Bendel ve Afifi (1977)'nin doğrusal regresyon ve Mickey ve Greenland (1989)'ın lojistik regresyon çalışmalarına dayalıdır.

Tek değişkenli yaklaşımla ilgili karşımıza çıkan bir problem ise, sonuç değişkeni (bağımlı değişken) ile zayıf bir ilişkisi bulunan değişkenin, diğer değişkenlerle birlikte modele dahil edildiğinde bağımlı değişkenin önemli bir tahmin edicisi olabilmesidir. Böyle bir ihtimalin olduğundan şüphelenildiğinde bağımsız değişkenin çok değişkenli modele alınabilmesi için değişkenlerin anlamlılık düzeyi yeterince büyük seçilmelidir.

Genel olarak, olası tüm değişkenleri kapsayan çok değişkenli modele karar vermek için toplam örneklem sayısına ve modele girmesi muhtemel değişkenlerin sayısına göre her bir sonuç grubundaki örnek sayısına bakılır. Veriler böyle bir analiz için yeterli olduğunda çok değişkenli modellemeye başlamak yararlı olabilir. Eğer veriler yetersiz ise bu yaklaşım sayısal olarak istikrarlı olmayan çok değişkenli bir model üretebilir. Sonuçların sabit olmamasından dolayı ise değişkenlerin seçimi için Wald istatistiği kullanılmamalıdır[11].

Değişken seçimi için diğer bir yaklaşım adımsal (Stepwise) metotdur. Adımsal metotta değişkenlerin modele alınması ya da çıkarılması tamamen istatistiksel kriterler doğrultusundadır. Bu metodun iki farklı uygulaması vardır. Bunlar aşağıda kısaca özetlenmiştir[11].

(a) Geriye doğru eleme testli ileriye dönük seçim (Forward Stepwise): Belirlenen bir istatistiksel kriter(p-değeri) tarafından ölçülerek gruplar arasındaki en iyi ayrımı sağlayan değişken lojistik regresyon modeline dahil edilir. Bir sonraki adımda

fonksiyona girecek değişken, daha önceden belirlenmiş olan istatistiksel kriter tarafından ölçülerek lojistik regresyon modeli için en fazla ayırıcı güce sahip olan değişkendir. İleriye dönük seçim yöntemi, lojistik regresyon modeline dahil edilecek değişken kalmayana kadar devam eder.

(b) İleriye doğru seçim testiyle geriye dönük eleme (Backward Stepwise): Bu yöntemde lojistik regresyon modeline tüm değişkenler dahil edilerek başlanır. Her bir adımda belirlenen bir istatistiksel kriter tarafından ölçüm yapılarak, bağımlı değişkene en az etkiyi yapan değişken modelden çıkarılır. Geriye doğru eleme yöntemi, modelden daha fazla değişken atılamayana kadar devam eder.

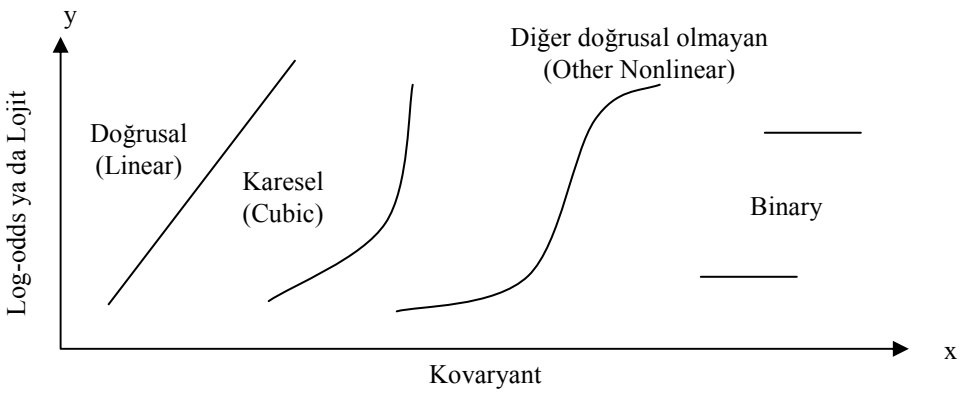
Lojistik regresyonda pek kullanılmayan alternatif bir seçim metodu da en iyi alt küme (Best Subset) seçim yöntemidir.

(3) Çok değişkenli modelin kurulmasıyla birlikte modele dahil edilen her değişkenin katsayısının önemliliği Wald test istatistiği ile test edilir. Bu kritere göre modele katkı sağlamayan değişkenler modelden çıkartılarak kalan değişkenlerle yeni bir model kurulur. Yeni model olabilirlik oran testi kullanılarak eski model ile karşılaştırılır.

Tek değişkenli analiz sonucunda modelde çok fazla bağımsız değişken kalıyorsa en iyi modeli elde edebilmek için adımsal metotları kullanmak yararlı olur. Örneğin ileriye doğru adımsal seçim işleminde modele dahil olma kriteri olarak p-değerinin 0,25 veya 0,50 den küçük olması şeklinde belirlenebilir. Bu süreç modelde sadece istatistiksel olarak önemli olan değişkenler kalıncaya kadar devam eder.

(4) Modelde kalacak değişkenler belirlendikten sonra bu değişkenler arasındaki etkileşim terimlerinin de modele dahil edilmesinin gerekli olup olmadığı araştırılmalıdır. Kesikli değişkenlerin kategorilerinin uygunluğu, tek değişkenli analiz sırasında incelenmiştir. Değişken seçimi aşaması sürekli bağımsız değişkenlerin lineer ilişki içinde olduğu varsayımı altında yapılır ve değişkenin

modelde kalıp kalmayacağına karar verme işlemiyle benzerdir. Sürekli bağımsız değişkenle lojit arasındaki birkaç farklı ilişki tipi Şekil 3.4 de gösterilmektedir.



Şekil 3.4. Sürekli bağımsız değişken ile lojit arasındaki birkaç farklı ilişki tipi

3.5.2. Adımsal lojistik regresyon

İleriye ve geriye doğru seçim yöntemlerinin bileşimi sonucunda adımsal seçim yöntemi ortaya çıkmıştır. Çok değişkenli lojistik regresyon modelinde katsayıların yorumu doğrusal (lineer) regresyon modeline benzer şekilde yapılır. Fakat çok değişkenli modelde herhangi bir x değişkeninin katsayısı, diğer bütün değişkenler sabitken logaritmik olabilirlik değerinde meydana gelen farkı verir. Bunun amacı, "ilgilenilen bağımsız değişkenin modele dahil edildiğinde mi yoksa dahil edilmediğinde mi model daha anlamlı olur?" sorusunun cevabını bulmaktır.

Değişken seçme işleminin adımsal olarak yapılması lineer regresyon analizinde yaygın olarak kullanılmaktadır. Yazılım programlarının çoğu ya ayrı bir program olarak yada bu analiz tipini gerçekleştirecek bir seçeneğe sahiptirler. Bir zamanlar model oluşturmak için kullanılan en yaygın yöntem olan adımsal lojistik regresyon, son yıllarda tümevarımsal metotlardan değişkenlerin amaçsal seçimlerine doğru bir kayma olmasından dolayı geri planda kalmıştır. Ancak adımsal lojistik regresyon analizinin halen en kullanışlı yöntem olduğu düşünülmektedir.

Hosmer, Wang, Lin ve Lemeshow(1978) yaptıkları çalışmalarda model yapılandırmak için sıklıkla kullanılan adımsal regresyon metodunun oldukça yararlı ve etkili bir analiz aracı olduğuna değinmişlerdir. Özellikle yapılan analizlerde üzerinde çalışılan bağımlı değişkenin yeni olması durumu ile karşılaşıldı zaman (Örneğin; AIDS) bağımsız değişkenlerin hangilerinin önemli olduğunun bilinmemesi problemi ile karşılaşılır. Bu gibi durumlarda olası birçok bağımsız değişken toplanır ve adımsal seçim işlemi uygulanarak hızlı ve etkili bir şekilde çok sayıda değişken arasından seçilen bağımsız değişkenlerle lojistik regresyon modelleri kurulur[12].

Bir modelden değişkenin seçilmesi ya da silinmesi için her adımsal işlem, değişkenin önemini kontrol eden ve belli bir kurala göre değişkenleri dahil eden yada çıkaran bir istatistiksel algoritmaya sahiptir. Değişkenin önemi, değişken katsayısının istatistiksel anlamlılığının ölçülmesi ile tanımlanır. Kullanılan istatistik modelin varsayımlarına bağlıdır. Hataların normal dağıldığı varsayımı altında, doğrusal adımsal regresyonda F-testi kullanılır. Lojistik regresyonda ise hataların binom dağılımına sahip olduğu varsayılır ve anlamlılığı ki-kare testi olabilirlik oranına göre değerlendirilir. İstatistiksel olarak en önemli değişken, yöntemin herhangi bir adımında o değişkeni kapsamayan bir modelle kıyaslanması sonucunda en büyük değişimi sağlayan değişkendir. Yani adımsal işlemler yapılırken herhangi bir adımda olabilirlik oranı test istatistiği(G) en büyük olan değişken en önemli değişkendir şeklinde ifade edilir. G istatistiğinden yararlanılarak değişkenin önemli olup olmadığına karar verme işlemi p değeri aracılığıyla yapılır.

Adımsal lojistik regresyon analizinin her bir adımının neler olduğu, paket programlar tarafından gerçekleştirilen istatistiksel adımlar da göz önünde bulundurularak açıklanacaktır. Adımsal lojistik regresyon analizinde ileriye doğru seçimin ardından geriye doğru eleme algoritması p tane olası bağımsız değişkenin olduğu varsayımı altında şu şekildedir:

Adım(0): Sonuç değişkenlerini incelemek için istatistiksel öneme sahip p tane bağımsız değişken olduğu varsayılsın. Bu adım yalnızca sabit terimin bulunduğu

model kurularak ve kurulan bu modelin log-olabilirliği (L_0) elde edilerek başlar. Daha sonra her bir p değişken için tek değişkenli lojistik regresyon modeli kurulur ve bu modellerin log-olabilirlikleri (L_0) karşılaştırılır. Adım (0)' da x_j değişkenini içeren modelin log-olabilirlik değeri $L_j^{(0)}$ ile gösterilsin. (j) indisi modele eklenen değişkeni ifade ederken, üstte bulunan *(0)* adım sayısını yani seviyeyi ifade eder. Bu notasyonların yardımıyla adımsal lojistik regresyon konusu süresince adım sayısı ve değişkenler takip edilecektir.

x_j değişkenini içeren modelin sadece kesim noktasını kapsayan modele karşı olabilirlik oran test değeri $G_j^{(0)} = 2\left(L_j^{(0)} - L_0\right)$ ile gösterilir ve bunun gözlenme olasılığı $p_j^{(0)}$ ile ifade edilir. Bu p-değeri $p_j^{(0)} = \Pr\left[\chi^2_{(v)} > G_j^{(0)}\right]$ yardımı ile elde edilir. Eğer x_j sürekli bir değişken ise $v = 1$, x_j k-düzeyi olan kesikli bir değişkense $v = k - 1$ olarak kabul edilir. Modele dahil edilecek en önemli değişken en küçük p değerine sahip değişkendir. Eğer bu değeri x_{e_1} olarak ifade edersek $p_{e_1}^{(0)} = \min\left(p_j^{(0)}\right)$ olur. Buradaki e_1 indisi birinci sırada modele dahil olmaya aday değişken olarak tanımlanır. Örneğin, değişkenler arasında x_2 en küçük p değerine sahipse, $p_2^{(0)} = \min\left(p_j^{(0)}\right)$ olur ve $e_1 = 2$ değerini alır. Çünkü x_{e_1} değeri en önemli değişkeni ifade etmektedir. Fakat bu değişkenin istatistiksel olarak anlamlı olduğunun bir garantisi yoktur. Örneğin, $p_{e_1}^{(0)} = 0{,}83$ ise devam eden analizde daha küçük bir değerin olduğu sonucu çıkarılabilir. Bunun sebebi en önemli değişkenin sonuç değişkeni ile ilgili olmamasından kaynaklanmaktadır. Başka bir deyişle $p_{e_1}^{(0)} = 0{,}003$ ise bu değişkeni içeren lojistik regresyon modeli incelenir ve daha sonra modelde x_{e_1} olarak verilen diğer değişkenlerin önemli olup olmadığına bakılır.

Adımsal lojistik regresyon analizinde değişkenin önemli olup olmadığına karar verirken "alfa(α)" anlamlılık düzeyinin seçimi kritik bir karardır. p_E gerekli α düzeyinde modele kaç tane değişken dahil edileceğini belirlesin. Bendel ve Afifi(1977) p_E'nin seçimi konusunda çalışmalar yapmıştır. Bu çalışmaların

sonucunda $p_E = 0{,}05$ olarak belirlendiğinde çoğu zaman model için önemli olan değişkenlerin modelin dışında kaldığı gösterilmiştir. Bu yüzden p_E değerinin 0,15 ile 0,20 arasında seçilmesi önerilmiştir.

Bazen analizin amacı daha geniş olabilir ve daha fazla değişken içeren modeller, olası modeli daha iyi tanımlayabilir. Bu durumda p_E değerinin 0,25 ve üzerinde alınması daha doğru olabilir. p_E için hangi değer alınırsa alınsın, G için p değeri p_E'den daha küçük bir değer alıyorsa, o değişken modelde yer almak için yeterli öneme sahip kabul edilecektir. Bu sebeple $p_{e_1}^{(0)} < p_E$ ise program adım(1) e ilerler, aksi takdirde durur.

Adım(1): Adım (1), x_{e_1} değişkenini kapsayan lojistik regresyon modelinin kurulmasıyla başlar. $L_{e_1}^{(1)}$ bu modelin log-olabilirliğini göstermektedir. x_{e_1} modeldeyken geriye kalan p-1 değişkenin model için önemli olup olmadığına bakmak üzere x_{e_1} ve x_j' yi $\left(j = 1,...,p \;\; ve \;\; j \neq e_1\right)$ içeren p-1 tane lojistik regresyon modeli kurulur. x_{e_1} ve x_j'yi içeren modelin log-olabilirliği $\left(L_{e_1 j}^{(1)}\right)$ ve sadece x_{e_1}' i içeren modelin ki-kare oran test istatistiği $G_j^{(1)} = 2\left(L_{e_1 j}^{(1)} - L_{e_1}^{(1)}\right)$ ile gösterilir. $p_{e_2}^{(1)} = \min\left(p_j^{(1)}\right)$ olduğunda, adım(1) deki en küçük p değerine sahip değişken x_{e_2} olsun. Eğer bu değer p_E değerinden küçük ise ikinci adıma geçilir, yoksa işlem durur.

Adım(2): Adım (2), x_{e_1} ve x_{e_2} değişkenlerini içeren modelin kurulmasıyla başlar. x_{e_2} değişkeninin modele eklenmesiyle x_{e_1} değişkeninin önemini yitirmesi mümkündür. Bu nedenle adım (2) geriye doğru eleme kontrolünü de içerir. Genel hatlarıyla bu işlem, bir önceki adımda eklenen değişkenlerden birinin silinerek modelin kurulması ve silinen değişkenin önem kontrolünün yapılmasıdır. Bu işlemi yapmak için adım (2)'de x_{e_j} değişkeni modelden çıkarıldıktan sonra $L_{-e_j}^{(2)}$ notasyonu ile gösterilen

modelin log-olabilirliği bulunur. $p = p_{-e_j}^{(2)}$ olasılık değeri ile x_{e_j} değişkeninin tam modelle karşılaştırılmasının olabilirlik oran testi $G_{-e_j}^{(2)} = 2\left(L_{e_1e_2}^{(2)} - L_{-e_j}^{(2)}\right)$ hesaplanır.

Program modelden bir değişkenin çıkarılıp çıkarılmayacağını kesinleştirmek için, modelden çıkarıldığında en büyük p-değerine sahip olan değişkeni seçer. Bu değişkeni x_{r_2} ile gösterirsek, $p_{r_2}^{(2)} = \max\left(p_{-e_1}^{(2)} - p_{-e_2}^{(2)}\right)$ olur. x_{r_2} değişkeninin modelden çıkarılıp çıkarılmayacağına karar vermek amacıyla, modele katkıyı devam ettirmek için minimal düzeyi " p_R " , modelden çıkarılmayı ise "R" indisi ifade eder. $p_{r_2}^{(2)}$ değeri de önceden belirlenmiş olan ikinci bir "alfa" düzeyiyle karşılaştırılır. Programın diğer adımlarında, aynı değişkenin modele dahil edilmesini veya modelden çıkarılmasını önlemek için, p_R değeri daima p_E değerinden büyük olarak seçilmelidir.

Eğer modele alınan değişkenlerden birçoğunun tekrar modelden çıkarılmaması isteniyorsa p_R=0,9 değeri kullanılır[12]. Modele devamlı bir katılım gerekiyorsa daha etkili bir değer kullanılmalıdır. Örneğin, $p_E = 0{,}15$ yerine $p_E = 0{,}20$ değeri seçilebilir. Değişkenin model içindeki durumuna karar verirken maksimum p-değeri ile p_R değerinin karşılaştırılması gereklidir.Bu karşılaştırma sonucunda;

$p_{r_2}^{(2)} > p_r$ ise x_{r_2} değişkeni modelden çıkarılır

$p_{r_2}^{(2)} < p_r$ ise x_{r_2} değişkeni modelde kalır.

Her iki durumda da program değişken seçme işlemine devam eder.

Geriye doğru eleme işleminden sonra x_{e_1} ve x_{e_2} değişkenleri hala modelde ise; ileriye doğru seçim safhasına geçilerek x_{e_1} , x_{e_2} ve x_j ' yi $\left(j = 1,...,p \;\; ve \;\; j \neq e_1, e_2\right)$ kapsayan p-2 tane lojistik regresyon modeli kurulur. Program her bir model için log-olabilirlik değerini hesapladıktan sonra, yalnızca x_{e_1} ve x_{e_2} ' yi kapsayan modele

karşı olabilirlik oran testlerini hesaplar ve karşılık gelen p-değerlerini bulur. x_{e_3} minimum p-değerine sahip değişken olsun $\left(p_{e_3}^{(2)} = \min\left(p_j^{(2)}\right)\right)$.Eğer $p_{e_3}^{(2)} < p_E$ ise işlem adım(3)'e geçer aksi taktirde durur.

Adım(3): Adım (3), adım (2) ile aynı işleyişe sahiptir. Program bu aşamada geriye doğru eleme kontrolünü yapar ve ileriye doğru seçim işlemine geçer. Bu işlem aynı mantıkla son basamak olan adım (s)' e kadar devam eder.

Adım(s): Bu adıma iki şekilde geçilir.

(1) Bütün değişkenlerin p-değerleri, değişkenleri modele dahil edebilecek bir seviyede olduğunda

(2) Modeldeki tüm değişkenlerin p-değerleri p_R değerlerinden küçük olduğunda ve modele dahil edilmeyen değişkenlerin modele girebilmeleri için gerekli değerleri p_E değerinden büyük olduğunda geçilir.

Bu adımdaki modelde p_E ve p_R kriterine göre anlamlı olan değişkenler bulunur. Eğer p_E ve p_R'nin değerleri istatistiksel önemliliği belirlemek için daha az güvenilir değerler olarak seçilseydi (0,05 yerine 0,25), o zaman final modeli için değişkenler, adımsal regresyon işleminin sonuçlarını özetleyen bir tablodan seçilecekti.

Özet tablosundan değişken seçmenin iki yöntemi vardır. Bunlar adımsal doğrusal regresyonda genel olarak kullanılan metotlarla uyum göstermektedir. İlk metot, her adımda modele giriş yapan değişkenlerin p-değerine bağlı iken, ikinci metot şu an ki adımdaki modele karşılık son adımdaki modelin olabilirlik oran testine bağlıdır.

İşlemdeki herhangi bir adımın "q" ile gösterildiği varsayılsın. Birinci metotta, $p_{e_q}^{(q-1)}$ değeri daha önceden seçilmiş olan $\alpha = 0{,}15$ gibi bir anlamlılık düzeyi ile

karşılaştırılır. Eğer $p_{e_q}^{(q-1)}$ değeri α değerinden küçükse q adımına geçilir. $p_{e_q}^{(q-1)}$ değeri α değerinden büyükse işlem o adımda sona erer. Bu metotta da modele girme kriteri $x_{e_1}, x_{e_2}, ...x_{e_{q-1}}$ değişkenlerinin modelde olma şartı altında x_{e_q}' nun katsayısının önem testine bağlıdır. Testin serbestlik derecesi x_{e_q}' nun sürekli veya k kategorili kesikli bir değişken olmasına göre 1 ya da k-1' dir.

İkinci metotta model, şuan ki adım(q) ile bir önceki adım(q-1) değil, en son adım olan adım(s) ile karşılaştırılır. Bu iki modelin olabilirlik oran testini bulmak için p-değeri bulunur ve bulunan bu p-değeri α değerini geçinceye kadar işleme devam edilir. Bu şekilde adım(q)' dan adım(s)' e kadar modele eklenen değişkenlerin katsayılarının sıfıra eşit olup olmadığı test edilmiş olur. Her adımda serbestlik derecesi, birinci metottaki teste göre daha fazla olur. Yani ikinci metot birinci metoda göre daha fazla değişkeni modele dahil etme eğilimindedir. Bu yüzden de birinci metot daha fazla tercih edilmektedir.

Adımsal seçim işleminde hesaplanan p-değerinin geleneksel hipotez testinde kullanılan p-değerinden farklı olduğu iyi bilinmektedir. Burada kullanılan p-değerinin değişkenler arasındaki önemin göstergesi olduğu düşünülmelidir. Zengin bir model için adımsal seçim yöntemi tavsiye edilebilir.

Adımsal seçim işleminin en belirgin özelliği, daha önceden önemli olarak bilinen değişkenleri adım(0)' da modele dahil ederek başlıyor olmasıdır.

Adımsal seçim işleminin göz önünde bulundurulması gereken en önemli dezavantajı ise, her adımda modelde olmayan tüm değişkenlerin en çok olabilirlik tahminlerinin hesaplanmasıdır. Değişken sayısının çok olduğu büyük veri kümeleri için bu oldukça zaman kaybettirir ve maliyeti arttırabilir.

Freedman(1983) birçok değişkeni olan modelleri incelemiş ve doğrusal regresyon analizinin sonucunda modelin, önemli olarak belirlenen değişkenden değil, daha düşük önem arz eden değişkenlerden oluşabileceğine dikkat çekmiştir. Flack ve

Chang(1987) önemli değişkenlerin seçilme sıklığını incelmiş ve aynı sonuca varmışlardır.

Özetle adımsal seçim işlemi istatistiksel temellere dayanarak model için aday değişkenleri seçmektedir. Adımsal seçim yönteminin bu özelliği ise regresyon analizinde değişken seçimi için popüler bir yöntem olmasını sağlamıştır[12].

3.6. Model Uyumluluğunun Belirlenmesi

Modele gerekli tüm değişkenler alındıktan sonra tahmin edilen lojistik regresyon modelinin sonuç değişkenini tanımlamakta ne kadar etkili olduğunu uyum iyiliği testiyle test edilir[16].

Modelin uyum iyiliğine karar verilmek isteniyorsa ilk olarak model uyumuyla ne ifade edilmek istendiği hakkında bazı fikirlere sahip olunmalıdır. Bağımlı değişkeninin gözlenen değerleri $\mathbf{y}$ vektörü ile gösterilsin ve $\mathbf{y}'=(y_1,y_2,...,y_n)$ olsun. Model tarafından tahmin edilen değerler $\hat{\mathbf{y}}$ ile gösterilsin ve $\hat{\mathbf{y}}' = (\hat{y}_1, \hat{y}_2, ..., \hat{y}_n)$ olsun. Eğer $\mathbf{y}$ ve $\hat{\mathbf{y}}$ arasındaki uzaklık özet ölçüleri küçükse ve her bir $(y_i, \hat{y}_i)$, i=1,2,3...,n ikilisinin bu özet ölçülere katkısı sistematik değil ve modelin hata yapısına göre küçükse modelin uyumlu olduğuna karar verilir. Böylece, uygun bir modele tamamen karar vermek için hem $\mathbf{y}$ ve $\hat{\mathbf{y}}$ arasındaki uzaklığın özet ölçüsünün hem de bu ölçülerin her bir parçasının teker teker incelenmesi gereklidir[12].

Uyum iyiliği testinde kullanılan farklı istatistikler vardır. Bunlar; Hosmer-Lemeshow(G) istatistiği, ki-kare istatistiği, -2logL istatistiği, Pearson ki-kare istatistiği ve blok ki-kare istatistiğidir. Lojistik regresyon analizinin yapıldığı tüm paket programlar yukarıda saymış olduğumuz uyum iyiliği test istatistiklerinden en az bir tanesini içermektedir. Bu tezde uygulama kısmında yararlanılacak olan Hosmer-Lemeshow(G) istatistiği üzerinde ayrıntılı olarak durulmuştur.

3.6.1. Hosmer-Lemeshow (G) İstatistiği

Hosmer ve Lemeshow (1980), Lemeshow ve Hosmer (1982) tahmin edilen olasılık değerlerinin gruplandırılmasını önermişlerdir. J=n olduğu durumlarda veri matrisindeki n tane sütun, n tane tahmin edilen olasılık değerine karşılık gelmektedir. Aynı bağımsız değişken değerine sahip olan bireylerin olduğu durumda bunlardan sadece bir tanesi seçilir ve toplam birey sayısı "J" ile ifade edilir. Burada gruplama yapmaktaki amaç, mevcut dağılımı ki-kare dağılımına yaklaştırarak anlamlı ve yorumlanabilir bir model elde etmektir. Gruplama iki farklı şekilde yapılabilir:

(a) Tahmin edilen olasılıkların yüzdesi dikkate alınarak

(b) Tahmin edilen olasılıkların sabit değerleri dikkate alınarak.

Birinci metot gözlenen ve tahmin edilen beklenen frekansları karşılaştıran $\hat{C}$,

$$\hat{C} = \sum_{a=1}^{g} \sum_{l=1}^{10} \frac{(o_{kl} - e_{kl})^2}{e_{kl}}$$

istatistiğine dayanmaktadır. *g* grup sayısını, *o* gözlenen değeri, *e* beklenen değeri, *l* ise risk grubunu temsil etmektedir. Burada 10'lu risk grubu kullanılmaktadır. Yani tüm gözlemler 10 gruba ayrılır. Bu yöntem yeterli sayıda gözlem olduğunda geçerlidir. Olumsuz yönü ise gerçek değerlerin göz ardı edilmesidir[12].

İkinci metotta sabit kesim noktaları üzerine bir gruplandırma yapılmaktadır. a/10, a=1,2,…,9 değerleriyle tanımlanmış olan kesim noktaları kullanılarak 10 grup oluşturulmaktadır. Bu gruplar tahmin edilen olasılıklara göre sabit grupların oluşturularak bireylerin ilgili grupta yer almasını sağlamaktadır. Örneğin; birinci grup tahmin edilen olasılık değeri 0,1' e eşit veya daha küçük olan tüm bireyleri kapsarken, onuncu grup ise tahmin edilen olasılık değeri 0,9' a eşit veya daha küçük olan tüm bireyleri kapsamaktadır.

Gruplandırma yoluyla verileri azaltma işleminde, gruplardaki veri sayısının azalmasından dolayı uyumdan sapmalar görülebilir. Hosmer-Lemeshow testi yorumlama ve geniş veri kümesini rahat çözmek amacını hedefleyen hem en yaygın hem de en çok tercih edilen bir uyum testidir. Bu istatistik SPSS'de "Hosmer-Lemeshow G" olarak bilinmesinin yanı sıra Model ki-kare istatistiği olarak da adlandırılır.

Ayrıca geliştirilen bir modelin geçerliliğinin değerlendirilmesi için genellikle gerçek olasılıklar ile tahmin edilen olasılıklar arasındaki standart farka bakılır. Yeni modelin hangi açıdan etkili olduğu, ilişki ölçümü, değişik hataların dağılımları gibi göstergeler incelenir. Lojistik regresyonda hesaplanabilen hata türleri: standart hata, standart olmayan hata, sapma değeri, uzaklık değeri, cook uzaklığı ve DfBeta değeridir.

4. BOOTSTRAP YÖNTEMİ

İstatistikte herhangi bir yığının parametresini tahmin etmek için o yığına ait gözlemlerden yararlanılır. Üzerinde çalışılan yığının tüm gözlemlerini parametre tahmini için kullanmak hem zaman kaybına yol açacağı hem de maliyeti arttıracağı için yığını en iyi şekilde açıklayacak olan örnekten elde edilen verilerle çalışmak bu sorunları ortadan kaldırır. İstenilen büyüklük ve miktarda veri setleri oluşturmak için herhangi bir boyuttaki veri setinden gözlemler tesadüfi yer değiştirilerek yeniden örneklenebilir. Bu sayede veri setinden daha fazla bilgi alınabilir. Bu şekilde tanımlanan yöntem "*Bootstrap Yöntemi*" olarak adlandırılır[1].

Bootstrap yöntemi literatüre ilk kez Efron 'ın 1979 yılındaki makalesi ile tanıtılmıştır. Teorik gelişme Freedman (1981) ve Wu (1986) ile devam etti. Daha sonraki gelişmelerden kitaplaştırılanlar ise tarihsel sırasıyla Beran ve Ducharme (1991), Hall (1992), Mammen (1992), Efron ve Tibshirani (1993), Davison ve Hinkley (1997) ve teorik bir çalışma olan Shao ve Tu (1995)'dur[1].

Günümüzde bilgisayarların da gelişimiyle beraber çok sayıda araştırmaya konu olan bootstrap yönteminde temel düşünce, eldeki örneklemi yığın olarak varsayıp buradan belirli sayıda tekrarlı örnekleme yaparak ilgilenilen tahmin edicinin suni bir örnekleme dağılımını yaratmaktır[1].

Bir yeniden örnekleme tekniği olan bu yöntemden, sadece tahmin değerlerini ve standart hataları belirlemenin yanı sıra birçok alanda da faydalanılmaktadır. Zaman Serileri Analizi, Lineer Olmayan Regresyon Analizi, Kümeleme Analizi, Diskriminant Analizi, Lojistik Regresyon Analizi ve her türlü hipotez testini sınamak için kullanılabilmektedir.

Bu yöntemin temeli mevcut veri setinden çok daha büyük veri setleri üretmek için yeniden örnekleme yapmaktır. Bootstrap yönteminin geliştirilme amacı, örneklemin ortalamasını, standart hatasını hesaplamak ve güven aralıklarını oluşturmak olarak özetlenebilir.

Tesadüfi olarak gözlemlenmiş n tane gözlemin olasılık dağılımı F ile gösterilecek olur ise,

$$F \to (x_1, x_2, ..., x_n) \tag{4.1}$$

şeklinde ifade edilir. Her bir gözlemin seçilme olasılığı $\frac{1}{n}$ olan dağılıma deneysel dağılım denir ve bu dağılım $\hat{F}$ ile gösterilir. $\hat{F}$ deneysel dağılımından yerine koyma yöntemiyle seçilmiş n birimlik tesadüfi bir örnek,

$$x^* = (x_1^*, x_2^*, ..., x_n^*)$$

şeklinde gösterilecek olup bootstrap örneği olarak tanımlanır. İfadede bulunan yıldız işaretleri, gözlemlenmiş gerçek değerlerin içinden yerine koyma yöntemiyle oluşturulmuş olan örneği temsil etmektedir.

Herhangi bir $\hat{\theta} = s(x)$ istatistiğini hesaplamak için ele alınan n adet gözlemden meydana gelen veri seti $x = (x_1, x_2, ..., x_n)$ ise, orijinal veri setinde gözlemlerin $\frac{1}{n}$ olasılıkla tesadüfi olarak yerine koyma yöntemiyle seçilmesi ile bootstrap örnek veri seti $x^* = (x_1^*, x_2^*, ..., x_n^*)$ elde edilmektedir. $x = (x_1, x_2, ..., x_n)$ orjinal gözlemlerinden yapılan seçim sonucunda, bu gözlemlerin bazıları bootstrap örnekleminde birden fazla olabileceği gibi bazıları ise hiç bulunmayabilir. Bu işlem istenildiği kadar tekrarlanarak birbirinden farklı B tane bootstrap gözlemler seti oluşturulabilmektedir. İlgili istatistik ise bu yeni veri setleri kullanılarak hesaplanmaktadır.

Örneğin, n birimlik bir örneklemin gözlem değerleri(x^*), bootstrap örneklemi olmak üzere $\hat{\theta} = s(x)$ istatistiğinin standart hatası için bootstrap algoritması aşağıdaki gibi hesaplanmaktadır.

1. $x = (x_1, x_2, ..., x_n)$ veri setinden, n birimlik yerine koyma yöntemiyle seçilmiş B adet $x^{*1}, x^{*2}, ..., x^{*B}$ olarak adlandırılan bootstrap örneklemi oluşturulur.

2. Her bir bootstrap örnekleminin standart sapması

$$\hat{\theta}^*(b) = s\left(x^{*b}\right) \qquad b = 1,2,...,B$$

ile gösterilir. Burada "s" ile gösterilen ifade standart hatadır.

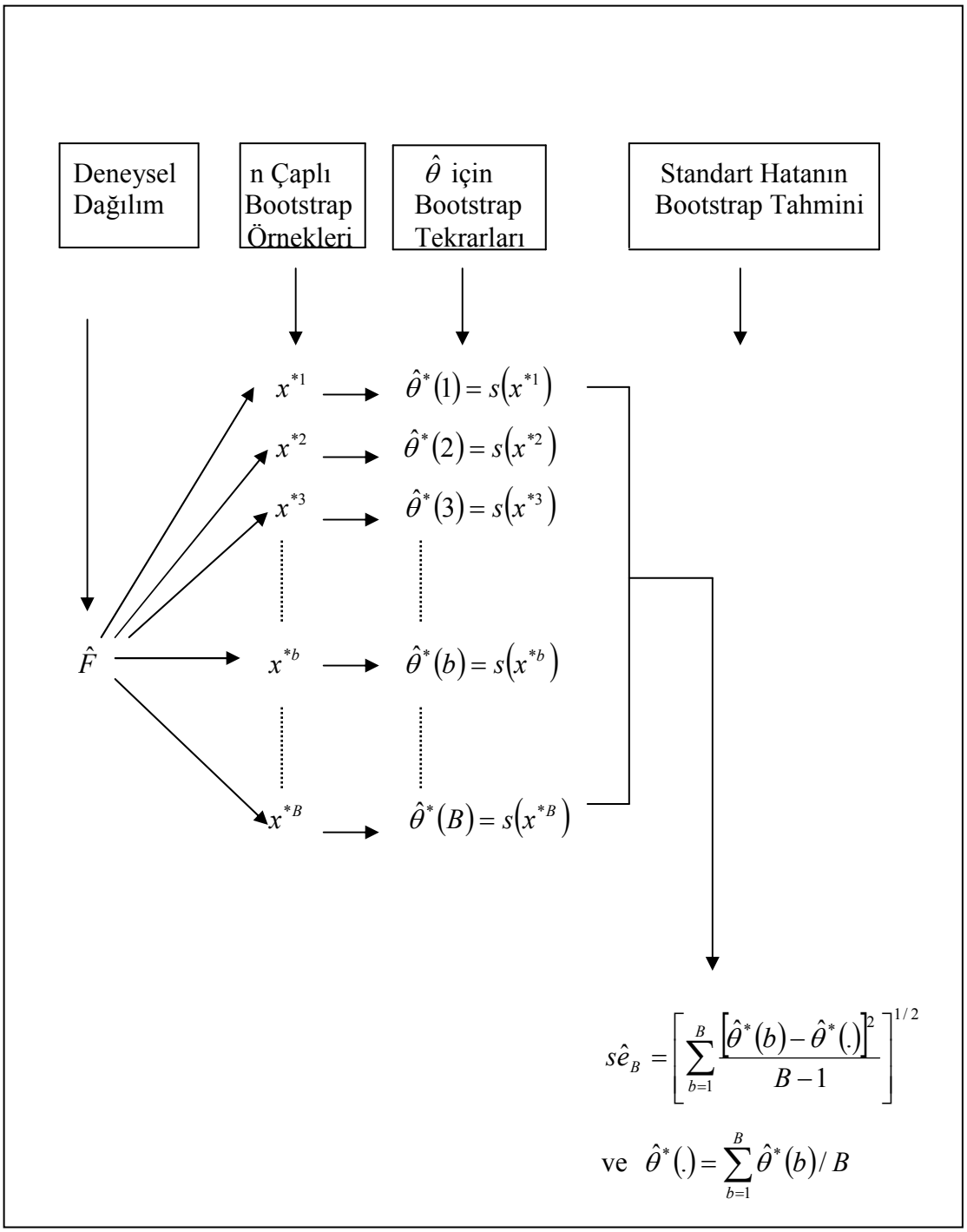
3. $\hat{\theta}$ istatistiğinin standart hatası olan $se_F\left(\hat{\theta}\right)$, B bootstrap örneklerinin sayısı olmak üzere,

$$s\hat{e}_B = \left\{ \sum_{b=1}^{B} \left(\hat{\theta}^*(b) - \hat{\theta}^*(.)\right)^2 / (B-1) \right\}^{1/2}$$

ile tahmin edilir.

Burada $s\hat{e}_B$, bootstrap örneklerinin örnek standart hatası olarak adlandırılır ve

$\hat{\theta}^*(.) = \sum_{b=1}^{B} \hat{\theta}^*(b) / B$ 'dır.

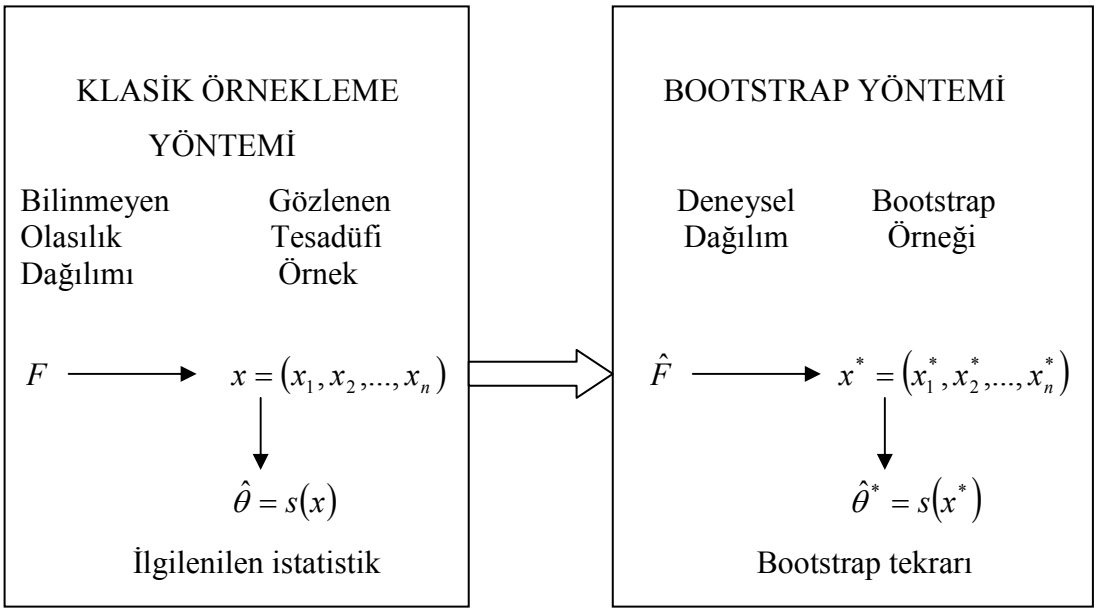


Şekil 4.1. $\hat{\theta} = s(x)$ istatistiğinin standart hata tahmini için bootstrap algoritması

Şekil 4.1' de görülen algoritma, standart hatanın bootstrap yöntemi ile hesaplanmasını açık bir biçimde anlatmaktadır. İlk adımda gözlemlenen değerlerden bootstrap örnekleri oluşturularak her örnek için standart hata tahmin değerleri

bulunmuştur. Daha sonra hesaplanan standart hata tahminlerinin ortalaması bulunmuş ve standart hata tahmin değerinden, hesaplanan ortalama standart hata değerler farkının karesi alınarak sapma miktarı elde edilmiştir. En son adımda, sapmaların karelerinin toplamı, bootstrap örnek sayısının bir eksiğine bölünmüştür.

4.1. Tek Örnekli Veri Setinde Bootstrap Tekniği



Şekil 4.2. Tek örnekli bir problem için bootstrap tekniği

Şekil 4.2' te tek örnekli bir problemde bootstrap metodunun uygulanışının şeması gösterilmektedir. Şeklin sol tarafı F dağılımına sahip tesadüfi örnekleme yoluyla gözlemlenen $x = (x_1, x_2, ..., x_n)$ veri seti kullanılarak elde edilmeye çalışılan istatistik değerinin verildiği klasik örnekleme yöntemidir. $\hat{\theta}$'nın istatistiksel özellikleri hakkında bilgi edinmek için $\hat{\theta} = s(x)$ istatistiği incelenir ve bu istatistiklerden biri de $se_F(\hat{\theta})$ standart hata değeri olabilir.

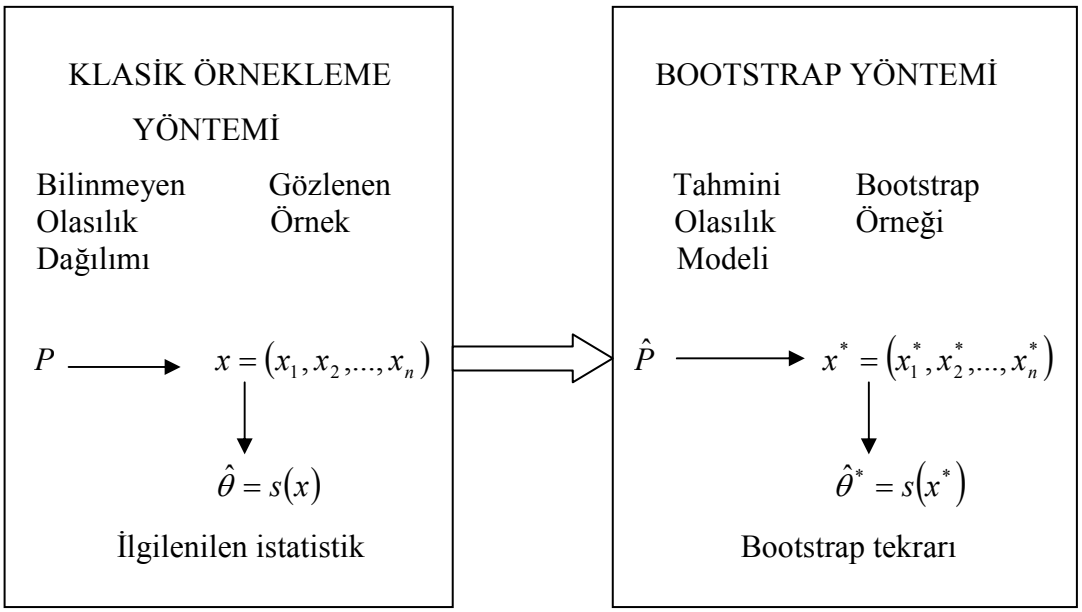
Şeklin sağ tarafında ise $\hat{F}$ dağılımından tesadüfi olarak $x^* = (x_1^*, x_2^*, ..., x_n^*)$ şeklinde bootstrap örneği oluşturularak $\hat{\theta}^* = s(x^*)$ istatistiği incelenmiştir.

Klasik örnekleme yönteminde sadece bir veri seti üzerinden hesaplama yapılabilirken, bootstrap yönteminin en büyük avantajı $\hat{\theta}^*$'ın istenilen sayıda hesaplanabilmesidir.

Daha karmaşık veri yapılarını kolaylaştırmak için,

$$P \to x \tag{4.2}$$

kullanılacaktır. Bu gösterimde bilinmeyen bir olasılık modeli olan P' nin gözlenen veri setinin x olduğu anlatılmaktadır.



Şekil 4.3. Tek örnekli bir problem için P olasılık dağılımı için bootstrap tekniği

Şekil 4.3' te tek örnekli bir problemde bootstrap metodunun uygulanışının şeması gösterilmektedir. Şeklin sol tarafı bilinmeyen P olasılık dağılımlı $x = (x_1, x_2, ..., x_n)$ veri seti kullanılarak tahmin edilen $\hat{\theta}$ istatistiğinin ifade edildiği klasik örnekleme yöntemidir. Şeklin sağ tarafında ise $\hat{P}$ olasılık dağılımından bootstrap yöntemiyle

tesadüfi olarak örneklenmiş $x^* = (x_1^*, x_2^*, ..., x_n^*)$ bootstrap örneği oluşturulmuş ve $\hat{\theta}^*$ istatistiği tahmin edilmiştir.

4.2. İki Örnekli Veri Setinde Bootstrap Tekniği

Eş. 4.2' de gösterilen P' nin F ve G gibi iki tane olasılık dağılımından oluştuğu düşünüldüğünde,

$$P = (F, G)$$

ile ifade edilir. Eş. 4.2' deki x veri setinin $z = (z_1, z_2, ..., z_m)$ ve $y = (y_1, y_2, ..., y_n)$ gözlemlerinden oluştuğu kabul edilirse,

$$x = (z, y) \tag{4.3}$$

şeklini alır. z veri seti F dağılımına y veri seti ise G dağılımına sahip olup,

$$F \to z \qquad \text{bağımsız} \qquad G \to y$$

olarak gösterilir. İki örnekli veri setinde bootstrap tekniğinin kullanımı için (z, y) veri setleri ayrı ayrı düşünülerek kendi bootstrap örnekleri oluşturulur. Daha sonra oluşturulan bu örneklerin birleştirilmesiyle x veri seti elde edilir. Burada,

$$\hat{P} = (\hat{F}, \hat{G})$$

ile gösterilir ve veri setinin bootstrap örneklemi

$$x^* = (z^*, y^*) \tag{3.4}$$

şeklinde ifade edilir. Böylece veri seti,

$$x^* = \left(z^*, y^*\right) = \left(z_{i1}, z_{i2}, ..., z_{in}, y_{j1}, y_{j2}, ..., y_{jm}\right) \tag{3.5}$$

olarak yazılabilir.

İki örnekli veri setinde bootstrap tekniğinin tek örnekli veri seti için bootstrap tekniğinden farkı, toplam gözlem sayısı m+n olmasına karşın n gözlem ve m gözlemin kendi içerisinde yerine koyma yöntemiyle tesadüfi olarak seçime tabi tutulmasıdır.

Bootstrap yönteminin iki farklı şekli bulunmakta olup bunlar, parametrik bootstrap yöntemi ve parametrik olmayan bootstrap yöntemidir.

4.3. Parametrik Bootstrap Tekniği

Parametrik yöntem, $\hat{\theta}$' nın örnekleme dağılımını ve varyansını tahmin etmenin en doğru şekli olarak tanımlanır. İlk olarak $\hat{f}_\theta(x)$ olasılık yoğunluk fonksiyonuna göre n büyüklüğünde B tane örnek çekilir. Her bir örnek varyansı, $\hat{\theta}$'nın varyansını tahmin etmeye yarar. Bu süreç parametrik bootstrap olarak adlandırılır.

4.3.1. Parametrik bootstrap tekniğinde en çok olabilirlik

Gözlemlerimiz için tanımlanan olasılık yoğunluk fonksiyonu,

$$X \sim f_\theta(x) \tag{4.6}$$

şeklinde gösterilir. Burada θ, X' in dağılımı belirleyici bir ya da birden fazla bilinmeyen parametreyi ifade eder. Ayrıca bu ifade, X için parametrik model olarak da adlandırılmaktadır. θ' nın eleman sayısı p ile gösterilir ve X'in μ ortalamalı ve σ^2 varyanslı bir normal dağılıma sahip olduğu varsayıldığında,

$\theta = (\mu, \sigma^2)$

'den p=2 olur. Buradan dağılımın olasılık yoğunluk fonksiyonu,

$$f_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-1/2\left(\frac{x-\mu}{\sigma}\right)^2}$$

şeklinde ifade edilir. En çok olabilirlik fonksiyonu ise,

$$L(\theta; x) = \prod_1^n f_\theta(x_i) \tag{4.7}$$

ile gösterilir. $L(\theta; x)$, θ'nın bir fonksiyonu şeklinde düşünülebilir. $L(\theta; x)$'in logaritması,

$$\ell(\theta; x) = \sum_1^n \ell(\theta; x_i) \tag{4.8}$$

şeklinde yazılır ve kısaca $\ell(\theta)$ olarak gösterilir. Bu ifade log-olabilirlik olarak ifade edilir ve her bir değer $\ell(\theta; x_i) = \log f_\theta(x_i)$ şeklinde olup log-olabilirliğin bir bileşeni olarak adlandırılır. En çok olabilirlik yönteminde $\ell(\theta; x_i)$' yi maksimum yapmak için $\theta = \hat{\theta}$ kabul edilir. $\hat{\theta}$'nın örnekleme dağılımı ve varyansını en doğru şekilde tahmin etmek için tanımlanan yönteme parametrik yöntem denir. Bu yöntemde, $\hat{f}_\theta(x)$ olasılık yoğunluk fonksiyonlu n büyüklüğünde B tane örnek seçilir. Seçilen her bir örneğin varyansı $\hat{\theta}$'nın varyansını tahmin etmek için kullanılır. Bütün bu adımlar ise parametrik bootstrap olarak adlandırılır.

4.4. Parametrik Olmayan Bootstrap Tekniği

Parametrik olmayan bootstrap tekniği ile parametrik bootstrap tekniği arasındaki en önemli fark, parametrik bootstrap tekniği için parametrik bir modelin olmasıdır. Parametrik olmayan bootstrap yöntemi için böyle bir model söz konusu değildir. Herhangi bir yığından bir X tesadüfi değişkeni için birbirinden bağımsız $x_1, x_2, ..., x_n$ örneğinin gözlemlendiği ve hiç bir parametrik modelin olmadığını varsayılsın.

F dağılımının birikimli dağılım fonksiyonunu elde etmek için $\hat{F}$ deneysel dağılımı kullanılır. Fakat F'in sadece parametrik bir model var olduğunda kullanıldığı bilinmektedir. Parametrik modelin olmadığı durumda ise verilerin simülasyonu ve gerekli özelliklerin deneysel hesaplamaları yapılmalıdır.

Deneysel dağılım fonksiyonunda orijinal veri grubu olan $x_1, x_2, ..., x_n$ kümesindeki değerlerin her birinin ortaya çıkma olasılığı eşittir. Bu sebeple her bir x^*, orijinal örnekten şansa bağlı olarak örneklenmiş bağımsız değerler olacaktır. Bu nedenle simülasyon örneği olan $(x_1^*, x_2^*, ..., x_n^*)$, orijinal verilerden yerine koyma metoduyla seçilen tesadüfi bir örnek olacaktır. Burada kolaylık sağlayan, verilerin homojen olmasıdır. Bu yeniden örnekleme yöntemi, parametrik olmayan bootstrap yöntemi olarak adlandırılır.

4.5. Regresyon Analizinde Bootstrap Tekniği

Tüm bilim dallarında en sık kullanılan ve en fazla yararı sağlayan istatistik metotlarının başında doğrusal regresyon analizi gelmektedir. Bu analizde kullanılan veri seti $x_1, x_2, ..., x_n$ şeklinde n tane gözlem değeri içermektedir. Burada her bir x_i gözlemi için,

$$x_i = (c_i, y_i)$$

yazıldığında, y_i bağımlı değişkeni, c_i ise $c_i = c_{i1}, c_{i2}, ..., c_{ip}$ ile ifade edilen bağımsız değişkenlerin oluşturduğu 1xp vektörünü gösterir. Eğer bağımsız değişkenlerin değerleri biliniyor ise, bağımlı değişkenin beklenen değeri,

$$\mu_i = E(y_i / c_i) \qquad (i = 1,2,...,n)$$

olur. Doğrusal regresyon modelinin en önemli varsayımı μ_i' in c_i bağımsız değişkeninin doğrusal bir fonksiyonu olduğudur ve,

$$\mu_i = c_i \beta = \sum_{j=1}^{p} c_{ij} \beta_j \tag{4.9}$$

ile gösterilir. Analizin amacı, bilinmeyen $\beta = (\beta_1, \beta_2, ..., \beta_p)^T$ regresyon parametrelerini $x_1, x_2, ..., x_n$ gözlem verilerini kullanarak tahminde bulunmaktır. Doğrusal model,

$$y_i = c_i \beta + \varepsilon_i \qquad (i = 1,2,...,n) \tag{4.10}$$

şeklinde ifade edilir.

Eş. 4.11' de gösterilen ε_i hata terimlerinin beklenen değeri sıfır olmak üzere bilinmeyen F dağılımına sahip olduğunda,

$$F \to (\varepsilon_1, \varepsilon_2, ...\varepsilon_n) = \varepsilon \qquad [E_F(\varepsilon) = 0] \tag{4.11}$$

yazılır[9]. Buradan Eş. 4.10 ve Eş. 4.11'den hareketle,

$$E(y_i / c_i) = E(c_i \beta + \varepsilon_i / c_i) = E(c_i \beta / c_i) + E(\varepsilon_i / c_i)$$
$$= c_i \beta$$

bulunur. Bu elde edilen formül Eş. 4.10' da verilen doğrusallık varsayımıdır.

Eş. 4.10' da verilen modeldeki parametrelerin tahmin değerleri EKK yöntemi kullanılarak, C bağımsız değişken matrisi olmak üzere,

$$\hat{\beta} = \left(C^T C\right)^{-1} C^T y$$

ile hesaplanır.

Bootstrap yöntemi regresyon analizinde iki farklı şekilde kullanılabilir. Birincisi doğrusal regresyonun olasılık modelini $P \to x$ şeklinde yazmaktır[9]. Burada P,

$$P = \left(\beta, F\right)$$

olarak iki elemandan meydana gelmektedir. β regresyon katsayılarını, F ise Eş. 4.10' da verilen hata terimlerinin dağılımını göstermektedir. β değerleri bilinmediğinden, EKK yöntemi ile elde edilen $\hat{\beta}$ tahminleri kullanılarak hata terimleri,

$$\hat{\varepsilon}_i = y_i - c_i\hat{\beta} \qquad \left(i = 1,2,...,n\right) \tag{4.12}$$

ile hesaplanır ve hata terimlerinin deneysel dağılımı elde edilir. Hata terimleri bulunduktan sonra bootstrap örneklerini oluşturmak için yerine koyma yöntemiyle tesadüfi olarak seçim yapılır. Eş. 4.12' den, bağımlı değişkene ilişkin tahminler,

$$\hat{y}_i = c_i\hat{\beta} + \hat{\varepsilon}_i \qquad \left(i = 1,2,...,n\right)$$

ile elde edilir. Bağımlı değişken değerlerinin bulunması sırasında bağımsız değişkenlerin değişmediği kabul edilir[9]. Bootstrap örnekleri üzerinden tahmin edilecek $\hat{\beta}^*$ değerleri ise,

$$\hat{\beta}^* = \left(C^T C\right)^{-1} C^T y^* \tag{4.13}$$

ile bulunur.

Yöntemin regresyon analizinde ikinci tür kullanımında, C bağımsız değişken matrisini, y ise bağımlı değişkeni göstermek üzere (c_i, y_i) şeklinde tanımlanan x veri setine bootstrap yöntemi uygulanır. $x_i = \left(c_i, y_i\right)$ ise, bootstrap yönteminin uygulanması sonunda 1' den n' e kadar tesadüfi örneği ifade eden $i_1, i_2, ..., i_n$ için x^* bootstrap veri seti,

$$x^* = \left\{\left(c_{i_1}, y_{i_1}\right), \left(c_{i_2}, y_{i_2}\right), ..., \left(c_{i_n}, y_{i_n}\right)\right\}$$

ile gösterilir. Her bir bootstrap örneği için hesaplanacak olan regresyon katsayıları,

$$\hat{\beta}^* = \left(C^{*T} C^*\right)^{-1} C^{*T} y^*$$

olur.

İki yöntemden hangisinin daha iyi sonuç verdiği regresyon modelinin ne kadar doğru olduğu ile bağlantılıdır. Eş. 4.13' de oluşturulan regresyon modeli, hata terimlerinin bağımsız değişkenlere göre değişmediğini varsaymaktadır. Yani bağımsız değişkenler ne olursa olsun hata terimlerinin dağılımının değişmediği varsayılır. Bu varsayımın gerçekleşmesi ise zordur. Hangi yöntemin kullanılacağına karar vermek için bağımsız değişkenin sabit olup olmadığına bakmak yeterli olacaktır. Bağımsız değişkenler sabit olarak kabul edilmiş ise hata terimleri yöntemi, bağımsız değişkenler tesadüfi olarak seçiliyor ise ikinci olarak anlatılan yöntem olan x veri setini kullanmak daha iyi sonuç verir[9].

5.UYGULAMA

5.1. Giriş

Lojistik regresyon analizi, bağımsız değişkenlerin sürekli ve kategorik olarak bir arada bulunduğu, dağılımları üzerinde hiçbir kısıtlamanın bulunmadığı bir yöntemdir. Bu yüzden bağımlı değişkenin bireyde hastalığın var olup olmadığının araştırıldığı tıp alanındaki çalışmalarda sıkça kullanılmaktadır. Bu çalışmada hipertansiyon şikayeti ile İstanbul Özel Sante-Plus Hastanesi Kardiyoloji Bölümü' ne başvuran 148 kişiden elde edilen veriler üzerinde lojistik regresyon analizi yapılarak, hastaların hipertansiyon riski taşıyıp taşımadıklarına ilişkin açıklayıcı bir model oluşturulmaya çalışılmıştır. Modeli oluşturan bağımsız değişkenler, geriye doğru adımsal eleme yöntemi kullanılarak SPSS 17 paket programı ile belirlenmiştir. Daha sonra aynı veri seti üzerinde farklı tekrar sayıları için bootstrap tekniği uygulanarak parametre tahminleri yapılmış ve güven aralıkları oluşturulmuştur. Bootstrap tekniğinin uygulamasında S-PLUS 6.1 paket programı kullanılmıştır. En son adım olarak da lojistik regresyon sonucu elde edilen parametrelerin tahmin değerleri ile bootstrap sonucu elde edilen parametrelerin tahmin değerleri karşılaştırılmıştır.

Üzerinde çalışılan yığının tüm gözlemlerini parametre tahmini için kullanmak hem zaman kaybına yol açacak hem de maliyeti arttıracaktır. Yeniden örnekleme yöntemlerinden biri olan bootstrap yönteminin bilgisayar destekli olarak kullanımı bu sorunları ortadan kaldırıp büyük avantajlar sağlamaktadır. Hastaların hipertansiyon riski taşıyıp taşımama riski üzerine yapılan bu çalışmada, bootstrap yönteminden elde edilen sonuçların lojistik regresyon analizi sonucunda elde edilen parametre tahminlerinin standart hataları ile karşılaştırılarak daha güvenilir parametre tahminlere ulaşmak için etkili olup olmadığı incelenmiştir. Ayrıca bootstrap tekrar sayılarının büyümesinin sonuca olan etkisi karşılaştırmalı olarak ele alınmıştır.

5.2. Hipertansiyon Hakkında Genel Bilgiler

Hipertansiyon basit bir ifadeyle yüksek kan basıncı demektir. Kan basıncı, kanı kalpten dokulara taşıyan damarlarda oluşan basınçtır. Kan basıncı ölçülürken iki değere bakılır. Bunlar, büyük tansiyon(sistolik kan basıncı) ve küçük tansiyon (diyastolik kan basıncı) dur. Büyük tansiyon kalbin kasılması sırasında ölçülen kan basıncı, küçük tansiyon ise kalbin gevşemesi esnasında ölçülen kan basıncıdır. Büyük tansiyon için kan basıncının 140 mmHg, küçük tansiyon için 90 mmHg' in üzerinde bulunması hipertansiyon olarak tanımlanmaktadır. Hipertansiyon tanısı konulması için iki değerden bir tanesinin yüksek çıkması yeterlidir. Yorgunluk, bulantı, görme bozuklukları, fazla terleme, ciltte kızarma ve solukluk, burun kanaması, endişe ve sinirlilik, çarpıntı, baş dönmesi, baş ağrısı, kulaklarda çınlama ve uğultu gibi etmenler hipertansiyonun belirtileridir.

Hipertansiyon çok yaygın bir hastalık olmakla birlikte, hastaların azımsanmayacak bir kısmının kan basınç değerlerinin yüksek olduğunun farkında olmaması hastalığın önemini bir kat daha artırmaktadır. Sanayileşmiş ülkelerde yetişkin nüfusun %10-20 kadarında hipertansiyon görülmektedir. Hipertansiyona siyah ırkta ve kadınlarda daha sık rastlanmaktadır. Yaş, cinsiyet, ırk, fiziksel durum (istirahat, efor) gibi etmenler kan basıncını etkilemektedir. Kişinin yaşının hipertansiyona etkisi damarlarda yaşlanmaya bağlı olarak ortaya çıkan esneklik kaybıdır. Tuz kullanımı, aşırı beslenme, hareketsiz yaşam ve stresin tansiyon üzerinde olumsuz etkileri vardır. Hipertansiyon, kalp hastalıkları için önemli bir risk faktörüdür. Tedavi edilmediği takdirde öldürücü sonuçlar doğurabilmektedir. Kalbi zorlayarak kalp yetmezliğine ve damarları zorlayarak damar sertliğine yol açmaktadır. Hipertansiyonlu hastalarda beyin kanaması, felç, koroner arter hastalığı, ani ölüm, kalp krizi, ritim bozuklukları, böbrek yetmezliği ve retinopati (görme bozukluğuna yol açan göz bozukluğu) en sık gözlenen hastalıklardan birkaçıdır.

Nedenlerine göre iki tip hipertansiyon vardır. Bunlar esansiyel ve sekonder hipertansiyon olarak adlandırılmaktadır. Hastada çıkış nedeni bilinmeyen hipertansiyon tipine esansiyel hipertansiyon denir ve vakaların yaklaşık %90' ını

kapsamaktadır. Çıkış nedenleri bilinmese de çeşitli risk faktörlerinden söz edilebilir. Genetik yatkınlık, siyah ırk, menopoz ve stres bu risk faktörlerine örnek olarak sayılabilir. Bu grupta bulunan hastalar genellikle orta yaşlı, kilolu, sınırlı fiziksel aktiviteye sahip, fazla tuz tüketen, fazlaca alkol tüketen, sigara içen kimselerdir. Hastada çıkış nedeni bilinen hipertansiyon tipine ise sekonder hipertansiyon denir. Bu grup vakaların yaklaşık %10' unu kapsamaktadır. Böbrek kökenli olanlar en yaygın görülenleridir. Bu grupta hipertansiyon, böbrek ve böbreküstü hastalıkları, hormonel hastalıklar ile doğum kontrol hapı kullanımından kaynaklanabilmektedir.

Hipertansiyon tedavisi için, sigarayı bırakmak, kilo vermek(özellikle karın bölgesinden), düzenli egzersiz yapmak, alkolü azaltmak, tuz alımını kısıtlamak(günde 2gr) ve stresle başa çıkmayı öğrenmek gibi önlemler önerilmektedir. Bütün bu önlemler tansiyonu düşürmez ise doktor kontrolü altında ilaç kullanımına başlamak gerekmektedir[13].

5.3. Uygulamada Kullanılan Değişkenler

Bu çalışmada bağımlı değişken hastada hipertansiyon rahatsızlığının görülüp görülmediğini gösteren kategorik bir değişkendir. $Y_i = 1$ hastada hipertansiyonun varlığını, $Y_i = 0$ ise yokluğunu ifade etmektedir. Uygulamada "hiptansy" olarak gösterilmiştir.

Yaş: Sürekli bir değişken olup, hipertansiyon üzerinde etkili olan risk faktörlerinden biri olabilir. Çünkü yaşlanmayla beraber damarlar esnekliğini kaybetmektedir. Yıl cinsinden ölçüm yapılmıştır.

Cinsiyet: Kesikli bir değişken olup, 0: Kadın, 1: Erkek şeklinde kodlamıştır.

Boy: Sürekli bir değişken olup hastanın fiziksel özelliklerinin hipertansiyona etkisi olduğu düşünülmektedir. Santimetre cinsinden ölçüm yapılmıştır.

Kilo: Sürekli bir değişken olup hastanın fiziksel özelliklerinin hipertansiyona etkisi olduğu düşünülmektedir. Kilogram cinsinden ölçüm yapılmıştır.

Beden kitle indeksi (Bki): Sürekli bir değişken olup tıbbın üzerinde anlaştığı ve yaygın olarak kullandığı vücut ağırlığı değerlendirme ölçüsüdür. Uygulamada bki olarak gösterilecek olan beden kitle indeksi, vücut ağırlığının boyun karesine bölünmesiyle bulunur. Bulunan değer 18,5 in altındaysa birey zayıf, 18,5- 25 arasında normal, 25- 30 arasında kilolu, 30'un üstünde şişman (obez) sayılmaktadır. Beden kitle indeksinin vücudun dengesi anlamına gelmesi, hipertansiyona etki eden önemli bir risk faktörü olabileceğini göstermektedir.

Hiperlipidemi(hiplipid): Kesikli bir değişken olup, kanda yağ oranının normalden daha yüksek olma durumudur. Uygulamada hiplipid olarak gösterilmiştir. Hiperlipideminin varlığı 1, yokluğu ise 0 olarak kodlanmıştır.

Koroner arter(kroarte): Kesikli bir değişken olup, kalbin beslenmesini sağlayan atardamarlara verilen addır. Sağlıklı koroner arterler elastiktir; iç yüzeyleri düzdür ve içinden kan rahatça akar. Koroner arter hastalığında ise damar duvarı kalınlaşır, daha az elastiktir ve içerisinde plaklar oluşarak damarlar daralır. Kalbe yeterince kan ve oksijen gitmez. Uygulamada kroarte olarak gösterilmiş ve koroner arter hastalığın varlığı 1, yokluğu ise 0 olarak kodlanmıştır.

Diyabet süresi(diabsure): Sürekli bir değişken olup, hipertansiyon üzerinde etkili olan risk faktörlerinden biri olabilir. Yıl cinsinden ölçüm yapılmıştır. Uygulamada diabsure olarak gösterilmiştir.

İnsülin: Kesikli bir değişkendir. Pankreas tarafından üretilen bir hormon olup kan şekerini düşürücü etki yapar. İnsülin tedavisinin amacı vücutta eksik olan insülini yerine koyarak kan şekeri değerlerini normal değerlere getirebilmektir. Genellikle vücudumuz insüline ihtiyaç duymaya başladığında pankreasın insülin üreten dokusunun en az %80'i hasar görmüştür ve pankreasın insülin üreten dokusu (beta hücreleri) kendini yenileyemez. Bu nedenle vücudumuzda yeterince üretilemeyen bu

hormonu insülin enjeksiyonları ile dışarıdan sürekli yerine koymamız gerekir. Uygulamada insülin ilacı kullanan bireyler 1, kullanmayanlar ise 0 olarak kodlanmıştır.

Yoğun insülin(yoguins): Kesikli bir değişken olup uygulamada yoguins olarak gösterilmiştir. Yoğun insülin tedavisi günde enaz üç defa insülin enjeksiyonu ya da insülin pompası kullanımını, günde en az dört defa kan şekeri kontrolünü ve tüketilen besin maddelerine dikkat edilmesini içeren bir yöntemdir. Bu tedavide amaç günde bir veya iki kez insülin enjeksiyonu ile sağlanan kan şekeri kontrolünden daha iyi ve daha normale yakın bir kan şekeri kontrolü sağlamaktır. Bu nedenden dolayı yoğun insülin değişkeni uygulamada kullanıma göre ikiye ayrılmıştır. Hastanın insülini yüksek seviyede kullanımı 1, düşük seviyede kullanımı 0 olarak kodlanmıştır.

Metformin(metfrmn): Kesikli bir değişkendir. Metformin şeker hastalığının tedavisinde kullanılan bir ilaç olup, insülin duyarlılığını arttırmak için kullanılır. Karaciğerden glikoz çıkışını azaltır ve böylece açlık plazma glikozunun düşmesini sağlar. Metformin kullanmaya başlamadan önce karaciğer fonksiyonları kontrol edilmelidir. Uygulamada metfrmn olarak gösterilmiş ve kullanıldığı durum 1, kullanılmadığı durum 0 olarak kodlanmıştır.

Hba1c: Sürekli bir değişken olup kırmızı kan hücrelerinde glikozun bağlı olduğu hemoglobin yüzdesini gösteren bir ölçü birimidir. Hemoglobin eritrosit denilen kırmızı kan hücrelerinde oksijeni bağlar ve taşınmasını sağlar. Kısaca HbA1c son 2-3 ay içindeki ortalama kan glikozu düzeyini verir. Diyabetin mikrovasküler komplikasyonlarının gelişimi ve ilerlemesinin habercisi olarak da adlandırılabilir. HbA1c düzeyindeki %1'lik değişime kan glukoz düeyinde yaklaşık %30'luk bir değişiklik olduğunu yansıtır. %6,5'den küçük değerler kan şekeri düzeninin iyi seyrettiğini, % 7,0 üstü değerler kan şekerinin kötü seyrettiğini gösterir.

Üre: Sürekli bir değişkendir. Vücutta proteinlerin yakılması sonucu oluşan amonyak, karaciğerde karbondioksitle üreye dönüşür. Kana geçen üre, idrar yoluyla dışarıya atılır. Çünkü kandaki oranı diğer azotlu maddelere göre çok daha fazladır. Normal

miktarı % 30 mg olarak kabul edilmiş olup % 50 mg' ın üstü anormal olarak kabul edilir. Yaşlandıkça, böbreklerin üreyi vücuttan atma kabiliyeti de azalır. 40 yaşından itibaren, her yıl böbreklerin süzme kabiliyeti % 1 oranında düşmektedir. Bu sebepten dolayı 75-80 yaşındaki bir kişide kandaki üre miktarının % 65-75 mg bulunmasın normal olarak kabul edilebilir. Kandaki üre miktarının normal değerin üzerinde olması durumuna üremi adı verilir.

Kreatin: Sürekli bir değişkendir. Kas hücrelerinde yağları indirgeyerek enerji desteği sağlayan organik bir asittir. Böbrekte, karaciğerde ve pankreasta sentezlenir. Kan kreatin düzeyinin artışı böbreğin yetersiz çalıştığının bir göstergesidir. Vücuttan günlük kreatin atım miktarı yaklaşık 1-2 gr/gün kadardır.

Tk(total kollesterol): Sürekli bir değişkendir. Kolesterol tüm vücutta bulunan yaşam için gerekli bir çeşit yağdır. Yağ asitlerinin metabolizması ve vücut içinde taşınması sırasında kolesterol molekülleri rol alır. Hormonların üretiminde büyük önemi vardır. Bu yüzdendir ki, bu hayati yağ molekülü karaciğer tarafından daimi olarak üretilmektedir. Fakat hayvansal gıdaların fazla alınması kolesterol seviyesini yükseltir. Kan kolesterol düzeyinin yüksek olması kalp damar hastalığı riskini arttırır. Kanda 200mg/dl değerinden küçük olması normal, bu değerin aşılması ise yüksek kolesterol olarak adlandırılır.

Tg(trigliserit): Sürekli bir değişkendir. Vücutta trigliserit seviyesi yüksek ise ateroskleroz(damar sertliği) ve buna bağlı koroner kalp hastalıkları görülebilir. Trigliserit değerinin 150mg/dl altında olması gerekmektedir. Bunun üzerindeki değerler için ilaç ya da diyet tedavisi gerekebilir.

Hdl(yüksek yoğunluklu lipoprotein): Sürekli bir değişkendir. Vücuttaki dokulardan karaciğere kolesterol taşıyan bir lipoproteindir. Hdl arterlerde oluşan kolesterolu alıp vücuttan atılmak üzere karaciğere taşıdığı için bu lipoproteinde bulunan kolesterol "iyi kolesterol" olarak anılır. Epidemiyolojik çalışmalarda 60 mg/dl üstünde Hdl seviyesinin kardiyovasküler hastalıklara(koroner arter hastalığı gibi) karşı koruyucu

bir etkisi olduğu görülmüştür. Düşük Hdl düzeylerinin ise aterosklerotik hastalıklar için pozitif risk faktörüdür.

Ldl(düşük yoğunluklu lipoprotein): Sürekli bir değişkendir. Ldl seviyesi ile kalp hastalıkları arasındaki bağlantıdan dolayı "kötü" kolesterol olarak anılır. Ldl' nin başlıca işlevi, kolesterol ve trigliserit üreten hücre ve dokulardan bu molekülleri alıp bunları gereksinimi olan hücre ve dokulara taşımaktır. Vücuttaki toplam kolesterolün %70'i Ldl 'de bulunmaktadır. Kanda Ldl seviyesinin 130 mg/dl' nin altında olması istenilen düzeydir.

Vldl(çok düşük yoğunluklu lipoprotein): Sürekli bir değişkendir. Vldl, karaciğerde oluştuktan sonra taşıdıkları trigliseriti vücuttaki çeşitli dokulara aktarırlar ve bu sürecin sonunda Ldl 'ye dönüşürler. Karaciğer, kolesterol ve trigliseritlerin sentezlendiği başlıca organdır. Bu organın ihtiyacını aşan kolesterol ve trigliseritler Vldl tanecikleri olarak kana salınırlar. Yüksek düzeyde Vldl, aterosklerozun hızlanmasına yol açabilir.

5.4. Geriye Doğru Adımsal Eleme Yöntemi Uygulaması

Lojistik regresyon analizinde çok değişkenli model kurulurken, modele dahil edilecek değişkenleri seçmek için kullanılan α anlamlılık düzeyinin seçimi regresyon analizine göre farklıdır. Değişken seçiminde p değerinin 0,25 olarak seçilmesi, Bendel ve Afifi (1977)'nin doğrusal regresyon ve Mickey ve Greenland (1989)'ın lojistik regresyon üzerine yapmış oldukları çalışmalarda ortaya konmuştur. α anlamlılık düzeyinin geleneksel olarak kullanılan 0,05 olarak seçimi, istatistiksel olarak önemli olan değişkenlerin modele dahil edilememesine yol açmaktadır. Bu çalışmada anlamlılık düzeyi 0,25 olarak belirlenmiş ve en iyi modeli kurabilmek için olabilirlik oran test istatistiğine dayalı olan geriye doğru adımsal eleme yöntemi kullanılmıştır. 14. adımda ulaşılan modelde yer alacak bağımsız değişken kümesi Çizelge 5.1' de verilmiştir. Geriye doğru adımsal eleme yönteminin 1. adımında modele bütün değişkenler dahil edilmiştir. İnsülin değişkeninin p değeri daha önceden belirlenmiş olan kritere göre en büyük olarak saptanmış olup modele

katkısının en az olduğu belirlenerek modelden çıkartılmış ve 2. adıma geçilmiştir. 2. adımda tekrar p değerlerinin karşılaştırılması sonucu diabsure değişkeninin diğer değişkenlere göre önemlik seviyesi az bulunmuş ve elemine edilerek diğer adıma geçilmiştir. 14. adım sonucunda modelden atılacak değişken kalmamış olup yaş, cinsiyet, boy, bki, kroarte ve hba1c değişkenleri modele katkısı en yüksek olan değişkenler olarak belirlenmiştir. Daha sonra bu değişkenlerle,

Hipertansiyon = -12,177 + 0,105 yas+ 1,382 cinsiyet + 0,035 boy + 0,092 bki
-1,501 koroner arter - 0,113 hba1c

lojistik regresyon modeli kurulmuştur.

Çizelge 5.1. Lojistik regresyon modelinde geriye doğru eliminasyon yöntemi 14.Adım

Denkleme Alınan Değişkenler							
		B	S.E.	Wald	df	Sig.	Exp(B)
Adım 14	Yaş	,105	,025	17,442	1	,000	1,111
	Cinsiyet(1)	1,382	,572	5,843	1	,016	3,981
	Boy	,035	,028	1,580	1	,209	1,036
	Bki	,092	,039	5,492	1	,019	1,096
	Kroarte(1)	-1,501	1,142	1,727	1	,189	,223
	Hba1c	-,113	,094	1,428	1	,232	,894
	Constant	-12,177	5,524	4,859	1	,028	,000

Kurulan modelin uygunluğu Hosmer-Lemeshow uyum iyiliği testi ile incelenmiştir. Hosmer ve Lemeshov test istatistiği ile sabit terim hariç tüm katsayıların model üzerinde belirleyici olup olmadığının test edilmesi amaçlanmıştır. İlgili hipotezler,

H_0 : Parametreler model açısından belirleyicidir.

H_1 : Parametreler model açısından belirleyici değildir.

şeklinde ifade edilerek, hesaplanan ki kare istatistiğinin belirlenen anlamlılık düzeyinde(0,25) ki-kare tablo değerinden küçük olduğu görülmüştür. Bu nedenle sıfır hipotezi kabul edilmiş ve modele dahil edilen değişkenlerin model için uyumlu olduğu ortaya çıkmıştır. (Bkz. Çizelge 5.2)

Çizelge 5.2. Hosmer ve Lemeshov test sonuçları

Hosmer Lemeshow Testi

Adım	Ki-kare	df	p
1	8,144	8	,420
2	8,140	8	,420
3	5,948	8	,653
4	5,272	8	,728
5	4,280	8	,831
6	4,342	8	,825
7	11,761	8	,162
8	6,970	8	,540
9	5,531	8	,700
10	10,346	8	,242
11	13,380	8	,099
12	4,755	8	,783
13	4,206	8	,838
14	15,104	8	,057

Geriye dönük adımsal eleme yöntemine göre tüm adımların doğru sınıflama oranının gösterildiği Çizelge 5.3 de seçilen en iyi model olan son adımdaki modelin doğru sınıflama oranı % 76,4 olarak hesaplanmıştır.

Çizelge 5.3. Adımsal sınıflama tablosu

Sınıflandırma Tablosu[a]

Gözlem			Tahmin Edilen		
			hipertansiyon		
			YOK	VAR	Doğru Sınıflandırma %
Adım 1	hipertansiyon	YOK	31	22	58,5
		VAR	13	82	86,3
	Toplam Yüzde				76,4
Adım 2	hipertansiyon	YOK	31	22	58,5
		VAR	13	82	86,3
	Toplam Yüzde				76,4
Adım 3	hipertansiyon	YOK	31	22	58,5
		VAR	13	82	86,3
	Toplam Yüzde				76,4
Adım 4	hipertansiyon	YOK	31	22	58,5
		VAR	13	82	86,3
	Toplam Yüzde				76,4
Adım 5	hipertansiyon	YOK	30	23	56,6
		VAR	13	82	86,3
	Toplam Yüzde				75,7
Adım 6	hipertansiyon	YOK	31	22	58,5
		VAR	13	82	86,3
	Toplam Yüzde				76,4
Adım 7	hipertansiyon	YOK	32	21	60,4
		VAR	13	82	86,3
	Toplam Yüzde				77,0
Adım 8	hipertansiyon	YOK	32	21	60,4
		VAR	13	82	86,3
	Toplam Yüzde				77,0
Adım 9	hipertansiyon	YOK	32	21	60,4
		VAR	14	81	85,3
	Toplam Yüzde				76,4
Adım10	hipertansiyon	YOK	31	22	58,5
		VAR	14	81	85,3
	Toplam Yüzde				75,7
Adım 11	hipertansiyon	YOK	31	22	58,5
		VAR	14	81	85,3
	Toplam Yüzde				75,7
Adım 12	hipertansiyon	YOK	31	22	58,5
		VAR	14	81	85,3
	Toplam Yüzde				75,7
Adım 13	hipertansiyon	YOK	30	23	56,6
		VAR	14	81	85,3
	Toplam Yüzde				75,0
Adım14	hipertansiyon	YOK	31	22	58,5
		VAR	13	82	86,3
	Toplam Yüzde				76,4

a. The cut value is ,500

5.5. Kategorik Değişken Analizi

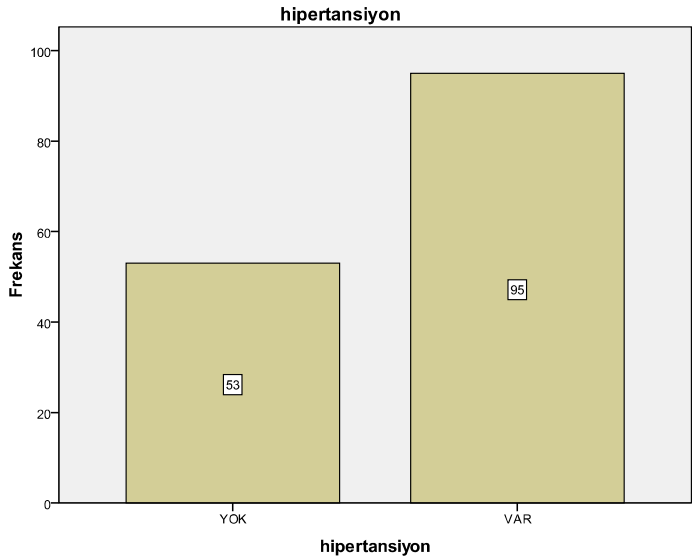
Araştırmaya dahil edilen kategorik değişkenlerin frekans dağılımları ve yüzdesel gösterimleri aşağıda yer alan çizelge ve şekillerde verilmiştir.

Çizelge 5.4. Hipertansiyon değişkeninin frekans tablosu

hipertansiyon

	Frekans	Yüzde	Geçerli Yüzde	Birikimli Yüzde
YOK	53	35,8	35,8	35,8
VAR	95	64,2	64,2	100,0
Toplam	148	100,0	100,0	

Çizelge 5.4' de görüldüğü gibi araştırma grubunda yer alan 148 hastanın % 64,2 ine hipertansiyon hastası olduğu teşhisi, % 35,8 sine hipertansiyon hastası olmadığı teşhisi konulmuş olup frekans dağılımı Şekil 5.1 de gösterilmektedir.



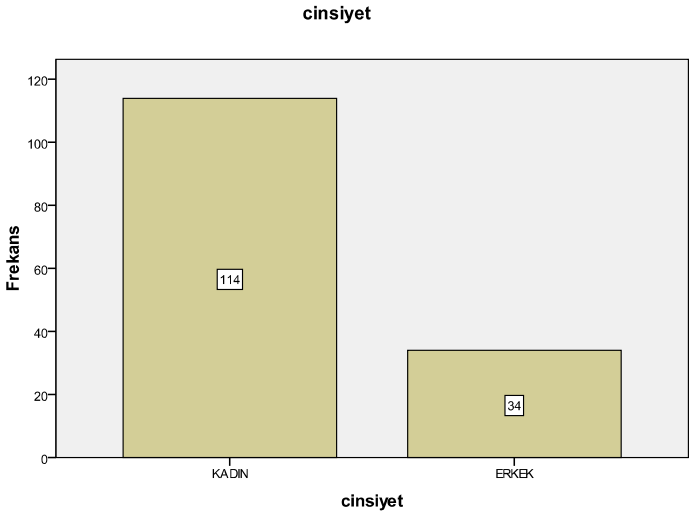
Şekil 5.1. Hipertansiyon değişkeninin frekans dağılım grafiği

Çizelge 5.5. Cinsiyet değişkeninin frekans tablosu

cinsiyet

	Frekans	Yüzde	Geçerli Yüzde	Birikimli Yüzde
KADIN	114	77,0	77,0	77,0
ERKEK	34	23,0	23,0	100,0
Toplam	148	100,0	100,0	

Çizelge 5.5' de görüldüğü gibi araştırma grubunda yer alan 148 hastanın % 77 si kadın, % 23 ü ise erkek olup frekans dağılımı Şekil 5.2 de gösterilmektedir.



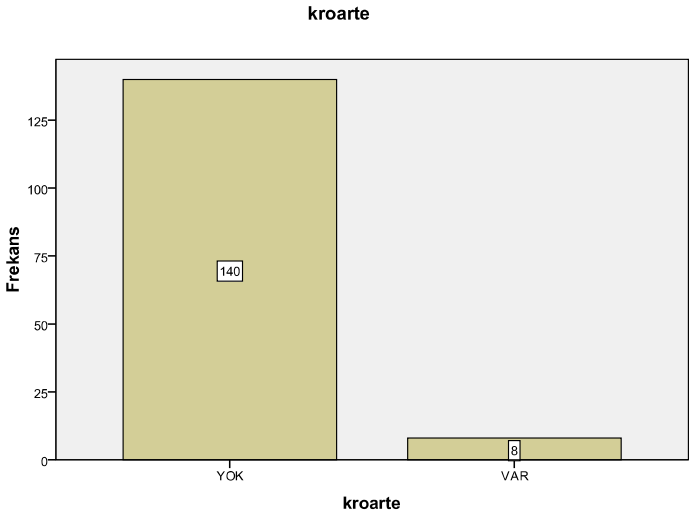
Şekil 5.2. Cinsiyet değişkeninin frekans dağılım grafiği

Çizelge 5.6. Kroarte değişkeninin frekans tablosu

kroarte

	Frekans	Yüzde	Geçerli Yüzde	Birikimli Yüzde
YOK	140	94,6	94,6	94,6
VAR	8	5,4	5,4	100,0
Toplam	148	100,0	100,0	

Çizelge 5.6' da, araştırma grubunda yer alan 148 hastanın % 94,6 sında kroner arter hastalığının olduğu , % 5,4 ünde ise olmadığı teşhisi konulmuş olup frekans dağılımı Şekil 5.3' de gösterilmektedir.



Şekil 5.3. Kroarte değişkeninin frekans dağılım grafiği

5.6. Sürekli Değişken Analizi

Araştırmaya dahil edilen sürekli değişkenlerin tümünün ortalaması, maksimum değerleri, minimum değerleri, ortalamaları, standart sapmaları, basıklık ve çarpıklık değerleri Çizelge 5.7' de verilmiştir.

Çizelge 5.7 Sürekli değişkenlerin tanımlayıcı istatistikleri

Tanımlayıcı İstatistikler

	N	Minimum	Maksimum	Ortalama		Std. Sapma	Çarpıklık		Basıklık	
	İstatistik	İstatistik	İstatistik	İstatistik	Std. Hata	İstatistik	İstatistik	Std. Hata	İstatistik	Std. Hata
yaş	148	30,00	78,00	53,2568	,82284	10,01028	,286	,199	-,341	,396
boy	148	51,00	190,00	159,5338	1,00940	12,27985	-4,446	,199	41,062	,396
kilo	148	7,00	145,00	86,3176	1,44011	17,51968	,186	,199	3,154	,396
bki	148	25,14	52,75	33,7728	,45651	5,55363	,757	,199	,416	,396
diabsure	148	,50	25,00	5,6385	,37500	4,56205	1,316	,199	1,924	,396
hba1c	148	5,00	14,10	8,3884	,17849	2,17145	,636	,199	-,553	,396
ure	148	10,00	138,00	30,3588	1,13455	13,80234	3,913	,199	25,514	,396
kreatin	148	,00	1,80	,8526	,01859	,22615	,956	,199	4,422	,396
tk	148	2,70	327,00	199,8547	3,92548	47,75554	-,401	,199	2,971	,396
tg	148	37,00	951,00	192,2365	10,73814	130,63510	2,365	,199	8,451	,396
hdl	148	8,00	118,00	48,2568	1,13024	13,75001	1,302	,199	4,781	,396
ldl	148	51,00	234,00	119,1014	3,08742	37,56006	,630	,199	-,037	,396
vldl	148	1,00	160,00	37,9865	2,04347	24,85988	1,967	,199	5,264	,396

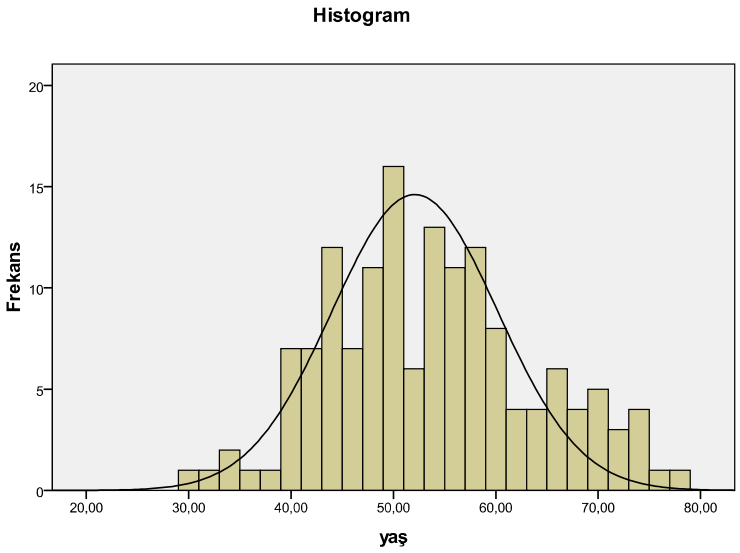
Çizelge 5.8. Yaş değişkeninin tanımlayıcı istatistikleri

İstatistik

yaş

N	Geçerli	148
	Kayıp	0
Ortalamanın Std. Hatası		,82284
Std. Sapma		10,01028
Varyans		100,206
Çarpıklık		,286
Çarpıklığın Std. Hatası		,199
Basıklık		-,341
Basıklığın Std. Hatası		,396
Minimum		30,00
Maksimum		78,00

Uygulamaya dahil edilen 148 kişinin yaş ortalaması 53 olup standart sapması 10,01'dir. Çarpıklık, basıklık istatistikleri, en küçük ve en büyük değerleri Çizelge 5.8' de verilmiştir. Değişkenin dağılımına ilişkin histogram ise Şekil 5.4' de gösterilmektedir.



Şekil 5.4. Yaş değişkeninin histogram grafiği

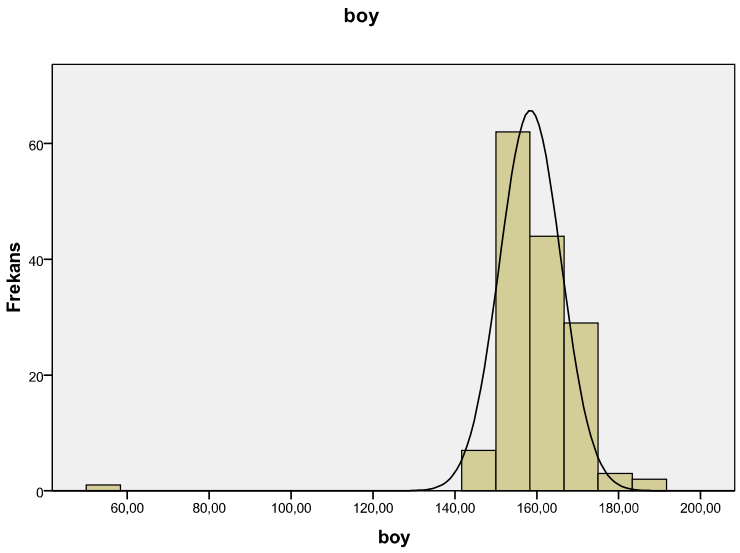
Çizelge 5.9. Boy değişkeninin tanımlayıcı istatistikleri

İstatistik

boy

N	Geçerli	148
	Kayıp	0
Ortalamanın Std. Hatası		1,00940
Std. Sapma		12,27985
Varyans		150,795
Çarpıklık		-4,446
Çarpıklığın Std. Hatası		,199
Basıklık		41,062
Basıklığın Std. Hatası		,396
Minimum		51,00
Maksimum		190,00

Uygulamaya dahil edilen 148 kişinin boy ortalaması 1,59 cm olup standart sapması 12,27'dir. Çarpıklık, basıklık istatistikleri, en küçük ve en büyük değerleri Çizelge 5.9' da verilmiştir. Değişkenin dağılımına ilişkin histogram ise Şekil 5.5' de gösterilmektedir.



Şekil 5.5. Boy değişkeninin histogram grafiği

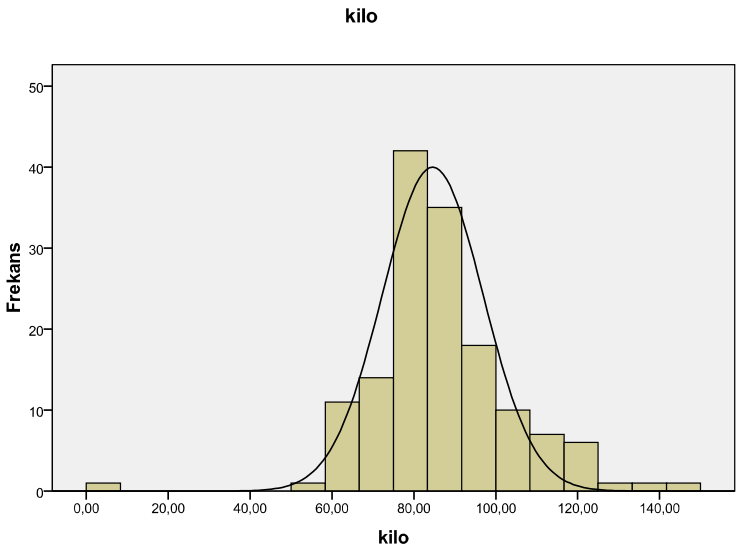
Çizelge 5.10. Kilo değişkeninin tanımlayıcı istatistikleri

İstatistik

kilo

N	Geçerli	148
	Kayıp	0
Ortalamanın Std. Hatası		1,44011
Std. Sapma		17,51968
Varyans		306,939
Çarpıklık		,186
Çarpıklığın Std. Hatası		,199
Basıklık		3,154
Basıklığın Std. Hatası		,396
Minimum		7,00
Maksimum		145,00

Uygulamaya dahil edilen 148 kişinin kilolarının ortalaması 86,31kg olup standart sapması 17,51'dir. Çarpıklık, basıklık istatistikleri, en küçük ve en büyük değerleri Çizelge 5.10' da verilmiştir. Değişkenin dağılımına ilişkin histogram ise Şekil 5.6' da gösterilmektedir.



Şekil 5.6. Kilo değişkeninin histogram grafiği

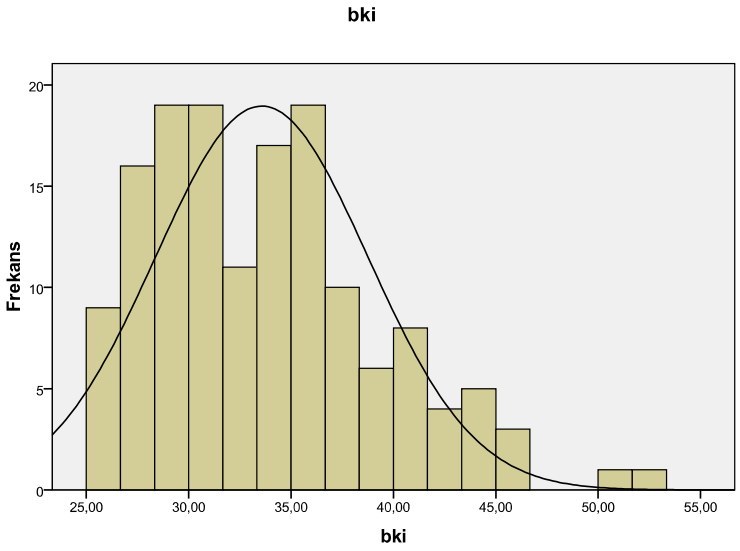
Çizelge 5.11. Bki değişkeninin tanımlayıcı istatistikleri

İstatistik

bki

N	Geçerli	148
	Kayıp	0
Ortalamanın Std.Hatası		,45651
Std. Sapma		5,55363
Varyans		30,843
Çarpıklık		,757
Çarpıklığın Std. Hatası		,199
Basıklık		,416
Basıklığın Std. Hatası		,396
Minimum		25,14
Maksimum		52,75

Uygulamaya dahil edilen 148 kişinin bki(beden kitle indeksi) ortalaması 33,77 olup standart sapması 5,55'dir. Çarpıklık, basıklık istatistikleri, en küçük ve en büyük değerleri Çizelge 5.11' de verilmiştir. Değişkenin dağılımına ilişkin histogram ise Şekil 5.7' de gösterilmektedir.



Şekil 5.7. Bki değişkeninin histogram grafiği

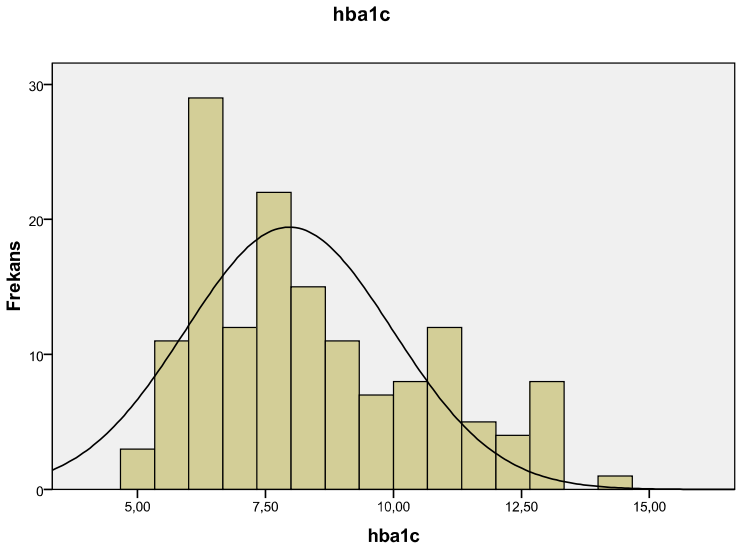
Çizelge 5.12. Hba1c değişkeninin tanımlayıcı istatistikleri

İstatistik

hba1c

N	Geçerli	148
	Kayıp	0
Ortalamanın Std. Hatası		,17849
Std. Sapma		2,17145
Varyans		4,715
Çarpıklık		,636
Çarpıklığın Std. Hatası		,199
Kurtosis		-,553
Basıklığın Std. Hatası		,396
Minimum		5,00
Maksimum		14,10

Uygulamaya dahil edilen 148 kişinin hba1c(Hemoglobin a1c) testinin ortalaması 8,38 olup standart sapması 2,17'dir. Çarpıklık, basıklık istatistikleri, en küçük ve en büyük değerleri Çizelge 5.12' de verilmiştir. Değişkenin dağılımına ilişkin histogram ise Şekil 5.8' de gösterilmektedir.



Şekil 5.8. Hba1c değişkeninin histogram grafiği

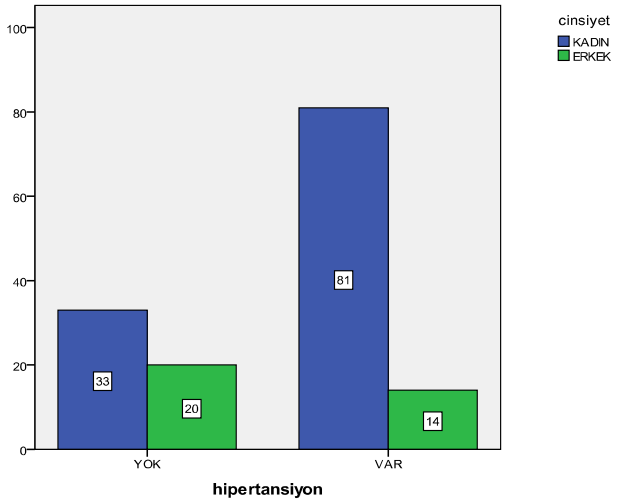
5.7. Kategorik Değişkenlerin Çapraz Tablo Analizi

Çizelge 5.13. Hipertansiyon ve cinsiyet değişkenlerinin çapraz tablosu

hipertansiyon * cinsiyet Çapraz Tablosu

			cinsiyet		
			KADIN	ERKEK	Toplam
hipertansiyon	YOK	Sayısı	33	20	53
		Toplam Yüzde	22,3%	13,5%	35,8%
	VAR	Sayısı	81	14	95
		Toplam Yüzde	54,7%	9,5%	64,2%
Toplam		Sayısı	114	34	148
		Toplam Yüzde	77,0%	23,0%	100,0%

Çizelge 5.13' de görüldüğü gibi uygulamaya dahil edilen 148 kişiden %22,3 ünü hipertansiyon teşhisi konulmamış, %54,7 sini hipertansiyon teşhisi konulmuş kadınlar oluşturmaktadır. %14 ünü hipertansiyon teşhisi konulmamış, % 10 unu ise hipertansiyon teşhisi konulmuş erkekler oluşturmaktadır. Şekil 5.9' da kadın ve erkeklerin hipertansiyon teşhisinin konulup konulmamasına göre oluşturulmuş grafik görülmektedir.



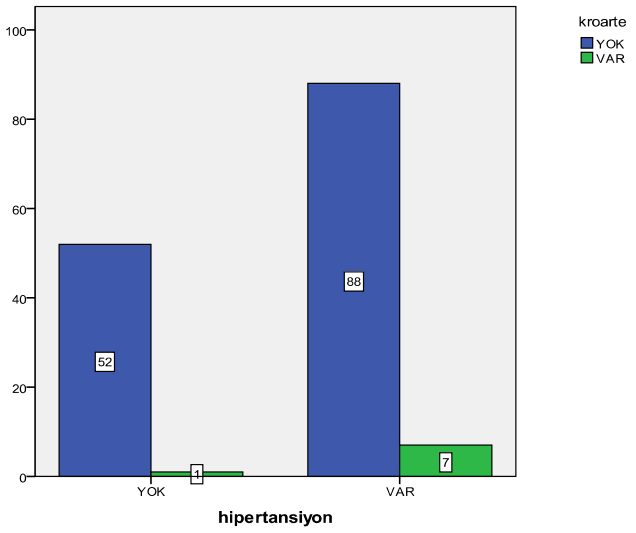
Şekil 5.9. Hipertansiyon ve cinsiyet değişkenlerinin dağılım grafiği

Çizelge 5.14. Hipertansiyon ve kroarte değişkenlerinin çapraz tablosu

hipertansiyon * kroarte Çapraz Tablosu

			kroarte		
			YOK	VAR	Toplam
hipertansiyon	YOK	Sayısı	52	1	53
		Toplam Yüzde	35,1%	,7%	35,8%
	VAR	Sayısı	88	7	95
		Toplam Yüzde	59,5%	4,7%	64,2%
Toplam		Sayısı	140	8	148
		Toplam Yüzde	94,6%	5,4%	100,0%

Çizelge 5.14' de görüldüğü gibi uygulamaya dahil edilen 148 kişiden %35,1 ini hipertansiyon teşhisi konulmamış, %59,5 ini hipertansiyon teşhisi konulmuş kroner arter hastalığını taşımayan kişiler oluşturmaktadır. %0,7 sini hipertansiyon teşhisi konulmamış, %4,7 sini ise hipertansiyon teşhisi konulmuş kroner arter hastalığını taşıyan kişiler oluşturmaktadır. Şekil 5.10' da kroner arter hastası olan ya da kroner arter hastası olmayan kişilere hipertansiyon teşhisinin konulup konulmamasına göre oluşturulmuş grafik görülmektedir.



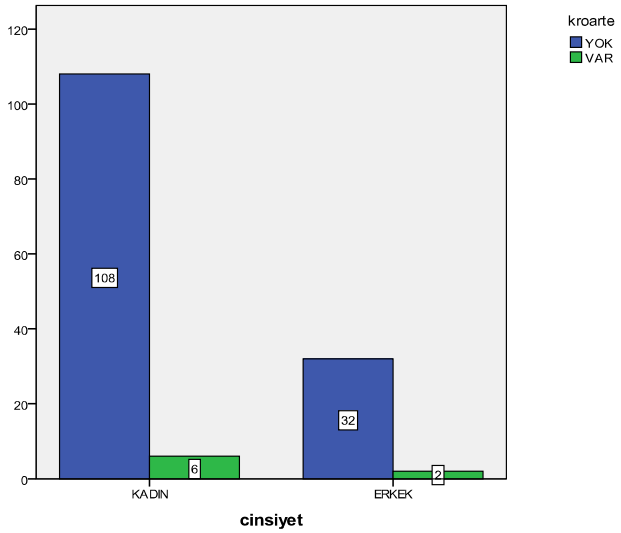
Şekil 5.10. Hipertansiyon ve kroarte değişkenlerinin dağılım grafiği

Çizelge 5.15. Cinsiyet ve kroarte değişkenlerinin çapraz tablosu

cinsiyet * kroarte Çapraz Tablosu

			kroarte		
			YOK	VAR	Toplam
cinsiyet	KADIN	Sayısı	108	6	114
		Toplam Yüzde	73,0%	4,1%	77,0%
	ERKEK	Sayısı	32	2	34
		Toplam Yüzde	21,6%	1,4%	23,0%
Toplam		Sayısı	140	8	148
		Toplam Yüzde	94,6%	5,4%	100,0%

Çizelge 5.15' de görüldüğü gibi uygulamaya dahil edilen 148 kişiden %73 ünü kroner arter hastalığı bulunmayan, %4,1 ini ise bulunan kadın bireyler oluşturmaktadır. %21,6 sını kroner arter hastalığı bulunmayan, %1,4 ünü ise kroner arter hastalığı bulunan erkek bireyler oluşturmaktadır. Şekil 5.11' de kroner arter hastalığı bulunan ve bulunmayan kişilerin cinsiyetlerine göre oluşturulmuş grafik görülmektedir.



Şekil 5.11. Cinsiyet ve kroarte değişkenlerinin dağılım grafiği

5.8. Bootstrap Yöntemi Uygulama Sonuçları

Hipertansiyon şikayeti ile İstanbul Özel Sante-Plus Hastanesi Kardiyoloji Bölümü' ne başvuran 148 kişiden elde edilen veriler üzerinde bootstrap yeniden örnekleme tekniği uygulanarak aynı hacimde(n=148) yeni veri setleri oluşturulmuştur. Oluşturulan veri setlerinde bootstrap tekrar sayıları B=50, B=100, B=250, B= 500 ve B= 1000 olarak alınmış olup her biri için parametre tahminleri yapılarak lojistik regresyon modelleri kurulmuştur. Kurulan modellerdeki parametre tahminlerinin standart hata değerleri ve güven aralıkları hesaplanarak geriye doğru adımsal eleme yöntemi sonucunda elde edilen modeldeki parametre tahminleri ile karşılaştırılmıştır. Farklı bootstrap tekrar sayılarının kullanılmasının sebebi, elde edilen parametre tahmin değerlerinin tekrar sayısı arttıkça standart hatalarının nasıl değişim gösterdiğini göstermek ve veri setinin yığını temsil gücünü incelemektir.

Çizelge 5.16. Bootstrap uygulaması sonucunda elde edilen katsayıların lojistik regresyon sonuçlarıyla karşılaştırılması

DEĞİŞKEN	KATSAYI	B=50		B=100		B=250		B=500		B=1000	
		DEĞER	SAPMA	DEĞER	SAPMA	DEĞER	SAPMA	DEĞER	SAPMA	DEĞER	SAPMA
sabit	-12,177	-12,57351	-0,39651	-12,79776	-0,62076	-12,80608	-0,62908	-12,80683	-0,62983	-12,99635	-0,81935
yaş	,105	0,10697	0,00197	0,10790	0,0029	0,10913	0,00413	0,11022	0,00522	0,10930	0,0043
cinsiyet	1,382	1,41728	0,03528	1,42019	0,03819	1,41590	0,0339	1,42094	0,03894	1,40521	0,02321
boy	,035	0,03501	0,00001	0,03623	0,00123	0,03625	0,00125	0,03570	0,0007	0,03703	0,00203
bki	,092	0,09720	0,0052	0,09864	0,00664	0,09675	0,00475	0,09752	0,00552	0,09711	0,00511
kroarte	-1,501	-2,22327	-0,72227	-1,93182	-0,43082	-2,35313	-0,85213	-2,37055	-0,86955	-2,31895	-0,81795
hba1c	-,113	-0,11521	-0,00221	-0,12130	-0,0083	-0,11944	-0,00644	-0,11942	-0,00642	-0,11494	-0,00194

Çizelge 5.17. Bootstrap uygulaması sonucunda elde edilen standart hataların lojistik regresyon sonuçlarıyla karşılaştırılması

DEĞİŞKEN	KATSAYI STD. HATA	B=50	B=100	B=250	B=500	B=1000
sabit	5,524	3,67031	3,54811	3,34525	3,23127	3,26707
yaş	,0257	0,02196	0,01760	0,02038	0,01879	0,01877
cinsiyet	,5724	0,47547	0,37171	0,38719	0,39767	0,41107
boy	,0281	0,01843	0,01806	0,01859	0,01688	0,01827
bki	,03991	0,03979	0,02964	0,03426	0,03427	0,03328
kroarte	1,1422	2,24767	1,86583	2,32397	2,35487	2,32982
hba1c	,0943	0,08562	0,05923	0,07608	0,07918	0,07587

Çizelge 5.18. % 95'lik bootstrap güven aralıkları (BCa: bootstrap güven aralıkları)

DEĞİŞKEN	KATSAYI	B=50		B=100		B=250		B=500		B=1000	
		% 5 BCa	%95 BCa	% 5 BCa	%95 BCa	% 5 BCa	%95 BCa	% 5 BCa	%95 BCa	% 5 BCa	%95 BCa
sabit	-12,177	-20,36963	-7,39339	-20,28679	-7,86509	-18,54469	-7,60996	-17,69657	-7,11380	-17,24029	-7,14737
yaş	0,105	0,06744	0,14358	0,07990	0,13361	0,06650	0,13422	0,07232	0,12993	0,07678	0,13591
cinsiyet	1,382	0,71471	2,25439	0,78393	1,95887	0,73633	1,92856	0,59616	1,96967	0,68969	2,01424
boy	0,035	0,01401	0,06955	0,00848	0,06571	0,01433	0,07221	0,01030	0,06592	0,00869	0,06722
bki	0,092	0,03466	0,16121	0,03002	0,12877	0,03030	0,14356	0,02878	0,13704	0,04077	0,14544
kroarte	-1,501	-7,20331	0,21547	-7,07260	0,29827	-7,61318	0,24109	-7,60478	0,01587	-7,7132	0,04648
hba1c	-0,113	-0,26049	0,02287	-0,20401	-0,01035	-0,22595	0,03397	-0,23386	0,02758	-0,22892	0,01578

Çizelge 5.16' da verilen katsayılarından;

Ana lojistik regresyon modeli:

Hipertansiyon = -12,177 + 0,105 yaş + 1,382 cinsiyet + 0,035 boy + 0,092 bki
-1,501 kroner arter -0,113 hba1c

B = 50 alınarak yapılan bootstrap örneklemesi sonucu elde edilen lojistik regresyon modeli:

Hipertansiyon = -12,57351 + 0,10697 yas + 1,41728 cinsiyet + 0,03501 boy
+0,09720 bki - 2,22327 kroner arter - 0,11521 hba1c

B = 100 alınarak yapılan bootstrap örneklemesi sonucu elde edilen lojistik regresyon modeli:

Hipertansiyon = -12,79776 + 0,10790 yas + 1,42019 cinsiyet + 0,03623 boy
+ 0,09864 bki - 1,93182 kroner arter - 0,12130 hba1c

B = 250 alınarak yapılan bootstrap örneklemesi sonucu elde edilen lojistik regresyon modeli:

Hipertansiyon = -12,80608 + 0,10913 yas + 1,41590 cinsiyet + 0,03625 boy
+ 0,09675 bki - 2,35313 kroner arter - 0,11944 hba1c

B = 500 alınarak yapılan bootstrap örneklemesi sonucu elde edilen lojistik regresyon modeli:

Hipertansiyon = -12,80683 + 0,11022 yas + 1,42094 cinsiyet + 0,03570 boy
+ 0,09752 bki - 2,37055 kroner arter - 0,11942 hba1c

B = 1000 alınarak yapılan bootstrap örneklemesi sonucu elde edilen lojistik regresyon modeli:

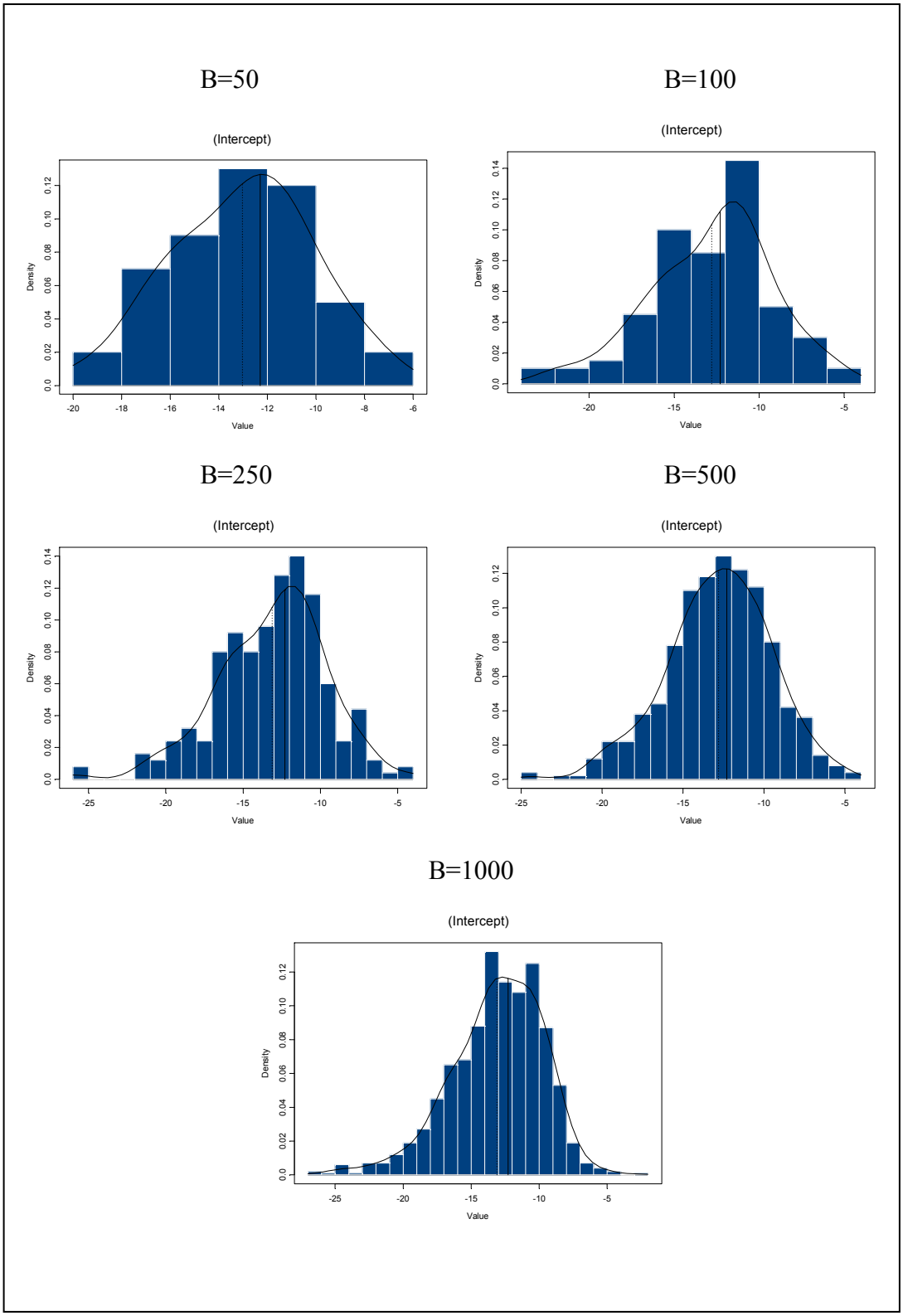
Hipertansiyon = -12,99635 + 0,10930 yas + 1,40521 cinsiyet + 0,03703 boy
+ 0,09711 bki - 2,31895 kroner arter - 0,11494 hba1c

şeklinde elde edilmiştir.

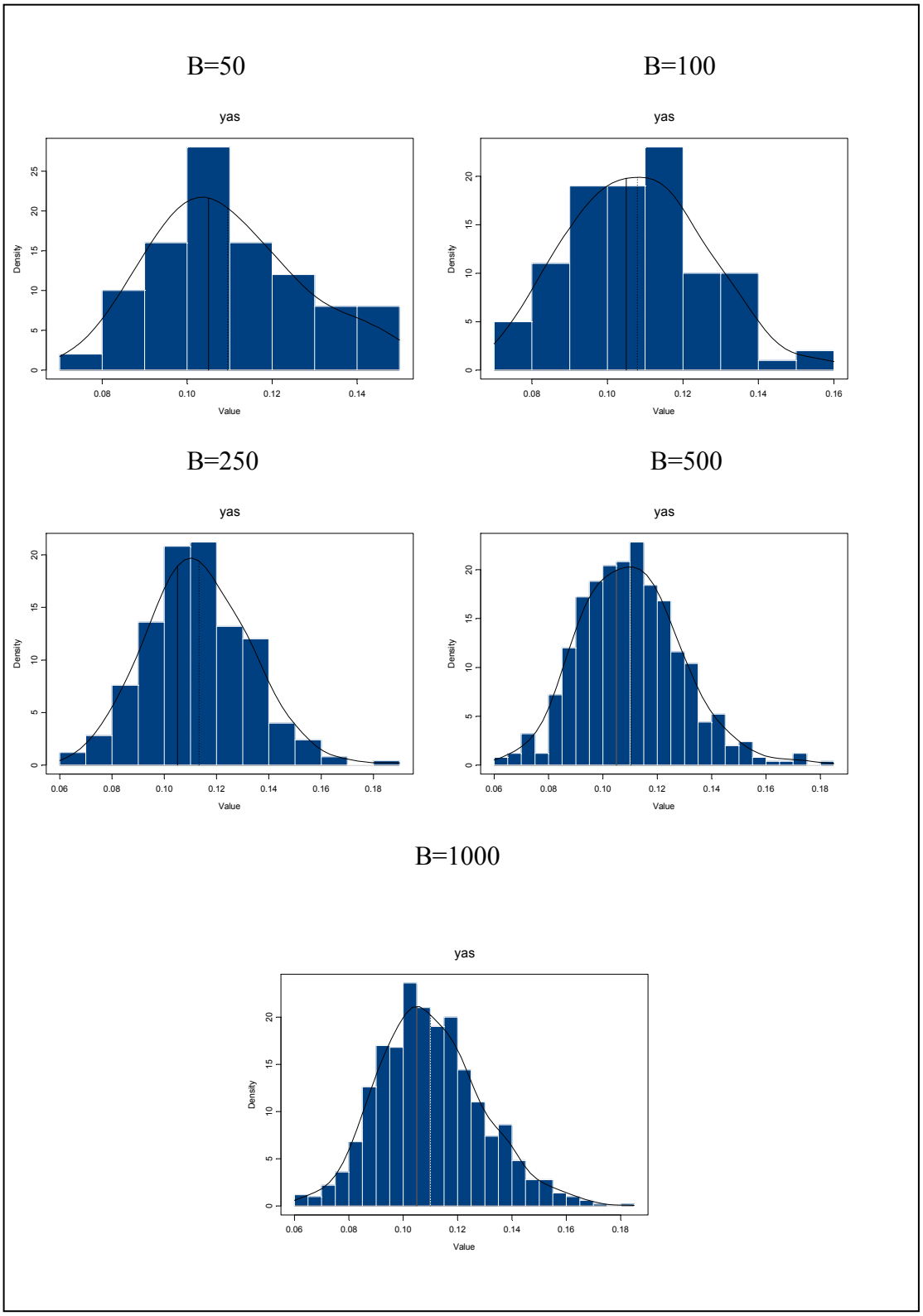
Bootstrap örneklemesi sonucu kurulan model katsayılarının geriye doğru eleme yöntemi sonucunda elde edilen model katsayılarından çok düşük düzeyde sapmalar gösterdiği Çizelge 5.16' da görülmektedir. Örneğin geriye doğru eleme yöntemi sonucunda sabit için elde edilen parametre tahmini -12,177 iken B=50 olarak alınan bootstrap örneklemesinde parametre tahmin sonucu -12,57351 olarak bulunmuştur. Aradaki sapma miktarının -0,39651 olarak hesaplanmıştır.

Çizelge 5.17' de bootstrap modellerinin standart hatalarının geriye doğru eleme yöntemi sonucunda elde edilen modelin standart hatalarıyla karşılaştırılması görülmektedir. Bu karşılaştırma sonucunda bootstrap modellerinin standart hatalarının geriye doğru eleme yöntemi sonucunda elde edilen modelin standart hatası düşük olduğu görülmektedir.

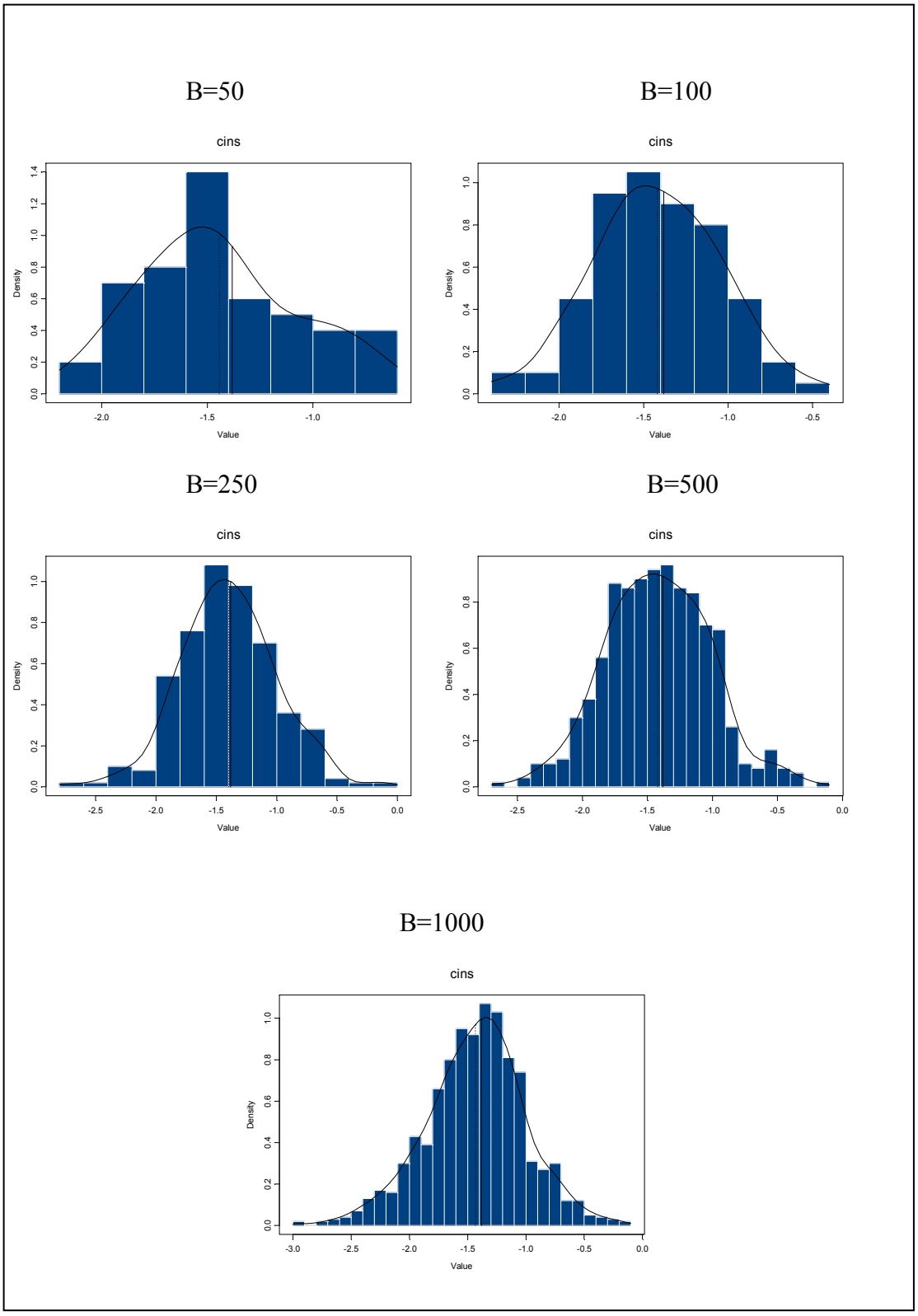
Bootstrap örneklemesinden elde edilen model katsayılarının güven aralıklarının, lojistik regresyon modeli katsayılarını kapsadığı ise Çizelge 5.18' de görülmektedir.



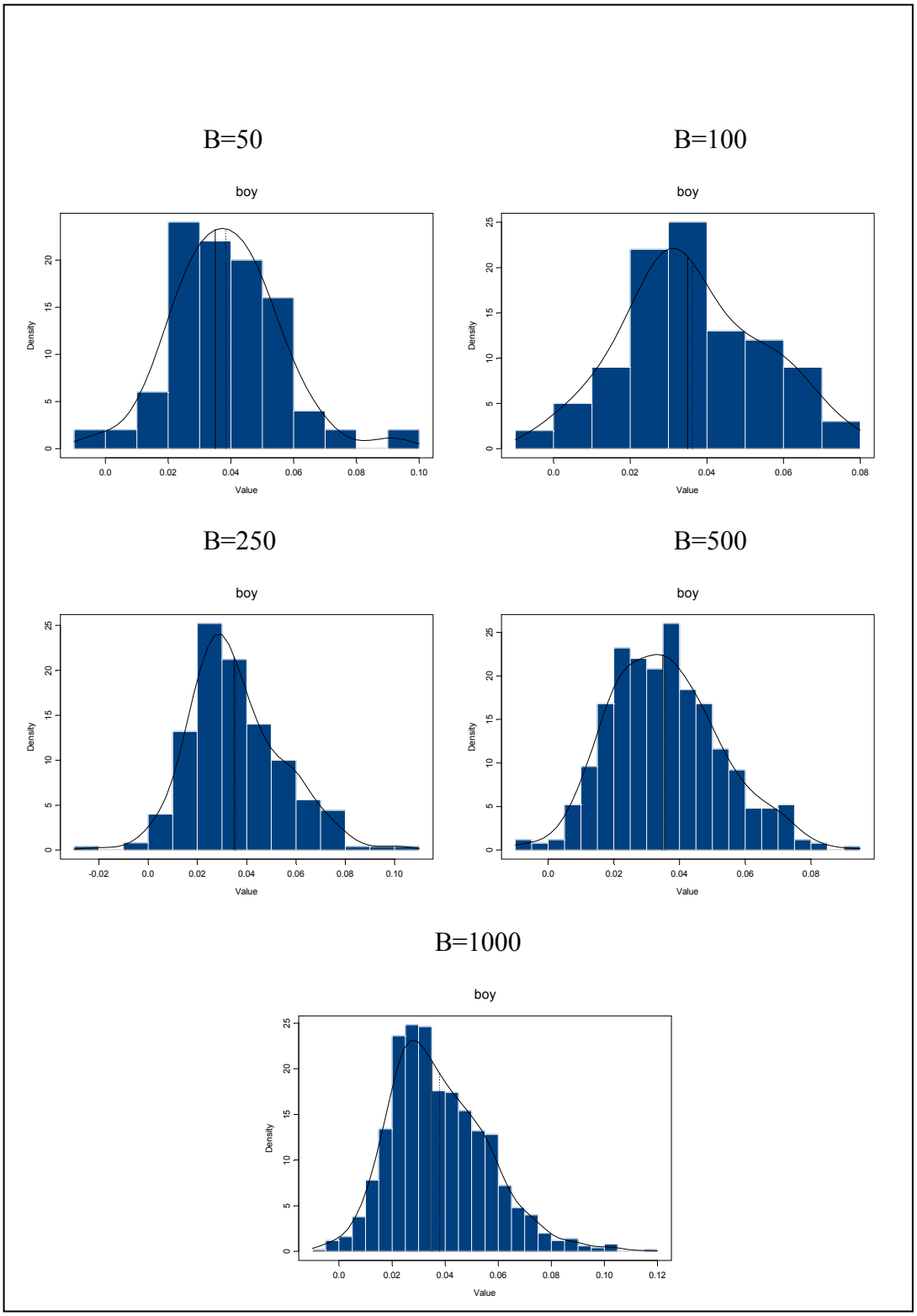
Şekil 5.12. Bootstrap yöntemi ile elde edilen sabitin histogram grafiği



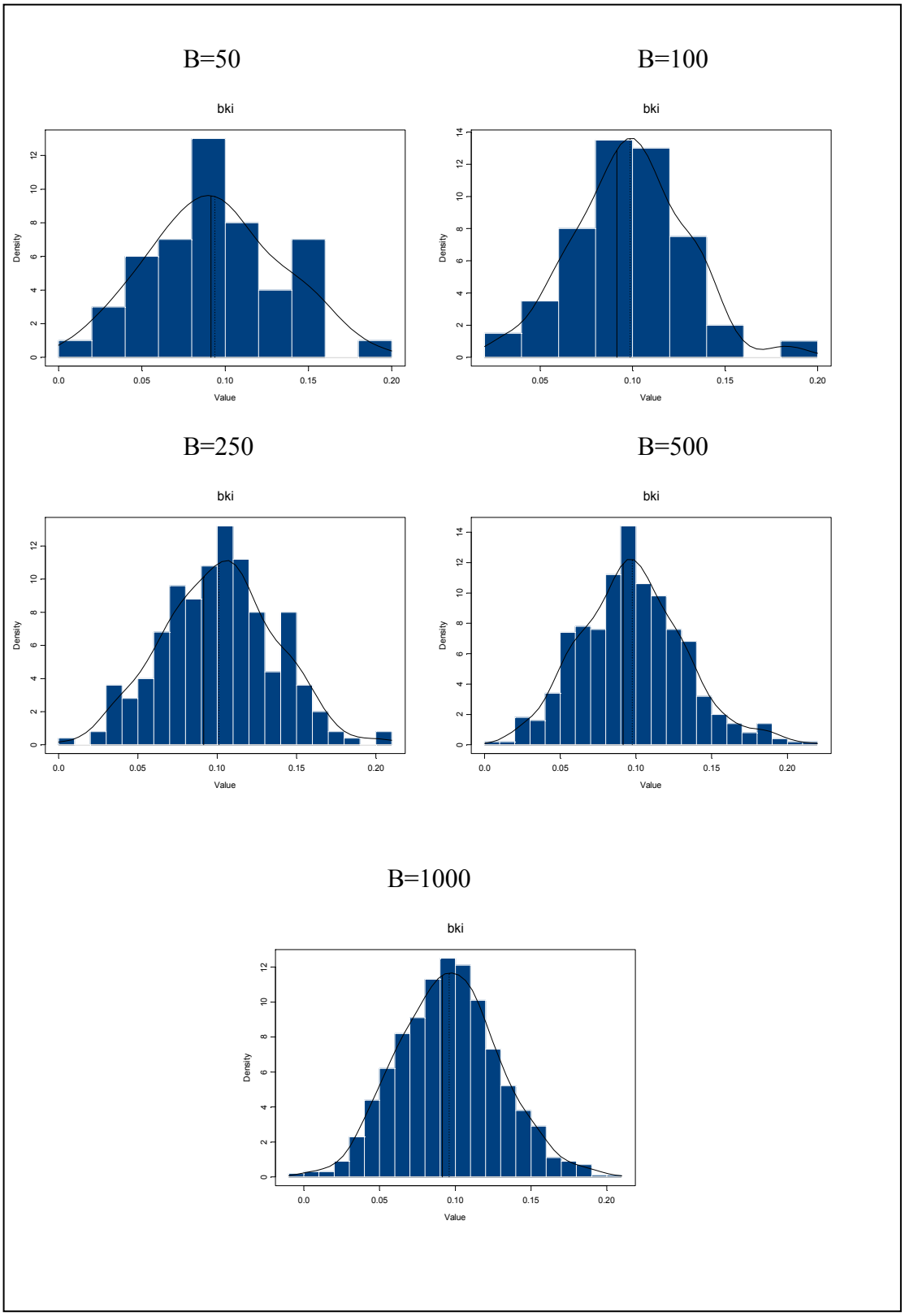
Şekil 5.13. Bootstrap yöntemi ile elde edilen yaş değişkeninin histogram grafiği



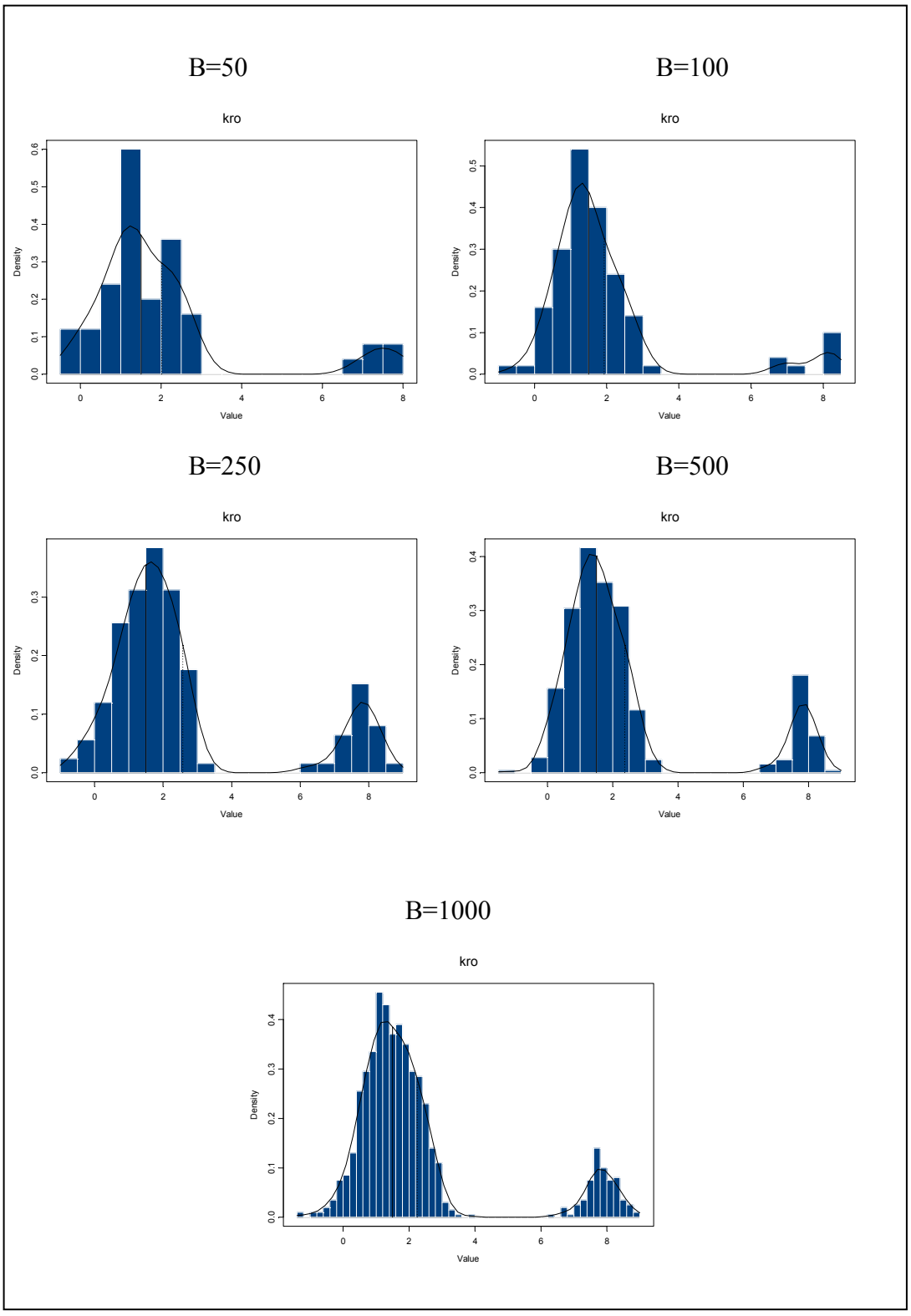
Şekil 5.14. Bootstrap yöntemi ile elde edilen cinsiyet değişkeninin histogram grafiği



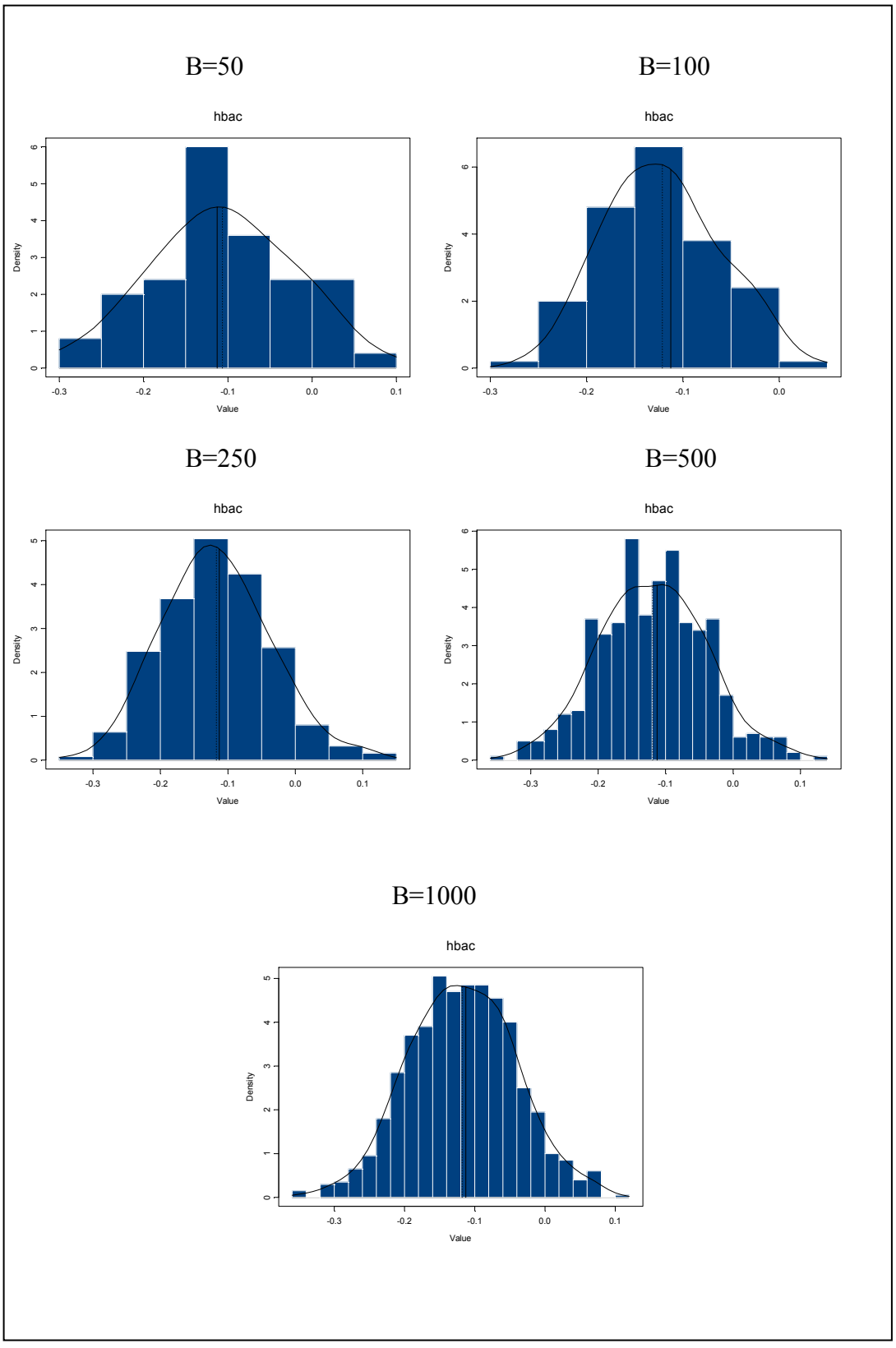
Şekil 5.15. Bootstrap yöntemi ile elde edilen boy değişkeninin histogram grafiği



Şekil 5.16. Bootstrap yöntemi ile elde edilen bki değişkeninin histogram grafiği



Şekil 5.17. Bootstrap yöntemi ile elde edilen kroarte değişkeninin histogram grafiği



Şekil 5.18. Bootstrap yöntemi ile elde edilen hba1c değişkeninin histogram grafiği

5.9. Sonuç

Lojistik Regresyon Analizinin kullanım amacı, en az değişkeni kullanarak en iyi uyuma sahip olacak şekilde bağımlı değişken ile bağımsız değişkenler arasındaki ilişkiyi tanımlayabilmek ve amaca yönelik kabul edilebilir bir model kurmaktır. Bu yöntemde, bağımsız değişkenlerin bağımlı değişken üzerindeki etkileri olasılık olarak hesaplanarak risk faktörlerinin olasılık olarak belirlenmesi sağlanır. Lojistik regresyonun diğer yöntemlerden farkı sonuç değişkeninin kesikli iki veya daha çok değer aldığı durumlarda kullanılıyor olmasıdır.

Lojistik regresyon analizinde bağımsız değişkenler sürekli değişkenler ve kategorik olarak bir arada kullanılabilmektedir. Bağımsız değişkenlerin dağılımları üzerinde hiçbir kısıtlayıcı varsayım bulunmamaktadır. Bu sebepten dolayı özellikle tıp alanında hastalığın birey üzerinde var olup olmadığının araştırıldığı çalışmalarda yoğun bir şekilde kullanılmaktadır. Bu amaç doğrultusunda yapılan bu çalışmada hastaneye hipertansiyon şikayeti ile başvuran bireylerde hastalığın olup olmadığına ilişkin bir ayrımsama modeli lojistik regresyon analizi kullanılarak elde edilmeye çalışılmıştır.

Diğer taraftan kurulmuş olan modelde değişken seçimi yapılarak, en az değişken kullanılarak hipertansiyon şikayeti ile gelen bireylerin hastalık grubunu doğru bir şekilde tahmin edecek en iyi model kurulmaya çalışılmıştır. Böylece az sayıda değişkenle zamandan kazanarak, işlem kolaylığının sağlanması hedeflenmiştir.

İstatistikte herhangi bir yığının parametrelerini tahmin etmek için o yığına ait gözlemlerden yararlanılır. Üzerinde çalışılan yığının tüm gözlemlerini parametre tahmini için kullanmak hem zaman kaybına yol açacağı hem de maliyeti arttıracağı için yığını en iyi şekilde açıklayacak olan örnekten elde edilen verilerle çalışmak bu sorunları ortadan kaldırır. İstenilen büyüklük ve miktarda veri setleri oluşturmak için herhangi bir boyuttaki veri setinden gözlemler tesadüfi olarak yer değiştirilerek yeniden örneklenebilir. Bu sayede veri setinden daha fazla bilgi alınabilir. Bu şekilde tanımlanan yöntem "*Bootstrap Yöntemi*" olarak adlandırılır.

Bu yöntemin temeli mevcut veri setinden çok daha büyük veri setleri üretmek için yeniden örnekleme yapmaktır. Bootstrap yönteminin geliştirilme amacı, örneklemin güven aralıklarını oluşturmak, standart hatasını küçültmek ve bunun sonucu olarak da olarak daha doğru tahminlere ulaşmak olarak özetlenebilir.

Bu çalışmada öncelikle lojistik regresyon analizi ve bootstrap yöntemi ele alınmıştır. Uygulama bölümünde ise hipertansiyon şikayeti ile hastaneye başvuran 148 kişinin lojistik regresyon analizinde geriye doğru eleme yöntemi kullanılarak parametre tahminleri yapılmış ve model kurulmuştur. Bu aşamadan sonra yeniden örnekleme tekniklerinden biri olan bootstrap yöntemiyle örnekler elde edilmiş ve lojistik regresyon analizi sonucunda elde edilen modelle karşılaştırılıp hangisinin daha etkili parametre tahminleri elde ettiği ortaya konulmaya çalışılmıştır.

Lojistik regresyon analizi sonuçları ile bootstrap uygulamasının sonuçları incelendiğinde, katsayı tahminleri arasında çok az bir fark olduğu görülmüştür. Bootstrap yönteminde parametre tahminlerinin standart hataları genellikle klasik örnekten elde edilen parametre tahminlerinin standart hatalarından, daha küçük bulunmuştur.

Örnek genişliğini artırarak yapılan tahmin ile daha küçük örnekten yeniden örnekleme yapılarak elde edilen tahmin arasında önemsenmeyecek kadar küçük sapmalar olmaktadır. Bu durumda büyük örnek üzerinde çalışarak tahmin yapılmaktansa, küçük örnekler ile yeniden örnekleme yöntemlerini uygulayarak çalışmak bizi benzer sonuçlara ulaşılabilmektedir. Ancak büyük örnek yerine küçük örnekler kullanıldığında yeniden örnekleme yaparak parametre değerlerini tahmin etmek her zaman olumlu sonuçlar vermeyebilir. Aynı örnek genişliğine sahip (n=148) örnekleme yöntemleri içinde klasik örnekleme yerine yeniden örnekleme yapmanın genellikle daha iyi sonuçlar verdiği ortaya çıkmıştır. Fakat bootstrap yöntemini her zaman güvenilir sonuçlar ortaya çıkarmayabilir. Yöntemin başarısı elde edilen verilerin yapısına ve deneysel dağılım fonksiyonunun ana kütlenin dağılımını iyi yansıtmasına bağlı olarak değişmektedir.

Uygulamada oluşturulan lojistik regresyon modeline dahil edilen yaş, cinsiyet, boy ve bki(beden kitle indeksi) değişkenlerinin hipertansiyona etkisinin anlamlı bulunması, fiziksel özelliklerin bu hastalıkla ağırlıklı olarak ilişkili olduğuna işaret etmektedir. Kilo değişkeninin beklenen oranda etkisinin görülmemesine rağmen, bki değişkeninin hesaplanmasında kullanılması ve bu değişken değerinin 30' un üzerinde olması halinde obezite olarak adlandırılması hipertansiyon hastalığıyla ilişkili olduğunu göstermektedir.

Koroner arter damar çapının normalliğinin gözlendiği kroarte kesikli değişkeninde, çap normal olmayan yapıdan(0) normal yapıya dönüştükçe(1) hipertansiyon hastalığının görülme riski azalmaktadır(hastalık var(1)' dan hastalık yok(0) a doğru).

Kandaki glikoz miktarının ölçüldüğü hba1c sürekli ölçüm değişkeninin de hipertansiyon hastalığını negatif yönde etkilediği tespit edilmiştir. Kandaki glikoz düzeyinin %6,5 değerinin altına düşmesi hipertansiyon riskini azaltmaktadır.

Geriye doğru eleme tekniğinin kullanıldığı ve 14. adım sonrası kurulan bu modelden elde edilen sonuçlar gerçek bulgular ile belirlenen anlam seviyesinde(0.25) paralellik göstermektedir.

Gözlenen veri setinin yığını temsil etme gücünün ölçüldüğü bootstrap yeniden örnekleme tekniğinde, parametrelerin standart hata değerlerinin daha küçük çıktığı görülmüştür.

Örneklem sayısının aynı veri seti hacminde (n=148) arttırılarak uygulandığı bootstrap tekniği zaman ve maliyet açısından yarar sağlayabilir. Analiz sonuçlarına göre, bu tekniğin uygulandığı veri setinden elde edilen parametre tahminlerinin, lojistik regresyon analizi sonucu elde edilen parametre tahminlerinden çok ufak sapmalar gösterdiği görülmüştür. Ayrıca lojistik model için elde edilen parametre tahmin değerlerinin, bootstrap örneklemesinin her aşamasında oluşturulan güven aralıklarının içerisinde yer alması veri setinin kitleyi temsil gücünün var olduğunu göstermektedir.

KAYNAKLAR

1. Aktükün, A., "Asal Bileşenler Analizinde Bootstrap Yaklaşımı", ***İstanbul Üniversitesi İktisat Fakültesi Ekonomi ve İstatistik Dergisi***, 1: 15-05, (2005).

2. Aldrich, H.J., Nelson, D.F., "Linear Probability Logit and Probit Models", *London*: ***Sage Publications***, 49-52 (1986).

3. Alpar, R., "Uygulamalı Çok Değişkenli İstatistiksel Yöntemlere Giriş-I" 2. baskı, Ankara, ***Nobel Yayınevi***, 89-91 (2003).

4. Anderson, J.A., "Robust Inference Using Logistic Models", ***Bulletion of International Statistical Institute,*** 48: 35-53 (1983).

5. Başarır, G., "Çok Değişkenli Verilerde Ayırsama Sorunu ve Lojistik Regresyon Analizi*",* Doktora tezi, ***Hacettepe Üniversitesi Fen Bilimleri Enstitüsü***, Ankara, 1-5, 20-21, 49-52 (1990).

6. Bircan, H., "Lojistik Regresyon Analizi ve Tıp Verileri Üzerine Bir Uygulama", ***Kocaeli Üniversitesi Sosyal Bilimler Dergisi,*** 2:185-208(2004).

7. Dobson, A.J., "An Introduction to Generalized Linear Models"*,* ***Chapmen Hall***, London, 121-124 (1990).

8. Dufy, D.E., "On Continuity - Corrected Residuals in Logistic Regression"*,* ***Biometrika***, 77(2) 287-293 (1990).

9. Efron, B., Tibshirani, R.J., "An Introduction to The Bootstrap", ***Chapman&Hall***, New York ,USA, 45-56, 88-92, 105-115, 296-307(1993).

10. Erdoğan, B.E., "Bankaların Mali Performanslarının Lojistik Regresyon ile Analizi ve İleriye Yönelik Tahmin", Doktora Tezi, **Marmara Üniversitesi Sosyal Bilimler Enstitüsü**, İstanbul, 24-25 (2002).

11. Hosmer, D.W., Lemeshow, Jr. S., "Applied Logistic Regression", ***John Wiley & Sons*** , New York , 1-29, 38-60, 63-66, 82-88 (1989).

12. Hosmer, D.W., Lemeshow, Jr.S., "Applied Logistic Regression", Second Edition, ***John Wiley & Sons***, New York, 11-17, 48-85, 116-128,143-147(2000).

13. İnternet: T.C. Boğaziçi Üniversitesi Mediko Sosyal Merkezi, "Hipertansiyon" http://www.mediko.boun.edu.tr/files/Hipertansiyon.htm(2010).

14. İnternet: Türk Toraks Derneği, "Korelasyon ve Regresyon Analizi" http://www.toraks.org.tr/mse-ppt-pdf/Kenan_KOSE3.pdf (2010).

15. İyit, N., "Lineer Olmayn Lojistik Regresyon Analizinde Model Kurma Stratejileri ve Bir Uygulaması" Yüksek Lisans Tezi, ***Selçuk Üniversitesi Fen Bilimleri Enstitüsü***, Konya, 53-57 (2003).

16. Kleinbaum, D.G., Klein, M., "Logistic Regression, A Self-Learning Text", Second Edition, ***Springer-Verlag***, New York, 4-9, 164-165 (2002).

17. Kleinbaum, D.G., Kupper, L.L., Muller, K.E., "Applied Regression Analysis and Other Multivariable Methods", Fourth Edition. ***Thomson Brooks/Cole***, Boston, 191-198 (2008).

18. Lesaffre, E., Albert, A., "A Multiple Group Logistic Regression Diagnostics", ***Applied Statistics***, 38, 3, 425-440 (1989).

19. McCullagh, P., Nelder, J.A., "Generalized Lineer Models", Second Edition, ***Chapmen Hall***, London, 28-30, 114-115, 171-174 (1989).

20. Menard, S., "Applied Logistic Regression Anaiysis", Second Edition, London: ***Sage Publications,*** 12-14 (2002).

21. Neter, J., Kunter, M.H., Nachtsheim, C.J., "Applied Linear Regression Models", Fourth Edition, ***The Mc Graw – Hill Companies, Inc.***, New York, 591-598 (2004).

22. Özdamar, K., "Paket Programlar ile İstatistiksel Veri Analizi 1", ***Kaan Kitabevi***, Eskişehir, 197-198, 461-462 (1997).

23. Şahin, M., "Lojistik Regresyon ve Biyolojik Alanlarda Kullanımı*"* Yüksek Lisans Tezi, ***Kahramanmaraş Sütçü İmam Üniversitesi Fen Bilimleri Enstitüsü***, Kahramanmaraş, 2-8 (1999).

24. Tatlıdil, H., "*Uygulamalı Çok Değişkenli İstatistiksel Analiz", **Cem Ofset***, Ankara, 292-293 (1992).

25. Tezcan, B., "Lojistik Regresyon Analizi ve Sigortacılık sektöründe bir uygulama*",* Yüksek Lisans Tezi, ***Marmara Üniversitesi Bankacılık ve Sigortacılık Enstitüsü***, İstanbul, 33-34 (2006).

26. Zhao, L., Chen, Y., Schaffner, D.W., "Comparison of Logistic Regression and Linear Regresyon in Modeling Percentage Data", ***Applied and Environmental Microbiology*** 5, 2129-2135 (2001).

EKLER

EK-1 Lojistik regresyon modelinde geriye doğru eliminasyon yöntemi adımları

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	yaş	,116	,031	14,233	1	,000	1,123	1,057	1,192
	cinsiyet(1)	1,520	,673	5,101	1	,024	4,571	1,222	17,089
	boy	,068	,078	,768	1	,381	1,071	,919	1,247
	kilo	-,030	,062	,231	1	,631	,971	,860	1,096
	bki	,167	,167	1,009	1	,315	1,182	,853	1,639
	hiplipid(1)	-,309	,496	,388	1	,533	,734	,278	1,941
	kroarte(1)	-1,947	1,255	2,406	1	,121	,143	,012	1,671
	diabsure	-,014	,059	,055	1	,814	,986	,879	1,107
	inslin(1)	,005	,697	,000	1	,994	1,005	,257	3,940
	yoguins(1)	-,378	,886	,182	1	,670	,685	,121	3,893
	metfrmn(1)	-,342	,588	,339	1	,560	,710	,224	2,246
	hba1c	-,131	,107	1,496	1	,221	,878	,712	1,082
	ure	,020	,029	,449	1	,503	1,020	,963	1,080
	kreatin	-,742	1,314	,319	1	,572	,476	,036	6,255
	tk	,008	,008	1,045	1	,307	1,008	,993	1,023
	tg	-,015	,010	2,147	1	,143	,985	,965	1,005
	hdl	-,019	,016	1,441	1	,230	,981	,950	1,012
	ldl	-,010	,009	1,221	1	,269	,990	,972	1,008
	vldl	,076	,052	2,167	1	,141	1,079	,975	1,194
	Constant	-16,271	13,052	1,554	1	,213	,000		
Step 2[a]	yaş	,116	,031	14,282	1	,000	1,123	1,057	1,192
	cinsiyet(1)	1,520	,670	5,154	1	,023	4,573	1,231	16,989
	boy	,068	,078	,768	1	,381	1,071	,919	1,247
	kilo	-,030	,062	,231	1	,631	,971	,860	1,096
	bki	,167	,166	1,011	1	,315	1,182	,853	1,638
	hiplipid(1)	-,309	,495	,388	1	,533	,734	,278	1,939
	kroarte(1)	-1,947	1,255	2,406	1	,121	,143	,012	1,671
	diabsure	-,014	,058	,057	1	,812	,986	,879	1,106
	yoguins(1)	-,374	,704	,282	1	,595	,688	,173	2,735
	metfrmn(1)	-,342	,587	,340	1	,560	,710	,225	2,245
	hba1c	-,131	,104	1,577	1	,209	,877	,715	1,076
	ure	,020	,029	,449	1	,503	1,020	,963	1,080
	kreatin	-,743	1,309	,322	1	,570	,476	,037	6,191
	tk	,008	,008	1,045	1	,307	1,008	,993	1,023
	tg	-,015	,010	2,147	1	,143	,985	,965	1,005
	hdl	-,019	,016	1,459	1	,227	,981	,951	1,012
	ldl	-,010	,009	1,225	1	,268	,990	,972	1,008
	vldl	,076	,052	2,167	1	,141	1,079	,975	1,194
	Constant	-16,264	13,021	1,560	1	,212	,000		
Step 3[a]	yaş	,114	,030	14,721	1	,000	1,121	1,057	1,188
	cinsiyet(1)	1,510	,666	5,145	1	,023	4,526	1,228	16,685
	boy	,067	,079	,732	1	,392	1,070	,917	1,248
	kilo	-,029	,063	,212	1	,645	,971	,859	1,099
	bki	,166	,169	,962	1	,327	1,180	,847	1,644
	hiplipid(1)	-,332	,486	,466	1	,495	,718	,277	1,860
	kroarte(1)	-1,922	1,247	2,377	1	,123	,146	,013	1,685
	yoguins(1)	-,338	,686	,243	1	,622	,713	,186	2,736
	metfrmn(1)	-,353	,586	,364	1	,547	,703	,223	2,213
	hba1c	-,135	,103	1,733	1	,188	,874	,714	1,068
	ure	,019	,029	,431	1	,511	1,019	,963	1,079
	kreatin	-,733	1,301	,317	1	,573	,481	,038	6,159
	tk	,007	,007	,999	1	,318	1,008	,993	1,022
	tg	-,015	,010	2,188	1	,139	,985	,965	1,005
	hdl	-,020	,016	1,525	1	,217	,981	,950	1,012
	ldl	-,010	,009	1,169	1	,280	,990	,973	1,008
	vldl	,077	,052	2,210	1	,137	1,080	,976	1,195
	Constant	-16,106	13,147	1,501	1	,221	,000		
Step 4[a]	yaş	,114	,030	14,769	1	,000	1,121	1,058	1,188
	cinsiyet(1)	1,505	,661	5,173	1	,023	4,502	1,231	16,462
	boy	,067	,079	,705	1	,401	1,069	,915	1,249
	kilo	-,029	,064	,209	1	,648	,971	,857	1,101
	bki	,166	,171	,940	1	,332	1,181	,844	1,652
	hiplipid(1)	-,329	,485	,459	1	,498	,720	,278	1,864
	kroarte(1)	-1,962	1,238	2,511	1	,113	,141	,012	1,592
	metfrmn(1)	-,304	,575	,280	1	,597	,738	,239	2,275
	hba1c	-,124	,100	1,541	1	,215	,884	,727	1,074
	ure	,021	,029	,501	1	,479	1,021	,964	1,081
	kreatin	-,764	1,301	,345	1	,557	,466	,036	5,964
	tk	,007	,007	,918	1	,338	1,007	,993	1,022
	tg	-,016	,010	2,234	1	,135	,985	,965	1,005
	hdl	-,019	,016	1,457	1	,227	,981	,951	1,012
	ldl	-,010	,009	1,160	1	,282	,990	,973	1,008
	vldl	,078	,052	2,260	1	,133	1,081	,977	1,197
	Constant	-16,332	13,247	1,520	1	,218	,000		
Step 5[a]	yaş	,117	,030	15,603	1	,000	1,124	1,061	1,191
	cinsiyet(1)	1,457	,654	4,970	1	,026	4,295	1,193	15,466
	boy	,067	,078	,740	1	,390	1,070	,917	1,248
	kilo	-,029	,063	,215	1	,643	,971	,859	1,098
	bki	,164	,168	,953	1	,329	1,179	,847	1,640
	hiplipid(1)	-,317	,485	,428	1	,513	,728	,282	1,883
	kroarte(1)	-1,818	1,196	2,310	1	,129	,162	,016	1,693
	hba1c	-,128	,100	1,658	1	,198	,880	,724	1,069
	ure	,020	,029	,492	1	,483	1,021	,964	1,080
	kreatin	-,861	1,281	,452	1	,502	,423	,034	5,204
	tk	,007	,007	,935	1	,334	1,007	,993	1,022
	tg	-,016	,010	2,292	1	,130	,984	,965	1,005
	hdl	-,019	,016	1,382	1	,240	,982	,952	1,012
	ldl	-,010	,009	1,203	1	,273	,990	,973	1,008
	vldl	,079	,051	2,375	1	,123	1,083	,979	1,198
	Constant	-16,650	13,121	1,610	1	,204	,000		
Step 6[a]	yaş	,118	,029	15,995	1	,000	1,125	1,062	1,192
	cinsiyet(1)	1,447	,632	5,237	1	,022	4,252	1,231	14,684
	boy	,035	,029	1,516	1	,218	1,036	,979	1,096
	bki	,089	,042	4,625	1	,032	1,094	1,008	1,186
	hiplipid(1)	-,280	,482	,337	1	,561	,756	,294	1,944
	kroarte(1)	-1,788	1,194	2,244	1	,134	,167	,016	1,736
	hba1c	-,127	,100	1,631	1	,202	,880	,724	1,070
	ure	,021	,029	,539	1	,463	1,022	,965	1,082
	kreatin	-,880	1,284	,470	1	,493	,415	,033	5,141
	tk	,007	,007	1,025	1	,311	1,007	,993	1,022
	tg	-,016	,010	2,351	1	,125	,984	,965	1,004
	hdl	-,019	,016	1,409	1	,235	,981	,951	1,012
	ldl	-,010	,009	1,208	1	,272	,990	,973	1,008
	vldl	,080	,051	2,431	1	,119	1,084	,980	1,199
	Constant	-11,631	5,790	4,035	1	,045	,000		

EK-1 **(**Devam**)** Lojistik regresyon modelinde geriye doğru eliminasyon yöntemi adımları

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 7[a]	yaş	,118	,029	16,233	1	,000	1,125	1,063	1,192
	cinsiyet(1)	1,444	,633	5,206	1	,023	4,237	1,226	14,648
	boy	,037	,029	1,578	1	,209	1,037	,980	1,099
	bki	,089	,041	4,682	1	,030	1,093	1,008	1,186
	kroarte(1)	-1,757	1,187	2,190	1	,139	,173	,017	1,768
	hba1c	-,128	,100	1,627	1	,202	,880	,723	1,071
	ure	,021	,029	,520	1	,471	1,021	,965	1,081
	kreatin	-,896	1,284	,487	1	,485	,408	,033	5,059
	tk	,008	,007	1,345	1	,246	1,008	,994	1,023
	tg	-,015	,010	2,212	1	,137	,986	,967	1,005
	hdl	-,018	,016	1,268	1	,260	,982	,953	1,013
	ldl	-,010	,009	1,168	1	,280	,990	,973	1,008
	vldl	,075	,049	2,314	1	,128	1,078	,979	1,188
	Constant	-12,277	5,799	4,481	1	,034	,000		
Step 8[a]	yaş	,114	,028	16,043	1	,000	1,121	1,060	1,185
	cinsiyet(1)	1,555	,618	6,325	1	,012	4,735	1,409	15,910
	boy	,035	,029	1,469	1	,226	1,036	,978	1,097
	bki	,085	,040	4,417	1	,036	1,089	1,006	1,178
	kroarte(1)	-1,685	1,184	2,027	1	,155	,185	,018	1,887
	hba1c	-,123	,099	1,539	1	,215	,884	,727	1,074
	ure	,013	,027	,245	1	,620	1,013	,961	1,069
	tk	,008	,007	1,173	1	,279	1,008	,994	1,022
	tg	-,014	,010	2,138	1	,144	,986	,967	1,005
	hdl	-,017	,016	1,188	1	,276	,983	,953	1,014
	ldl	-,009	,009	,961	1	,327	,992	,975	1,009
	vldl	,074	,049	2,248	1	,134	1,077	,978	1,186
	Constant	-12,464	5,796	4,624	1	,032	,000		
Step 9[a]	yaş	,118	,027	19,192	1	,000	1,126	1,068	1,187
	cinsiyet(1)	1,530	,615	6,186	1	,013	4,619	1,383	15,425
	boy	,036	,029	1,531	1	,216	1,037	,979	1,098
	bki	,084	,040	4,364	1	,037	1,088	1,005	1,178
	kroarte(1)	-1,698	1,185	2,052	1	,152	,183	,018	1,868
	hba1c	-,127	,099	1,645	1	,200	,881	,726	1,069
	tk	,007	,007	1,063	1	,303	1,007	,994	1,021
	tg	-,014	,010	2,040	1	,153	,986	,967	1,005
	hdl	-,016	,016	1,111	1	,292	,984	,954	1,014
	ldl	-,008	,009	,824	1	,364	,992	,976	1,009
	vldl	,072	,049	2,154	1	,142	1,075	,976	1,184
	Constant	-12,389	5,781	4,592	1	,032	,000		
Step 10[a]	yaş	,115	,027	18,872	1	,000	1,122	1,065	1,182
	cinsiyet(1)	1,544	,616	6,281	1	,012	4,681	1,400	15,654
	boy	,037	,029	1,620	1	,203	1,038	,980	1,100
	bki	,085	,040	4,449	1	,035	1,089	1,006	1,178
	kroarte(1)	-1,673	1,184	1,996	1	,158	,188	,018	1,911
	hba1c	-,122	,099	1,527	1	,217	,885	,730	1,074
	tk	,002	,005	,280	1	,597	1,002	,993	1,012
	tg	-,013	,010	1,942	1	,163	,987	,968	1,005
	hdl	-,018	,015	1,304	1	,253	,982	,953	1,013
	vldl	,072	,049	2,200	1	,138	1,075	,977	1,182
	Constant	-12,554	5,830	4,638	1	,031	,000		
Step 11[a]	yaş	,114	,026	18,817	1	,000	1,120	1,064	1,180
	cinsiyet(1)	1,593	,613	6,754	1	,009	4,920	1,479	16,364
	boy	,040	,030	1,792	1	,181	1,040	,982	1,102
	bki	,083	,040	4,272	1	,039	1,086	1,004	1,175
	kroarte(1)	-1,722	1,176	2,145	1	,143	,179	,018	1,791
	hba1c	-,115	,098	1,374	1	,241	,892	,736	1,080
	tg	-,013	,010	1,831	1	,176	,987	,968	1,006
	hdl	-,017	,016	1,198	1	,274	,983	,954	1,014
	vldl	,072	,049	2,169	1	,141	1,075	,976	1,183
	Constant	-12,400	5,869	4,465	1	,035	,000		
Step 12[a]	yaş	,112	,026	18,550	1	,000	1,118	1,063	1,177
	cinsiyet(1)	1,451	,591	6,036	1	,014	4,269	1,341	13,588
	boy	,039	,029	1,788	1	,181	1,040	,982	1,102
	bki	,086	,040	4,667	1	,031	1,089	1,008	1,178
	kroarte(1)	-1,684	1,172	2,065	1	,151	,186	,019	1,846
	hba1c	-,122	,096	1,613	1	,204	,885	,733	1,069
	tg	-,010	,009	1,265	1	,261	,990	,972	1,008
	vldl	,060	,047	1,630	1	,202	1,062	,968	1,164
	Constant	-13,133	5,814	5,102	1	,024	,000		
Step 13[a]	yaş	,107	,025	17,924	1	,000	1,113	1,059	1,170
	cinsiyet(1)	1,449	,583	6,168	1	,013	4,259	1,357	13,364
	boy	,036	,028	1,601	1	,206	1,036	,981	1,095
	bki	,088	,039	5,041	1	,025	1,092	1,011	1,179
	kroarte(1)	-1,467	1,151	1,625	1	,202	,231	,024	2,200
	hba1c	-,127	,096	1,765	1	,184	,881	,730	1,062
	vldl	,007	,009	,768	1	,381	1,007	,991	1,024
	Constant	-12,525	5,602	4,999	1	,025	,000		
Step 14[a]	yaş	,105	,025	17,442	1	,000	1,111	1,057	1,167
	cinsiyet(1)	1,382	,572	5,843	1	,016	3,981	1,299	12,204
	boy	,035	,028	1,580	1	,209	1,036	,981	1,094
	bki	,092	,039	5,492	1	,019	1,096	1,015	1,183
	kroarte(1)	-1,501	1,142	1,727	1	,189	,223	,024	2,091
	hba1c	-,113	,094	1,428	1	,232	,894	,743	1,075
	Constant	-12,177	5,524	4,859	1	,028	,000		

ÖZGEÇMİŞ

Kişisel Bilgiler

Soyadı, adı	: ATABEY, Özgür
Uyruğu	: T.C.
Doğum tarihi ve yeri	: 02.10.1982 Lüleburgaz
Medeni hali	: Bekar
Telefon	: 0 (312) 229 78 81
Faks	: 0 (312) 229 78 82
e-mail	: ozgur_atabey@hotmail.com

Eğitim

Derece	Eğitim Birimi	Mezuniyet tarihi
Yüksek lisans	Gazi Üniversitesi /İstatistik A.B.D.	2010
Lisans	Selçuk Üniversitesi/ İstatistik Bölümü	2004
Lise	Fatih Sultan Mehmet Lisesi	1999

İş Deneyimi

Yıl	Yer	Görev
2005-2006	Havelsan A.Ş.	Eğitim Uzmanı
2007-	Özata Medikal	Şirket Sahibi

Yabancı Dil

İngilizce

Hobiler

Kitap okumak, Müzik dinlemek, Spor yapmak