ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ

Selin SEVİNDİK

DİSKRİMİNANT ANALİZİ VE BAZI ALTERNATİF REGRESYON ANALİZLERİ

İSTATİSTİK ANABİLİM DALI

ADANA, 2018

ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

DİSKRİMİNANT ANALİZİ VE BAZI ALTERNATİF REGRESYON ANALİZLERİ

Selin SEVİNDİK

YÜKSEK LİSANS TEZİ

İSTATİSTİK ANABİLİM DALI

Bu Tez 15/02/2018 Tarihinde Aşağıdaki Jüri Üyeleri Tarafından Oybirliği/ Oyçokluğu İle Kabul Edilmiştir.

……	……	……..
Doç. Dr. Gülesen ÜSTÜNDAĞ ŞİRAY DANIŞMAN	Prof.Dr. Sadullah SAKALLIOĞLU ÜYE	Yrd. Doç. Dr. Ahmet ÇALIK ÜYE

Bu tez Enstitümüz İstatistik Anabilim Dalında hazırlanmıştır.
Kod No:

Prof. Dr. Mustafa GÖK
Enstitü Müdürü

Bu çalışma Ç. Ü. Bilimsel Araştırma Projeleri Birimi Tarafından Desteklenmiştir.
Proje No: FYL-2016-5808

ÖZ

YÜKSEK LİSANS TEZİ

DİSKRİMİNANT ANALİZİ VE BAZI ALTERNATİF REGRESYON ANALİZLERİ

Selin SEVİNDİK

ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
İSTATİSTİK ANABİLİM DALI

Danışman : Doç. Dr. Gülesen ÜSTÜNDAĞ ŞİRAY
Yıl: 2018, Sayfa:197
Jüri : Doç. Dr. Gülesen ÜSTÜNDAĞ ŞİRAY
: Prof. Dr. Sadullah SAKALLIOĞLU
: Yrd. Doç. Dr. Ahmet ÇALIK

Araştırılan yanıt değişkenin nitel ölçekli yapıya sahip olması sıkça rastlanan bir durumdur. Yanıt değişkenin yapısının iki veya daha fazla kategoriye sahip olduğu durumlarda, bu yanıt değişkenin tahminini elde etmek için lineer regresyon yöntemleri yerine sınıflandırma yöntemleri kullanılır. Diskriminant analizi, hatalı sınıflandırma olasılığını en aza indirgeyerek gözlemleri ait oldukları kategorilere ayırmak için kullanılan en temel sınıflandırma yöntemlerinden birisidir.

Bu tez çalışmasında diskriminant analizi ve bu analize alternatif olarak önerilen lojistik ve probit regresyon analizleri, iki ve ikiden fazla kategoriye sahip yanıt değişken için ayrı ayrı incelenmiştir. Bu amaçla, ilk olarak bazı sınıflandırma yöntemleri hakkında ön bilgi verilmiş ve neden lineer regresyonun uygulanamadığı açıklanmıştır. Değişkenlerin çok değişkenli normal dağılması ve ortak varyans-kovaryans matrisli olmaları varsayımlarını gerektiren lineer diskriminant analizi ile ortak varyans-kovaryans matrisli olma koşulunu gerektirmeyen karesel diskriminant analizi incelenmiştir. Bu varsayımların gerçekleşmediği koşullarda önerilen, lojistik ve probit regresyon analizleri sırasıyla ele alınmıştır. Yapılan uygulama ile elde edilen bulgular özetlenmiş ve bu yöntemlerin sınıflandırma doğrulukları karşılaştırılmıştır.

Anahtar Kelimeler: Lineer diskriminant analizi, Karesel diskriminant analizi, Lojistik regresyon analizi, Probit Regresyon analizi, Sınıflandırma

ABSTRACT

MSc THESIS

DISCRIMINANT ANALYSIS AND SOME ALTERNATIVE REGRESSION ANALYSIS

Selin SEVİNDİK

CUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES
DEPARTMENT OF STATISTICS

Supervisor : Assoc. Prof. Dr. Gülesen ÜSTÜNDAĞ ŞİRAY
Year: 2018, Page:197
Jury : Assoc. Prof. Dr. Gülesen ÜSTÜNDAĞ ŞİRAY
: Prof. Dr. Sadullah SAKALLIOĞLU
: Asst. Prof. Dr. Ahmet ÇALIK

It is a frequent occurrence that the response variable studied has a qualitative scale structure. In cases where the response variable structure has two or more categories, the classification methods are used instead of the linear regression methods to obtain an estimate of the response variable. Discriminant analysis is one of the most basic classification methods used to separate the observations into categories they belong to by minimizing the probability of incorrect classification.

In this thesis, in addition to the discriminant analysis, logistic and probit regression analyzes which are proposed as an alternative to discriminant analysis, are examined separately for response variables with two and more than two categories. For this purpose, firstly some preliminary information about the classification methods are given and it is described that why the linear regression can not be applied. Linear discriminant analysis which requires the assumption that the variables are multivariate normal distribution with common variance-covariance matrix and quadratic discriminant analysis that does not require common variance covariance matrix, are examined. The logistic and probit regression analyzes, which are proposed in the conditions where these assumptions do not take place, are given respectively. The findings obtained by the application are summarized and the classification accuracy of these methods are compared.

Key words: Linear discriminant analysis, Quadratic discriminant analysis, Logistic regression analysis, Probit regression analysis, Classification

GENİŞLETİLMİŞ ÖZET

Bir gözlem değeri için kategorik yapıda bir yanıt tahmin etmek, bu gözlemi bir gruba sınıflandırmak olarak adlandırlır. Yanıt değişkenin yapısının iki veya daha fazla kategoriye sahip olduğu böyle durumlarda, bu yanıt değişkenin tahminini elde etmek için lineer regresyon yöntemleri yerine sınıflandırma yöntemleri kullanılır. Diskriminant analizi sınıflandırma yöntemlerinin en başında gelen yöntemlerdendir ancak belirli varsayımsal kısıtlamalarından dolayı kendisine alternatif yöntemler önerilmektedir.

Bu tez çalışmasında en temel sınıflandırma yöntemlerinden biri olan diskriminant analizi ve bu analize alternatif olarak önerilen lojistik regresyon analizi ve probit regresyon analizi incelenmektedir.

Diskriminant analizi, gözlem (birey) üzerinden ölçümler alındığında bu ölçümlere bağlı olarak gözlemi, daha önceden bilinen sonlu sayıdaki farklı gruplardan birisine atayan istatistiksel bir sınıflandırma yöntemidir. Temel olarak, incelenen gözlemin gerçek grubunun belirlenmesinde kullanılacak diskriminant fonksiyonlarının bulunması ve bu fonksiyonları kullanarak oluşturulacak atama kuralına uygun biçimde söz konusu bireyin o grupta bulunma olasılıkları aracılığıyla atanacağı grubun belirlenmesi amaçlanır. Diskriminant fonksiyonları i_inci gruptaki n_i gözlemli k tane grup (örneklem) için, her bir gözlem vektörü $\boldsymbol{x}_{ij}$ kullanılarak

$$y_{ij} = \boldsymbol{b}'\boldsymbol{x}_{ij},\ i = 1, 2, \ldots, k,\ j = 1, 2, \ldots, n_i \qquad (1)$$

eşitliği ile ifade edilir. Daha yalın bir ifadeyle bu analiz, p açıklayıcı değişken sayısı olmak üzere

$$\boldsymbol{y} = \boldsymbol{b}'\boldsymbol{X} = b_1 x_1 + b_2 x_2 + ... + b_p x_p \quad (2)$$

eşitliğine benzer sınıflandırma fonksiyonları sağlamaktadır. Diskriminant analizini uygulayabilmek için;

1) Gruplar çok değişkenli normal dağılım göstermeli,
2) Değişkenlere ait varyans-kovaryanslar homojen olmalı,
3) Değişkenlerin ortalamaları ile varyansları arasında bir ilişki bulunmamalı,
4) Değişkenler arasında çoklu bağlantı (Multicolinearity) bulunmamalı,
5) Gruplar gereğinden fazla ve gereksiz değişken içermemeli

şeklindeki varsayımların sağlanması gerekmektedir. Bu varsayımların sağlanma durumuna göre diskriminant analizi, doğrusal (linear) ve karesel (quadratic) diskriminant analizi olarak iki ana gruba ayrılarak incelenmektedir. Lineer diskriminant analizi (LDA) çok değişkenli normal dağılım gösteren kitlelerden rasgele çekilmiş örneklem veri matrislerinin gruplar arası varyans-kovaryans matrislerinin ($S_1 = S_2 = ... = S_k$) eşit olması koşulunda uygulanabilmektedir. Gruplar arası varyans-kovaryans matrisleri eşit değil ise karesel diskriminant analizi (QDA) uygulanmaktadır. Gruplar arası varyans-kovaryansının eşit olup olmadığı Box's M testi ile kontrol edilirek veri yapılarına göre hangi diskriminant yönteminin uygulanması gerektiğine karar verilmektedir.

Diskriminant analizi yönteminin temel varsayımlarından olan, değişkenlerin çok değişkenli normal dağılması ve grupların ortak varyans-kovaryans matrisli olmaları varsayımlarının sağlanmaması durumunda diskriminant analizi yerine lojistik regresyon analizi (LRA) kullanılabilmektedir.

LRA, yanıt değişkenin ikili ya da çoklu, açıklayıcı değişkenlerin ise kategorik, sıralı veya sürekli olması durumunda belirli bir dağılım varsayımına

bağlı kalmaksızın yanıt değişken ile açıklayıcı değişkenler arasındaki neden-sonuç ilişkisinin belirlenmesinde ve aynı zamanda açıklayıcı değişkenlerin etkilerine dayanarak verilerin sınıflandırılmasında kullanılan bir yöntemdir. Lojistik regresyonun bir diğer özelliği de, elde edilen sonuçlarda değerlerin sadece hangi grupta olduklarını değil, aynı zamanda diskriminant analizine benzer olarak o grupta bulunma olasılıklarını da belirlemesidir.

Lojistik regresyon analizinin amacı, istatistikte kullanılan diğer model yapılandırma yöntemlerine benzer şekilde; en az değişkeni kullanarak, yanıt değişkeni ile açıklayıcı değişkenler arasındaki ilişkiyi tanımlayabilen en iyi uyuma sahip ve yorumlanabilen bir model bulmaktır. Lojistik regresyon analizinde, yanıt değişken doğrudan modellenememektedir. *k* tane grup için ikişerli karşılaştırma yapabilen (*k-1*) tane lojistik modele ihtiyaç duyulur. Bu durumda koşulu olasılıklar $\boldsymbol{y}_i$ 'nin *k* grubu için

$$\mathrm{P}(\boldsymbol{y}_i = j \mid \boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)}{\sum_{j=0}^{k-1} \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)}$$

ve lojistik fonksiyonlar $\boldsymbol{\beta}_j' = [\beta_{j0}, \beta_{j1}, \beta_{j2}, ..., \beta_{jp}]$ olmak üzere

$$\ln\left(\frac{\mathrm{P}(\boldsymbol{y}_i = j \mid \boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{y}_i = 0 \mid \boldsymbol{x}_i)}\right) = \boldsymbol{x}_i'\boldsymbol{\beta}_j \qquad j = 1, 2, ..., k \tag{3}$$

genel formunda ifade edilmektedir. Burada $\ln\left(\frac{\mathrm{P}(\boldsymbol{y}_i = j \mid \boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{y}_i = 0 \mid \boldsymbol{x}_i)}\right)$ lojit dönüşüm (log-odds oranı) olarak adlandırılmaktadır.

Probit regresyon analizi (PRA) lojistik regresyona alternatif olarak kullanılan bir model olup, kategorik verilerin analizlerinde kullanılan doğrusal olmayan modelleme yöntemleri üzerine yapılan çalışmalardan biridir. Lojit modelde olduğu gibi probit modelde de, yanıt değişkeni ikili ya da çoklu, açıklayıcı değişkenler ise kategorik, sıralı veya sürekli durumda olabilmektedir.

Probit regresyonda Φ, standart normal dağılımın kümülatif dağılım fonksiyonunu göstermek üzere, $\Phi^{-1}(p_i)$ den yararlanılmaktadır. $\Phi^{-1}(p_i)$ fonksiyonuna probit regresyon fonksiyonu denir. PRA da yanıt değişkeninin normal dağıldığı varsayımı yapılırken, LRA da bu değişkenin lojistik eğriye dayandığı varsayılmaktadır. Diğer bir ifadeyle, logit model lojistik birikimli dağılım fonksiyonunu kullanırken, probit model normal birikimli dağılım fonksiyonunu kullanmaktadır.

Bu ön bilgiler ışığında uygulama çalışmasında, gözlemlerin gruplara ayrılmasında kullanılan diskriminant analizi bu analize alternatif olan lojistik regresyon analizi ve probit regresyon analizi ile incelenip karşılaştırılmıştır. Gerçek veri seti üzerine ve bir simülasyon çalışması üzerine bu analizler yapılarak LDA varsayımını sağlayan verinin diğer yöntemlere göre daha iyi bir sınıflandırıcı olduğu görülmüştür. Klasik diskriminant varsayımlarını sağlamayan verilerde ise QDA'nin LDA'ya göre daha doğru sınıflandırmalar elde ettiği tespit edilmiştir. Ayrıca diskriminant analizine alternatif olarak önerilen LRA ve PRA nın sonuçlarının çalışılan veriler için neredeyse aynı sonuçları verdikleri ve yakın olasılık tahminlerine sahip oldukları görülmüştür.

TEŞEKKÜR

Bu tezin hazırlanmasında, konu seçiminden içeriğin belirlenmesine kadar her aşamasında bana destek olan, araştırma ve yazım sürecinde yardımlarının yanı sıra sabrını da benden esirgemeyen, öğrencisi olmaktan gurur duyduğum danışmanım sayın Doç. Dr. Gülesen ÜSTÜNDAĞ ŞİRAY'a teşekkürlerimi sunarım. Tez yazımındaki katkılarından dolayı değerli arkadaşım Gökhan ÖZATAŞ'a ayrıca teşekkürlerimi sunarım.

Her zaman yanımda olan, eğitim hayatım boyunca beni teşvik eden, her durumda bana güç veren, beni anlayışla karşılayarak sevgilerini ve sabırlarını benden esirgemeyen, bu çalışma boyunca maddi ve manevi desteklerini her zaman hissettiğim değerli aileme teşekkürü bir borç bilirim.

İÇİNDEKİLER

TABLOLAR DİZİNİ

ŞEKİLLER DİZİNİ

SİMGELER VE KISALTMALAR

EKK : En küçük kareler
EÇO : En çok olabilirlik
LDA : Lineer diskriminant analizi
QDA : Karesel diskriminant analizi
LRA : Lojistik regresyon analizi
PRA : Probit regresyon analizi
LDF : Lineer diskriminant fonksiyonu
QDF : Karesel diskriminant fonksiyonu
odds : Olabilirlik oranları
OR : odds oranı
ln(OR) : odds oranının logaritmik dönüşümü
LL : Log olabilirlik
-2LL : Ölçeklendirilmiş Sapma İstatistiği
D : Deviance istatistiği
MNL : Multinominal lojistik
MNP : Multinominal probit
AIC : Akaike bilgi kriteri
BIC : Bayesian bilgi kriteri
IIA : İlgisiz alternatiflerin bağımsızlığı
PI : Pelvis/leğen kemiği insidansı
PT : Pelvis/leğen kemiği eğimi
LuL : Lordoz/bel kemiği açısı
SS : Sakral/kuyruk sokumu eğimi
PR : Pelvis/leğen kemiği yarıçapı
DS : Spondilolistezi/Kayma derecesi

1. GİRİŞ

1.1. Sınıflandırmaya Genel Bakış

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim (0, \sigma^2 \boldsymbol{I}_n) \tag{1.1}$$

çoklu lineer regresyon modelini göz önüne alalım. Burada; $\boldsymbol{y} : n \times 1$ yanıt değişkenler vektörü, $\boldsymbol{X} : n \times p$ açıklayıcı değişkenlerin matrisi, $\boldsymbol{\beta} : p \times 1$ bilinmeyen regresyon katsayılarının (parametreler) vektörü, $\boldsymbol{\varepsilon} : n \times 1$ hataların vektörü, $E(\boldsymbol{\varepsilon}) = 0$ ve $Var(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_n$ dir.

Çoklu lineer regresyon modelinde, *y* yanıt değişkeninin nicel olduğu varsayılır. Ancak bunun yerine birçok durumda yanıt değişkeni niteldir. Örneğin, mavi, kahverengi ya da yeşil değerlerini alan göz rengi niteldir. Nitel değişkenler genellikle kategorik olarak adlandırılırlar; bu terimler birbirlerinin yerine kullanılmaktadır.

Tahmin edilen nitel yanıtlar için çalışılan yaklaşımlar "sınıflandırma" olarak bilinen işlemlerdir. Bir gözlem için nitel bir yanıt tahmin etmek, bu gözlemi "sınıflandırmak" olarak adlandırılır. Çünkü bu tahmin, gözlemi bir kategoriye ya da gruba atamayı içerir. Diğer bir yandan, genellikle sınıflandırma için kullanılan yöntemler, sınıflandırmanın temeli olarak, öncelikle nitel değişkenin her bir kategorisinin olasılığını tahmin eder. Bu anlamda bu yöntemler regresyon yöntemi gibi de davranmaktadır.

Nitel bir yanıtı tahmin etmek için kullanılan sınıflandırma teknikleri ya da sınıflandırıcılar mevcuttur. Bunların yaygın olarak kullanılanları; Bayes sınıflandırıcısı, lojistik regresyon, probit regresyon, diskriminant analizi ve K-en yakın komşuluklardır.

Sınıflandırma problemleri, belki de regresyon problemlerinden bile daha sık karşımıza çıkmaktadır. Örneğin:

1. Muhtemelen üç tıbbi koşuldan birine dayanan bir dizi belirti ile acil servise gelen bir kişinin, bu üç koşuldan hangisine sahip olduğu belirlenebilmelidir.
2. Online bir banka servisi, kullanıcının IP adresine, geçmiş işlem tarihine, vs, dayanarak site üzerinde gerçekleştirilen bir işlemin hileli olup olmadığını belirleyebilmelidir.
3. Bir biyolog hastalıklı olan ve olmayan hasta sayısı için DNA dizilimi verisine dayanarak, zararlı olan (hastalık yapıcı) veya olmayan DNA mutasyonlarını anlayabilmelidir.

Regresyonda olduğu gibi, sınıflandırma çerçevesinde bir sınıflandırıcı oluşturmak için kullanabilecek bir dizi $(x_1, y_1), \ldots, (x_n, y_n)$ deneysel gözlem mevcuttur. Sadece deneysel veri üzerinde değil, aynı zamanda deneyde kullanılmayan test edilecek gözlemler için de en iyi sınıflandırıcı elde edilmek istenir.

1.2. Bazı Sınıflandırma Yöntemleri

1.2.1. Diskriminant Analizi

Diskriminant analizi, veri kümesindeki değişkenlerin iki veya daha fazla gerçek gruplara ayrılmasına imkan verecek şekilde, birimlerin ya da gözlemlerin p-tane özelliği ele alınarak bu birimlerin gerçek gruplarına optimal düzeyde atanmalarını sağlayacak fonksiyonlar oluşturan bir yöntemdir. Örneğin A,B ve C gibi birbirine yakın özellikleri olan üç hastalık olsun. Her hastalığın *p* tane özelliğe (değişkene) göre bir değeri olacaktır. Diskriminant analizi ile her hastalığın *p* özelliğe (değişkene) göre diğer hastalıklardan farkını ortaya koyan ve grup özelliklerini belirleyen fonksiyonlar oluşturulması mümkündür. Bu fonksiyonlar yardımıyla yeni bir hastalığın gözlem vektörünün hangi grup özelliğini taşıdığı

yani hangi hastalık tanısına sahip olabileceği belirlenebilir ve doğru gruba atanması diskriminant analizi ile yapılabilir.

Diskriminant analizi, gruplandırma modellerinin geliştirilmesinde kullanılan ilk çok değişkenli istatistiksel gruplandırma yöntemidir.

Diskriminant analizi, genel anlamda ayırma işlevine sahip olup, bireylere ait p tane özellikten yararlanarak ait oldukları grupları (kitleleri) belirlemede veya mevcut grupları birbirinden ayıracak en iyi fonksiyonu bulmada kullanılan çok değişkenli istatistik tekniklerinden birisidir (Çamdeviren, 2000).

Diskriminant analizi üzerindeki ilk çalışmalar 1936 yılında Fisher tarafından yapılmıştır. Fisher, iki grubu birbirinden ayırabilmek amacıyla bağımsız değişkenlerin lineer bir bileşimi olan diskriminant fonksiyonlarını oluşturmuştur.

Birimlerin gruplanmasında bazı matematiksel eşitliklerden faydalanılır. Diskriminant fonksiyonu olarak adlandırılan bu eşitlikler birbirine en çok benzeyen grupları belirlemeye olanak sağlayacak şekilde grupların ortak özelliklerini belirlemek amacıyla kullanılmaktadır. Grupları ayırmak amacıyla kullanılan karakteristikler ise diskriminant değişkenleri olarak adlandırılmaktadır. Kısaca diskriminant analizi, iki veya daha fazla sayıdaki grubun farklılıklarının diskriminant değişkenleri vasıtasıyla ortaya konması işlemidir (Klecka, 1980).

Araştırmacı, hatalı sınıflandırma olasılığını en aza indirgeyerek gözlemleri ait oldukları gruplara ayırmak veya bu gözlemlerin çekilmiş oldukları grupları belirlemek isteyecektir (Johnson-Wichem 2002). Burada, belirlenecek grupların ortalamaları arasındaki farklılığın maksimum olması amaçlanır.

Diskriminant analizi yönteminin temel varsayımları, değişkenlerin çok değişkenli normal dağılması ve grupların ortak varyans-kovaryans matrisli olmalarıdır. Bu varsayımları sağlayan diskriminant fonksiyonu lineer diskriminant fonksiyonu (LDF) olarak adlandırılır. Bu varsayımların sağlanmaması durumunda alternatif fonksiyonlar kullanılabilir. Verilerin normal dağıldığı ancak, grupların

varyans-kovaryans matrislerinin farklı olmaları durumunda kullanılan fonksiyon karesel diskriminant fonksiyonu (QDF) olarak tanımlanır.

Diskriminant analizinin işlevleri iki ana başlık altında toplanabilir. Bunlardan ilki, çekildiği kitle bilinmeyen herhangi bir gözlemin (değişkenin) uygun kitleye (gruba) atanmasıdır. İkinci işlevi ise, gelecekte kullanılabilir fonksiyonlar vermesidir ki bu işlev nedeniyle diskriminant analizi kümeleme analizinden farklılaşmakta ve çok değişkenli regresyon analizine yaklaşmaktadır. Diskriminant analizinde küme (grup) sayısı bilinmekte, bu sayı analiz süresince değişmemekte ve araştırmacıdan gözlemleri bu kümelere sınıflandırması istenmektedir. (Tatlıdil, 2002).

1.2.2. Lojistik Regresyon Analizi

Lojistik modelin ilk olarak kullanımı, biyolojik deneylerin analizi için Berkson (1944) tarafından önerilmiştir. Cornfield (1962), lojistik regresyondaki katsayı tahmin işlemlerinde diskriminant fonksiyonu yaklaşımını ilk kez kullanarak popüler hale getirmiştir.

Lojistik modelini Cox (1970) geliştirmiş, çeşitli uygulamalarını yapmış, Halperin ve ark. (1971) ise lojistik regresyonun, bağımsız değişkenlere ait normal dağılım varsayımının yerine gelmediği durumlarda diskriminant analizine alternatif olarak gösterilebileceğini savunmuşlardır. Finney de (1971) lojistik regresyonu probit analizine alternatif olarak önermiştir.

Lojistik regresyon analizi, yanıt değişkenin ikili ya da çoklu, açıklayıcı değişkenlerin ise kategorik, sıralı veya sürekli olması durumunda belirli bir dağılım varsayımına bağlı kalmaksızın yanıt değişken ile açıklayıcı değişkenler arasındaki neden-sonuç ilişkisinin belirlenmesinde ve aynı zamanda açıklayıcı değişkenlerin etkilerine dayanarak verilerin sınıflandırılmasında kullanılan bir yöntemdir.

Lojistik regresyonun bir diğer özelliği de, elde edilen sonuçlarda değerlerin sadece hangi grupta olduklarını değil, aynı zamanda o grupta bulunma olasılıklarını da belirlemesidir.

Lojistik regresyon analizi, çeşitli varsayım bozulmaları (normal dağılıma, ortak kovaryansa sahip olmama gibi) durumunda diskriminant analizi ve çapraz sınıflandırma tablolarına alternatiftir. Ayrıca, yanıt değişkenin 0 ve 1 gibi ikili (dichotomous) ya da ikiden çok düzey içeren kesikli değişken (polychotomous) olması durumunda, normallik varsayımının veya diğer varsayımların bozulması nedeniyle doğrusal regresyon analizine alternatif olmaktadır (Elasan, 2010).

Açıklayıcı değişkenlerin bazılarının sürekli, bazılarının kesikli olması durumlarında diskriminant analizine alternatif olarak lojistik regresyon analizi önerilir (Tatlıdil, 2002).

Lojistik regresyon analizinin amacı, istatistikte kullanılan diğer model yapılandırma yöntemlerine benzer şekilde; en az değişkeni kullanarak, yanıt değişkeni ile açıklayıcı değişkenler arasındaki ilişkiyi tanımlayabilen en iyi uyuma sahip ve yorumlanabilen bir model bulmaktır.

1.2.3. Probit Regresyon Analizi

Probit regresyon analizi, lojistik regresyona alternatif olarak kullanılan bir analiz olup, bir veya daha fazla açıklayıcı değişkenin kategorik bir yanıt değişkeni (sağ, ölü; çalışıyor, çalışmıyor vb.) üzerindeki etkisini bulmak için kullanılır. Her iki analiz birbirlerine oldukça benzer ve her iki yöntem ile elde edilen olasılık tahminleri birbirlerine yakın değerdedir. Matematiksel uygulaması ve yorumlama kolaylığından dolayı lojistik model daha çok tercih edilmektedir. Ancak paket programlarının yaygınlaşması ile probit modelinde bahsedilen zorluğu aşılmaktadır. Kullanılacak olan paket programa ve araştırmacının tercihine göre iki model arasında tercih yapılabilir.

Lojistik regresyon analizinde log odds (olabilirlik oranları) kullanılırken, probit regresyon analizinde kümülatif normal dağılım kullanılmaktadır(Bek,2009). Probit regresyonda Φ, standart normal dağılımın kümülatif dağılım fonksiyonunu göstermek üzere $\Phi^{-1}(\pi)$ den yararlanılmaktadır. $\Phi^{-1}(\pi)$ fonksiyonuna probit denir.

1.2.4. Sınıflandırma Problemi İçin Uygulanan Bayes Kuralı

A ve *B* herhangi iki olay olsun. *A*'nın meydana geldiği verilmişken *B*'nin meydana gelme olasılığı, $P(B \mid A)$ olasılığını bulmak için, Bayes Kuralı aşağıdaki gibi ifade edilir:

$$P(B \mid A) = \frac{P(A \text{ ve } B)}{P(A)} \tag{1.2}$$

Bu formül; koşullu olasılığın, *A* ve *B*'nin ikisinin gerçekleşme olasılığının *A*'nın koşulsuz olasılığına bölünerek elde edildiğini göstermektedir. Bu her iki olayın birlikte meydana gelme olasılığını bulmak için olan, $P(A \text{ ve } B) = P(A)P(B \mid A)$ kuralının basitçe yeniden düzenlenen cebirsel bir ifadesidir.

Bir gözlem, $k \geq 2$ iken *k* grubun birine sınıflandırılmak istensin. Diğer bir deyişle, $\boldsymbol{y}$ nitel yanıt değişkeninin *k* olası farklı ve sırasız değer alabildiği varsayılsın. π_k, *k*_ıncı gruptan gelen rasgele seçilen bir gözlemin genel ya da önsel olasılığını göstersin; bu $\boldsymbol{y}$ yanıt değişkenin *k*_ıncı grubu ile ilişkili olarak verilen bir gözlemin olasılığıdır. $f_k(\mathrm{x}) = \mathrm{P}(\boldsymbol{X} = \mathrm{x} \mid \boldsymbol{y} = k)$, *k*_ıncı gruptan gelen bir gözlem için x'in yoğunluk fonksiyonunu göstersin. *k*_ıncı gruptaki bir gözlemin $\boldsymbol{X} \approx \mathrm{x}$'e sahip olma olasılığı yüksek ise, $f_k(\mathrm{x})$ nispeten büyüktür.

Ayrıca *k*_ıncı gruptaki bir gözlemin $\boldsymbol{X} \approx \mathrm{x}$ 'e sahip olma olasılığı çok düşük ise, $f_k(\mathrm{x})$ küçüktür.

Yukarıdaki Bayes kuralı notasyonunu kullanarak, *A* olayı= gözlenen x ve *B* olayı= $\boldsymbol{y}$ yanıt değişkenin *k*_ıncı gruptan gelen gözlem olarak alınırsa, ilgilenilen olasılık

$$P(\boldsymbol{y} = k \text{ 'nın üyesi} \mid \text{gözlenen x}) = \frac{P(\boldsymbol{y} = k \text{ 'nın üyesi ve gözlenen x})}{P(\text{gözlenen x})}$$

olarak bulunabilir. Bu durumda Bayes teoremi

$$\mathrm{P}(\boldsymbol{y} = k \mid \boldsymbol{X} = \mathrm{x}) = \frac{\pi_k f_k(\mathrm{x})}{\sum_{l=1}^{k} \pi_l f_l(\mathrm{x})} \tag{1.3}$$

olarak ifade edilir. $p_k(\mathrm{x}) = \mathrm{P}(\boldsymbol{y} = k \mid \boldsymbol{X} = \mathrm{x})$ kısaltmasını kullanarak, bu ifade doğrudan $p_k(\mathrm{x})$ hesaplamak yerine, kolaylıkla (1.3)'deki π_k ve $f_k(\mathrm{x})$'in tahminlerinin kullanılabileceğini önermektedir. Genel olarak, eğer kitleden alınan $\boldsymbol{y}$'lerin rasgele bir örneklemi mevcutsa, π_k'yı tahmin etmek kolaydır: *k*_ıncı gruba ait olan deneysel gözlemlerin bir kısmı kolaylıkla hesaplanabilir. Ancak $f_k(\mathrm{x})$'i tahmin etmek, bu yoğunluklar için bazı basit formlar varsayılmazsa daha zorlayıcı olur. $p_k(\mathrm{x})$ *k*_ıncı gruba ait bir $\boldsymbol{X} = \mathrm{x}$ gözleminin sonsal olasılığı olarak adlandırılır. Yani $p_k(\mathrm{x})$ *k*_ıncı gruba ait gözlemin, bu gözlemin tahmin değeri verilmişken olasılığıdır.

Bir gözlemi $p_k(\mathrm{x})$'i en büyük olan gruba sınıflandıran Bayes sınıflandırıcısı, tüm sınıflandırıcılar arasında mümkün olan en düşük hata oranına sahiptir. Bu elbette ki sadece (1.3)'deki terimlerin hepsi doğru olarak belirlenmişse

doğrudur. Bu nedenle, $f_k(\mathrm{x})$'i tahmin etmek için bir yol bulunursa, o zaman Bayes sınıflandırıcısına yaklaşan bir sınıflandırıcı geliştirilebilir.

1.2.5. K-En Yakın Komşuluk Kuralı

En eski parametrik olmayan sınıflandırma yöntemi, aynı zamanda *k*-en yakın komşu kuralı olarak ta bilinen, Fix ve Hodges'in (1951) en yakın komşu kuralıdır. Yöntem kavramsal olarak basittir. Uzaklık fonksiyonu

$$(\boldsymbol{x}_{\mathrm{i}} - \boldsymbol{x}_{\mathrm{j}})' \boldsymbol{S}^{-1} (\boldsymbol{x}_{\mathrm{i}} - \boldsymbol{x}_{\mathrm{j}}), \qquad j \neq i$$

kullanılarak bir $\boldsymbol{x}_{\mathrm{i}}$ gözlem vektöründen diğer tüm $\boldsymbol{x}_{\mathrm{j}}$ noktalarına olan uzaklık hesaplanır. Burada $\boldsymbol{S}$ ortak örneklem kovaryans matrisidir. $(\boldsymbol{x}_{\mathrm{i}} - \boldsymbol{x}_{\mathrm{j}})' \boldsymbol{S}^{-1} (\boldsymbol{x}_{\mathrm{i}} - \boldsymbol{x}_{\mathrm{j}})$ karesel uzaklık fonksiyonuna Mahalanobis uzaklığı denir (Mahalanobis, 1936). $\boldsymbol{x}_{\mathrm{i}}$'yi iki gruptan (grup 1: G_1 ve grup 2: G_2) birine sınıflandırmak için, $\boldsymbol{x}_{\mathrm{i}}$'ye en yakın *k* noktaları incelenmektedir ve *k* noktalarının çoğunluğu G_1' e ait ise $\boldsymbol{x}_{\mathrm{i}}$ G_1' e atanır; diğer durumda $\boldsymbol{x}_{\mathrm{i}}$ G_2' e atanır. $k = k_1 + k_2$ iken, k_1 nokta ile G_1' deki noktaların sayısı kalan k_2 nokta ile G_2' deki noktaların sayısı ifade edilirse, bu durumda kural şu şekilde ifade edilebilmektedir:

$$k_1 > k_2, \tag{1.4}$$

ise $\boldsymbol{x}_{\mathrm{i}}$, G_1'e ve diğer durumda G_2' e atanır. n_1 ve n_2 örneklem büyüklükleri farklı ise sayımlar yerine oranları kullanmak istenebilir:

$$\frac{k_1}{n_1} > \frac{k_2}{n_2}, \tag{1.5}$$

ise $\boldsymbol{x}_i$, G_1'e atanır. Bir başka iyileştirme önsel olasılıklar (π_i) göz önünde bulundurularak yapılabilmektedir:

$$\frac{k_1 / n_1}{k_2 / n_2} > \frac{\pi_2}{\pi_1}, \tag{1.6}$$

ise $\boldsymbol{x}_i$, G_1'e atanır. Bu kurallar kolaylıkla ikiden fazla gruba genişletilebilmektedir. Örneğin, (1.5) şu şekilde olur: gözlem en yüksek k_i / n_i oranına sahip olan gruba atanır, burada k_i sorudaki gözlemin en yakın k komşu arasında G_i'den gelen gözlem sayısıdır.

Karar *k* değerine göre uygulanmalıdır. Loftsgaarden ve Quesenberry (1965) genel bir n_i için $\sqrt{n_i}$'ye yakın k seçimi önermişlerdir. Uygulamada, birden fazla k değeri denenebilir ve en iyi hata oranlı olan kullanılabilmektedir.

1.3. Neden Lineer Regresyon Değil?

Nitel yanıt olması durumunda lineer regresyon uygun olmayacaktır. Örneğin, acil servisteki bir kadın hastanın belirtilerine dayanarak sağlık durumu tahmin etmek istensin. Bu basitleştirilmiş örnekte, 3 mümkün teşhis vardır: felç, aşırı dozda ilaç ve epilepsi nöbeti. Bu üç değişkeni kodlayarak, $\boldsymbol{y}$ nitel yanıt değişkeni aşağıdaki gibi nicel bir yanıt değişkeni olarak düşünülebilir:

$$y = \begin{cases} 1, & \text{felç ise} \\ 2, & \text{aşırı dozda ilaç ise} \\ 3, & \text{epilepsi nöbeti} \end{cases}$$

Bu kodlamayı kullanarak, $x_1,...x_p$ tahmin edicilerine dayanan $\boldsymbol{y}$ 'yi tahmin eden bir lineer regresyon modeli uydurmak için en küçük kareler (EKK) kullanılabilir. Maalesef ki bu kodlama, felç ve epilepsi nöbeti arasındaki aşırı doz ilacı açıklayarak ve felç ve aşırı dozda ilaç arasındaki farkın aşırı dozda ilaç ve epilepsi nöbeti arasındaki fark ile aynı olduğunu öne sürerek, sonuçlar üzerinde bir sıralama anlamına gelir. Uygulamada bu durumda olması gereken belirli bir neden yoktur. Örneğin, üç durum arasında tamamen farklı bir ilişki anlamına gelen, eşit derecede uygun bir kodlama seçilebilirdi:

$$y = \begin{cases} 1, & \text{epilepsi nöbeti ise} \\ 2, & \text{felç ise} \\ 3, & \text{aşırı dozda ilaç ise} \end{cases}$$

Bu kodlamaların her biri, sonuçta deney gözlemleri üzerinde farklı ön tahmin kümelerine neden olan, temelde farklı lineer modeller gösterirdi.

Yanıt değişkenlerinin değerleri doğal bir sıralamada alınsaydı, örneğin hafif, orta ve şiddetli gibi ayrıca hafif ve orta arasında hissedilen farklılık orta ve şiddetli arasındaki farklılık benzer olsaydı, bu durumda 1,2,3 kodlaması uygun olabilirdi. Ne yazık ki, genellikle ikiden daha fazla seviyeli nitel bir yanıt değişkeni, lineer regresyon için kolay olan nicel bir yanıta dönüştürme yolu doğal olmaz.

Bu durum iki değişkenli (iki seviyeli) nitel bir yanıt için daha iyidir. Örneğin, hastaların sağlık durumu için sadece iki olasılık olabilir: felç ve aşırı dozda ilaç. O zaman yanıt değişkenini aşağıdaki gibi kodlamak için, potansiyel olarak kukla (dummy) değişken kullanılır:

$$y = \begin{cases} 0, & \text{felç ise} \\ 1, & \text{aşırı dozda ilaç} \end{cases}$$

Bu durumda bu iki yanıta göre lineer regresyon uydurulabilir ve $\hat{\boldsymbol{y}} > 0.5$ ise aşırı dozda ilaç ve diğer durumda felçlik tahmin edilebilir. İki değişkenli durumda yukarıdaki kodlamayı tersine döndürsek bile aynı final tahminlerini oluşturacak olan lineer regresyonun gösterilmesi zor değildir.

Yukarıdaki gibi 0/1 kodlamalı ikili yanıt değişken için, EKK ile regresyon uygun olabilir: lineer regresyon kullanarak elde edilen $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ 'nın, bu özel durumdaki $p(\text{aşırı dozda ilaç}|x)$'nın gerçek bir tahmini olduğu gösterilebilir. Ancak, lineer regresyon kullanıyorsak tahminlerin bazıları [0,1] aralığının dışında olabilir, dolayısıyla olasılık olarak yorumlanmaları zor olacaktır. Yine de, tahminler bir sıralama sağlar ve kaba olasılık tahminleri gibi yorumlanabilirler. İlginç biçimde, ikili yanıtı tahmin etmek için lineer regresyon kullanırsak elde ettiğimiz sınıflandırmalar, lineer diskriminant analizi (LDA) yöntemi ile aynı olacaktır.

Kukla değişkeni yaklaşımı ikiden fazla seviyeli yanıt değişkeni karşılamak için kolaylıkla genişletilemez. Bu nedenlere göre, nitel yanıt değerler için tam olarak uyan bir sınıflandırma yöntemi kullanmak daha uygundur (James, Witten, Hastie ve Tibshirani, 2015).

2. DİSKRİMİNANT ANALİZİ

2.1. Grup Ayrıştırmasının Tanımı

Bir kitle ya da kitleden alınmış bir örneklemi temsil etmek için grup terimi kullanılır. Grupların ayrılmasında iki temel amaç vardır:

Grup ayrıştırmasının tanımı, değişkenlerin doğrusal fonksiyonlarındaki (diskriminant fonksiyonları) iki ya da daha fazla grup arasındaki farkları tanımlamak ve açıklamak için kullanılmaktadır. Tanımlayıcı diskriminant analizi amaçları, grupların ayrılmasıyla *p* değişkenin göreceli katkısını belirlemeyi ve grupların yapısını en iyi şekilde belirlemek için öngörülebilir noktalar üzerinde optimal düzlemi bulmayı içerir.

Gözlemlerin, değişkenlerin doğrusal ya da karesel fonksiyonlarındaki (sınıflandırma fonksiyonu) gruplara tahmini ya da atanması, grupların birine bir birey örneklem birimi atamak için kullanılır. Bir birey ya da nesne için gözlem vektöründeki ölçülmüş değerler, bireyin büyük olasılıkla ait olduğu grubu bulmak için sınıflandırma fonksiyonu tarafından hesaplanmaktadır.

Diskriminant fonksiyonları, grupları en iyi şekilde ayıran değişkenlerin doğrusal kombinasyonlarıdır. Bu fonksiyonlar iki grup için ve ikiden fazla grup için ayrı olarak incelenir.

2.2. İki Grup İçin Diskriminant Fonksiyonu

Karşılaştırılacak iki kitlenin aynı $\boldsymbol{\Sigma}$ kovaryans matrisine fakat $\boldsymbol{\mu}_1$ ve $\boldsymbol{\mu}_2$ farklı ortalama vektörlerine sahip olduğu varsayılsın. İki kitleden alınan $\boldsymbol{x}_{11}, \boldsymbol{x}_{12}, \ldots, \boldsymbol{x}_{1n_1}$ ve $\boldsymbol{x}_{21}, \boldsymbol{x}_{22}, \ldots, \boldsymbol{x}_{2n_2}$ örneklemleri ile çalışılır. Genel olarak her bir $\boldsymbol{x}_{ij}$ vektörü p-tane değişken üzerindeki ölçümleri oluşturur. Diskriminant fonksiyonu iki grup ortalama vektörleri arasındaki mesafeyi maksimum yapan bu

p-tane değişkenin lineer kombinasyonudur. Bu lineer kombinasyon $\boldsymbol{y} = \boldsymbol{b}'\boldsymbol{X}$, her bir gözlem vektörünü bir skalere dönüştürür.

$$y_{1i} = \boldsymbol{b}'\boldsymbol{x}_{1i} = b_1 x_{1i1} + b_2 x_{1i2} + ... + b_p x_{1ip}, \qquad i = 1, 2, ..., n_1$$

$$y_{2i} = \boldsymbol{b}'\boldsymbol{x}_{2i} = b_1 x_{2i1} + b_2 x_{2i2} + ... + b_p x_{2ip}, \qquad i = 1, 2, ..., n_2$$

Dolayısıyla iki örneklemdeki $n_1 + n_2$ tane

$$\begin{matrix} \boldsymbol{x}_{11} & \boldsymbol{x}_{21} \\ \boldsymbol{x}_{12} & \boldsymbol{x}_{22} \\ \vdots & \vdots \\ \boldsymbol{x}_{1n_1} & \boldsymbol{x}_{2n_2} \end{matrix}$$

gözlem vektörleri,

$$\begin{matrix} y_{11} & y_{21} \\ y_{12} & y_{22} \\ \vdots & \vdots \\ y_{1n_1} & y_{2n_2} \end{matrix}$$

skalerlerine dönüştürülür.

$$\bar{\boldsymbol{x}}_1 = \sum_{i=1}^{n_1} \boldsymbol{x}_{1i} / n_1 \qquad \text{ve} \qquad \bar{\boldsymbol{x}}_2 = \sum_{i=1}^{n_2} \boldsymbol{x}_{2i} / n_2$$

iken

$$\bar{y}_1 = \sum_{i=1}^{n_1} y_{1i} / n_1 = \boldsymbol{b}'\bar{\boldsymbol{x}}_1 \qquad \text{ve} \qquad \bar{y}_2 = \boldsymbol{b}'\bar{\boldsymbol{x}}_2$$

ortalamaları bulunur. Daha sonra $(\bar{y}_1 - \bar{y}_2)/s_y$ standartlaştırılmış farkını maksimize eden bir $\boldsymbol{b}$ vektörü bulunmak istenir.

$(\bar{y}_1 - \bar{y}_2)/s_y$ ifadesi negatif olabileceği için $(\bar{y}_1 - \bar{y}_2)^2 / s_y^2$ karesel uzaklığı kullanılır. $\boldsymbol{S}_1$ ve $\boldsymbol{S}_2$ grupların kovaryans matrisleri, $\boldsymbol{S}$ ortak örneklem kovaryans matrisi $\boldsymbol{S} = \dfrac{(n_1-1)\boldsymbol{S}_1 + (n_2-1)\boldsymbol{S}_2}{n_1+n_2-2}$ iken, $\boldsymbol{y}_i$'nin örneklem varyansı $s_y^2 = \dfrac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \boldsymbol{b}'\boldsymbol{S}\boldsymbol{b}$ kullanılarak bu uzaklık aşağıdaki gibi ifade edilir:

$$\frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2} = \frac{[\boldsymbol{b}'(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)]^2}{\boldsymbol{b}'\boldsymbol{S}\boldsymbol{b}}. \tag{2.1}$$

(2.1)'in maksimumu

$$\boldsymbol{b} = \boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) \tag{2.2}$$

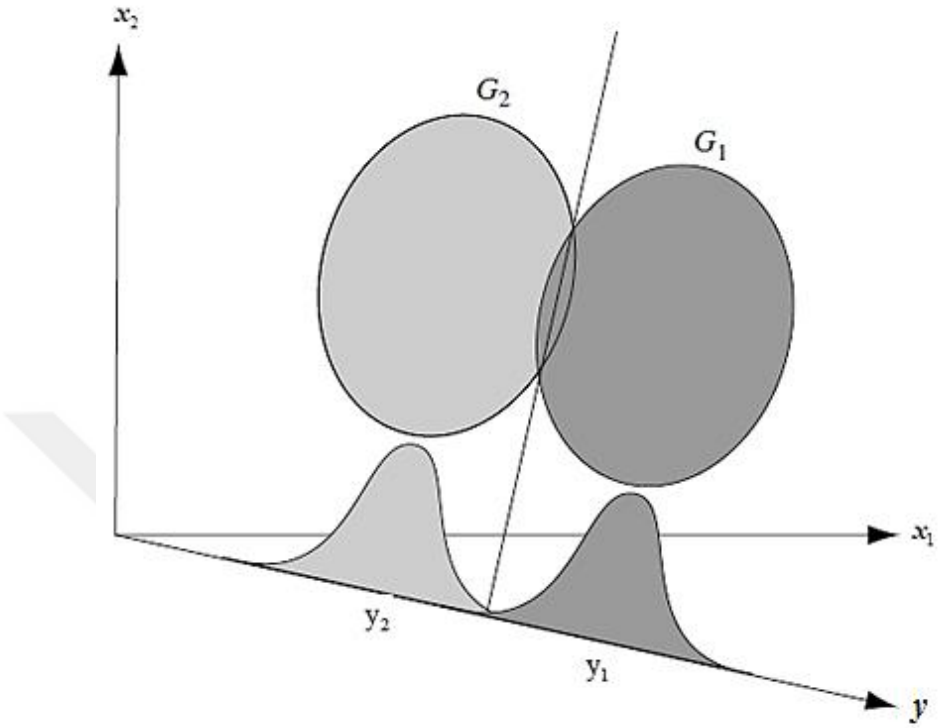
olduğunda ya da $\boldsymbol{b}$ vektörü $\boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$'in herhangi bir katı olduğunda ortaya çıkar. Bu nedenle maksimize edilen $\boldsymbol{b}$ vektörü tek değildir. Ancak yönü tektir, diğer bir deyişle $b_1, b_2, \ldots, b_p$ bağıl değerleri ya da oranları tektir ve $\boldsymbol{y} = \boldsymbol{b}'\boldsymbol{X}$, $(\bar{y}_1 - \bar{y}_2)^2 / s_y^2$'yi maksimize eden doğru üzerinde $\boldsymbol{X}$ noktaları üzerine izdüşer. Dikkat edilmelidir ki $\boldsymbol{S}^{-1}$'nin var olması için, $n_1 + n_2 - 2 > p$ olmalıdır.

$\boldsymbol{b} = \boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$ ile verilen optimum yön, $(\bar{y}_1 - \bar{y}_2)^2 / s_y^2$ karesel uzaklığı $\bar{\boldsymbol{x}}_1$ ve $\bar{\boldsymbol{x}}_2$ arasındaki standartlaştırılmış uzaklığa eşit olduğundan, $\bar{\boldsymbol{x}}_1$ ve $\bar{\boldsymbol{x}}_2$ 'nın birleştiği çizgiye etkili bir şekilde paraleldir. Bu durum (2.2), (2.1) yerine yazılarak görülebilir. $\boldsymbol{y} = \boldsymbol{b}'\boldsymbol{X}$ için

$$\frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2} = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) \tag{2.3}$$

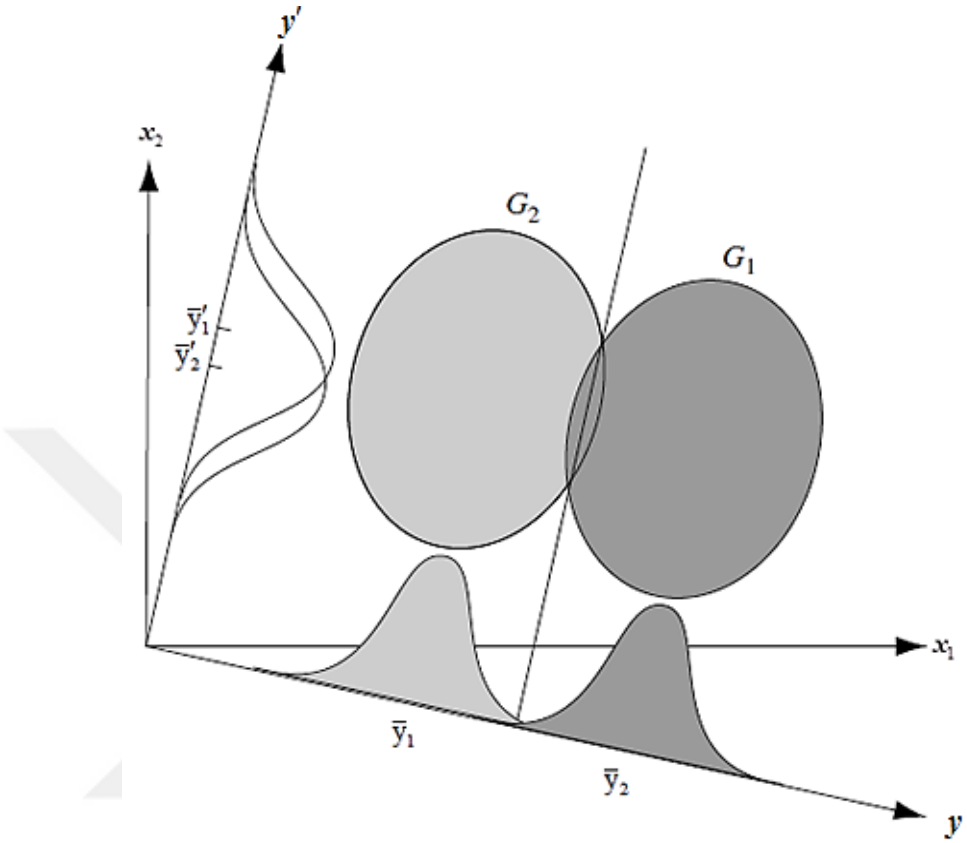
dir, burada $\boldsymbol{b} = \boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$ dir. $(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$ karesel uzaklığı Mahalanobis uzaklığı olarak adlandırılır. $\boldsymbol{b}' = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\boldsymbol{S}^{-1}$ olduğundan, (2.3) eşitliği $(\bar{y}_1 - \bar{y}_2)^2 / s_y^2 = \boldsymbol{b}'(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$ olarak yazılabilir ve $\boldsymbol{b} = \boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$ ile temsil edilenden başka herhangi bir yön, $\boldsymbol{b}'\bar{\boldsymbol{x}}_1$ ve $\boldsymbol{b}'\bar{\boldsymbol{x}}_2$ arasında daha küçük bir fark verecektir.

Şekil 2.1, (2.2)' de verilen $\boldsymbol{b}$ vektörü için $\boldsymbol{y} = \boldsymbol{b}'\boldsymbol{X}$ diskriminant fonksiyonu tarafından tek boyutla temsil edilen iki değişkenli normal (p=2) iki grubun ayrılmasını göstermektedir. Bu şekildeki kitle kovaryans matrisleri eşittir. $y_{1i} = \boldsymbol{b}'\boldsymbol{x}_{1i} = b_1 x_{1i1} + b_2 x_{1i2}$ ve $y_{2i} = \boldsymbol{b}'\boldsymbol{x}_{2i} = b_1 x_{2i1} + b_2 x_{2i2}$ doğrusal kombinasyonları, iki grubun optimum ayrışma doğrusu üzerinde $\boldsymbol{x}_{1i}$ ve $\boldsymbol{x}_{2i}$ noktalarını çizer (yansıtır). $\boldsymbol{x}_1$ ve $\boldsymbol{x}_2$ değişkenleri iki değişkenli normal oldukları için, $\boldsymbol{y} = b_1\boldsymbol{x}_1 + b_2\boldsymbol{x}_2 = \boldsymbol{b}'\boldsymbol{X}$ doğrusal kombinasyonu tek değişkenli normaldir. Bu nedenle bu fonksiyon $\boldsymbol{y}$ ile temsil edilen doğru boyunca iki tek değişkenli normal yoğunluklar ile gösterilmektedir.



Şekil 2.1. İki grup diskriminant analizi (Rencher, 2002)

Diskriminant fonksiyonu doğrusu $\boldsymbol{y}$ de kesişen iki elipsin kesişim noktalarını birleştiren doğrudaki nokta, doğru üzerinde gösterilen noktaların maksimum ayrıştırma (minimum çakışma) noktasıdır. Şekil 2.1'de gösterildiği gibi iki kitle $\boldsymbol{\Sigma}$ ortak kovaryans matrisli çok değişkenli normal ise, tüm olası grup ayrıştırmalarının yeni bir tek boyutta ifade edilebileceği gösterilebilir.



Şekil 2.2. Diskriminant fonksiyonu ile elde dilen ayrıştırma (Rencher, 2002)

Şekil 2.2'de, diskriminant fonksiyonu ile elde edilen en iyi ayrıştırma gösterilmektedir. $\boldsymbol{y}'$ ile ifade edilen diğer yönlerdeki gösterim, dönüştürülmüş $\bar{y}'_1$ ve $\bar{y}'_2$ ortalamaları arasında daha küçük bir standartlaştırılmış uzaklık ve ayrıca gösterilen noktalar arasında daha büyük bir çakışma verir.

2.2.1. İki Grup Diskriminant Analizi ve Çoklu Regresyon Arasındaki İlişki

İki grup diskriminant analizi ve çoklu regresyon analizi arasındaki bağlantı ilk kez Fisher (1936) tarafından belirtilmiştir. Flury and Riedwyl (1985) bu ilişki hakkında daha fazla görüş vermişlerdir.

Çoklu regresyon ve iki grup için diskriminant analizi arasındaki ortak bağlantı Hotelling T^2 istatistiği olarak bilinen T^2'nin hesaplaması açısından ortaya konulur. İki örneklem *p* değişken için genelleştirilmiş T^2

$$T^2 = [n_1 n_2 / (n_1 + n_2)](\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)' \boldsymbol{S}^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$$

şeklinde hesaplanır. İki örneklem T^2'de her bir gözlem vektörü $\boldsymbol{x}_{1i}$ ve $\boldsymbol{x}_{2i}$ için

Grup 1'deki her bir $x_{11}, x_{12}, ..., x_{1n_1}$ için $y_i = \dfrac{n_2}{n_1 + n_2}$

Grup 2'deki her bir $x_{21}, x_{22}, ..., x_{2n_2}$ için $y_i = -\dfrac{n_2}{n_1 + n_2}$

olacak şekilde bir kukla değişken tanımlanır. Bu durumda $\boldsymbol{y}$, tüm $n_1 + n_2$ gözlem için $\bar{y} = 0$ olacak şekilde, grup 1 ve 2'yi belirleyen bir gruplama değişkeni olur. Ayrıca $\boldsymbol{y}$ değişkeni x'ler için uydurulduğunda (fitted) tahmin edilen regresyon denklemi

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \boldsymbol{x}_{i1} + \hat{\beta}_2 \boldsymbol{x}_{i2} + ... + \hat{\beta}_p \boldsymbol{x}_{ip} \qquad (2.4)$$

olarak yazılır. Burada *i* tüm $n_1 + n_2$ gözlem üzerinde değişir ve β_0'ın EKK tahmini

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{\boldsymbol{x}}_1 + \hat{\beta}_2 \bar{\boldsymbol{x}}_2 + ... + \hat{\beta}_p \bar{\boldsymbol{x}}_p$$

şeklindedir. Bu tahmin regresyon denkleminde yerine konularak,

$$\hat{y}_i = \bar{y} + \hat{\beta}_1(x_{i1} - \bar{x}_1) + \hat{\beta}_2(x_{i2} - \bar{x}_2) + \ldots + \hat{\beta}_p(x_{ip} - \bar{x}_p)$$
$$= \hat{\beta}_1(x_{i1} - \bar{x}_1) + \hat{\beta}_2(x_{i2} - \bar{x}_2) + \ldots + \hat{\beta}_p(x_{ip} - \bar{x}_p) \quad (\bar{y} = 0) \tag{2.5}$$

elde edilir. (2.5) eşitliğinin katsayılar vektörü $\boldsymbol{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)'$, diskriminant fonksiyonu katsayılar vektörü $\boldsymbol{b} = \boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$ ile doğru orantılıdır:

$$\boldsymbol{\beta} = \frac{n_1 n_2}{(n_1 + n_2)(n_1 + n_2 - 2 + T^2)} \boldsymbol{b}. \tag{2.6}$$

R^2 çoklu korelasyon katsayısı olmak üzere, T^2 ile R^2 arasındaki ilişkisi

$$T^2 = (n_1 + n_2 - 2)\frac{R^2}{1 - R^2} \tag{2.7}$$

olarak hesaplanır. Buradan R^2'nin $\boldsymbol{\beta}$ ve T^2 ile ilişkisi

$$R^2 = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\boldsymbol{\beta} = \frac{T^2}{n_1 + n_2 - 2 + T^2} \tag{2.8}$$

eşitliği ile elde edilir.

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_p x_{ip}$ çoklu regresyon denkleminde p tane değişkenden en az birinin modele katkısı olup olmadığını test etmek için

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$$

hipotezi test edilir. H_0 hipotezi regresyon kareler ortalamasının hata kareler ortalamasına oranı olarak ifade edilen

$$F = \frac{MS_R}{MS_{\text{Res}}} = \frac{n-p-1}{p}\frac{R^2}{1-R^2} \sim F_{\alpha,p,n-p-1}$$

istatistiği kullanılarak elde edilebilir. Bu mantıkla T^2 ile R^2 arasındaki ilişki kullanılarak, grupları ayırmak için $p+q$ değişkenin q tanesinin anlama katkısı olmadığı hipotezleri için regresyona göre test istatistiği

$$F = \frac{n_1+n_2-p-q-1}{q}\frac{R^2_{p+q}-R^2_p}{1-R^2_{p+q}}$$

ile elde edilebilir, burada R^2_{p+q} ve R^2_p sırasıyla $p+q$ ve p değişkenli regresyonlardan elde edilen çoklu korelasyon katsayılarıdır.

2.3. İkiden Fazla Grup İçin Diskriminant Fonksiyonları

2.3.1. Diskriminant Fonksiyonları

İkiden fazla grup olması durumunda kullanılan diskriminant analizi teknikleri, iki grup için geliştirilenlerin genelleştirilmiş biçimleridir. Bu durumda da yine p değişkenli, ikiden fazla sonlu sayıda kitle bulunmaktadır. Gözlemler, kitleler arasında ayırma gücü en büyük olacak şekilde belli sayıda doğrusal bağlantılar yardımıyla sınıflandırılmaktadır. Gözlemlere ait p değişkeni, p boyutlu uzayda tanımlayacak öyle eksenler bulunmalı ki veri toplulukları bu eksenler boyunca olabildiğince ayrılabilsin. İkiden fazla grup için diskriminant analizinde, çok değişkenli gözlemleri en iyi şekilde k gruba ayıran değişkenlerin doğrusal

kombinasyonlarını bulmakla ilgilenilir. İkiden fazla grup için diskriminant analizi, çeşitli amaçlara hizmet eder:

i) İki boyutlu bir grafikte grup ayrımı incelenir. İkiden fazla grup olması durumunda, grup ayrımını tanımlamak birden daha fazla diskriminant fonksiyonu gerektirir. *p*-boyutlu uzaydaki noktalar ilk olarak iki diskriminant fonksiyonu ile gösterilen iki-boyutlu uzay üzerine yansıtılırsa, grupların nasıl ayrılacağının mümkün olan en iyi görünümü elde edilir.
ii) Grupları neredeyse orjinal küme kadar iyi ayıran, orijinal değişkenlerin bir kümesi bulunur.
iii) Grup ayrışımına bağıl katkıları açısından değişkenler sıralanır.
iv) Diskriminant fonksiyonları ile ifade edilen yeni boyutlar yorumlanır.
v) Sabit-etkiler çok değişkenli varyans analizi (MANOVA) takip edilir.

iii. ve iv. amaçlar yakından ilişkilidir. İlk dört amaçtan herhangi biri, v. amacı gerçekleştirmek için kullanılabilir. Çoklu iç ilişki ve aykırı gözlemlerin varlığında yararlı olabilen diskriminant fonksiyonlarının alternatif tahmin edicileri de mevcuttur (Rencher, 1998).

İki grup olması durumunda tek bir ayırıcı fonksiyon grupları birbirinden ayırırken, çok grup olması durumunda tek ayırıcı fonksiyon tüm grupları ayırmada yeterli olamamaktadır. Bu nedenle ikinci, hatta üçüncü ayırıcı fonksiyonlara ihtiyaç duyulmaktadır. Bulunacak ayırıcı fonksiyonlardan ilki, gruplar arasındaki en büyük ayırımı sağlayan olacaktır. İkinci fonksiyon, ilki ile ilişkisi olmayan ve ilk ayırıcı fonksiyondan sonra gruplar arasında en iyi ayırımı sağlayan bağlantı olacaktır. Diğer fonksiyonlar da benzer biçimde yorumlanacaktır (Tatlıdil, 2002).

*i*_inci gruptaki n_i gözlemli *k* tane grup (örneklem) için, her bir gözlem vektörü $\boldsymbol{x}_{ij}$ kullanılarak $y_{ij} = \boldsymbol{b}'\boldsymbol{x}_{ij}$, $i = 1, 2, ..., k$, $j = 1, 2, ..., n_i$ elde edilir ve

$\bar{y}_i = \boldsymbol{b}'\bar{\boldsymbol{x}}_i$ ortalamaları bulunur, burada $\bar{\boldsymbol{x}}_i = \sum_{j=1}^{n_i} \boldsymbol{x}_{ij} / n_i$ 'dir. İki grup durumunda olduğu gibi, $\bar{y}_1, \bar{y}_2, ..., \bar{y}_k$'ları maksimum düzeyde ayıran $\boldsymbol{b}$ vektörü aranır. $\bar{y}_1, \bar{y}_2, ..., \bar{y}_k$ arasındaki ayrılmayı açıklamak için, (2.1) ayırma kriteri *k*-grup durumuna genelleştirilir: $\boldsymbol{b}'(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\boldsymbol{b}$ olduğundan (2.1)

$$\frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2} = \frac{[\boldsymbol{b}'(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)]^2}{\boldsymbol{b}'\boldsymbol{S}\boldsymbol{b}} = \frac{\boldsymbol{b}'(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\boldsymbol{b}}{\boldsymbol{b}'\boldsymbol{S}\boldsymbol{b}} \tag{2.9}$$

formunda ifade edilir.

(2.9) eşitliğini *k* gruba genişletmek için, $(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'$'nin yerine MANOVA'dan gelen gruplar arası kareler toplamı matris gösterimi olarak

$$\boldsymbol{H} = n\sum_{i=1}^{k} (\bar{y}_{i.} - \bar{y}_{..})(\bar{y}_{i.} - \bar{y}_{..})'$$

ve $\boldsymbol{S}$ yerine grup içi kareler toplamı matris gösterimi olarak

$$\boldsymbol{E} = \sum_{i=1}^{k}\sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.})(y_{ij} - \bar{y}_{i.})'$$

kullanılarak, Fisher tarafından tanımlanan iki varyans oranı

$$\lambda = \frac{\boldsymbol{b}'\boldsymbol{H}\boldsymbol{b}}{\boldsymbol{b}'\boldsymbol{E}\boldsymbol{b}} \tag{2.10}$$

elde edilir. Ayrıca

$$\text{SSH}=n\sum_{i=1}^{k}(\bar{y}_{i.}-\bar{y}_{..})^2 \qquad i=1,2,...,k \tag{2.11}$$

ve

$$\text{SSE}=\sum_{ij}(y_{ij}-\bar{y}_{i.})^2 \qquad i=1,2,...,k \qquad j=1,2,...,n_i \tag{2.12}$$

olmak üzere

$$\lambda=\frac{\text{SSH}(\boldsymbol{y})}{\text{SSE}(\boldsymbol{y})} \tag{2.13}$$

olarak da ifade edilebilir, burada $\text{SSH}(\boldsymbol{y})$ ve $\text{SSE}(\boldsymbol{y})$ $\boldsymbol{y}$ için gruplar arasındaki ve grup içi kareler toplamıdır. Buradan elde edilecek λ_i özdeğerlerine karşılık gelen özvektörler yukarıda belirtilen koşulları sağlayan ayırıcı fonksiyonlar olacaktır.

(2.10) aşağıdaki formda yazılabilir:

$$\boldsymbol{b}'\boldsymbol{H}\boldsymbol{b}=\lambda\boldsymbol{b}'\boldsymbol{E}\boldsymbol{b},$$
$$\boldsymbol{b}'(\boldsymbol{H}\boldsymbol{b}-\lambda\boldsymbol{E}\boldsymbol{b})=0. \tag{2.14}$$

Maksimum λ'da $\boldsymbol{b}$'nın aldığı değer için yapılan araştırmada (2.14)'ün çözümleri olan λ ve $\boldsymbol{b}$ değerleri incelenir. (2.14)'te $\boldsymbol{b}'=\boldsymbol{0}'$ çözümüne $\lambda=0/0$ değerini verdiğinden izin verilmez. Diğer çözümler

$$\boldsymbol{H}\boldsymbol{b}-\lambda\boldsymbol{E}\boldsymbol{b}=0 \tag{2.15}$$

eşitliğinden bulunur,

$$(\boldsymbol{E}^{-1}\boldsymbol{H} - \lambda \boldsymbol{I})\boldsymbol{b} = 0 \qquad (2.16)$$

formunda yazılabilir. (2.16)'nin çözümleri $\lambda_1, \lambda_2, ..., \lambda_s$, $\boldsymbol{E}^{-1}\boldsymbol{H}$ 'nin özdeğerleridir ve $\boldsymbol{b}_1, \boldsymbol{b}_2, ..., \boldsymbol{b}_s$ sırasıyla bu özdeğerlere karşılık gelen özvektörlerdir. Özdeğerlerin $\lambda_1 > \lambda_2 > ... > \lambda_s$ şeklinde sıralandıkları düşünülür. Özdeğerlerin sayısı (sıfır olmayan) *s*, $\boldsymbol{H}$ matrisinin rankıdır ve s=min (*k*-1, *p*) olmak üzere *k*-1 ya da *p*'den daha küçük olarak bulunur. Böylece en büyük özdeğer λ_1, (2.10)'daki $\lambda = \boldsymbol{b}'\boldsymbol{H}\boldsymbol{b} / \boldsymbol{b}'\boldsymbol{E}\boldsymbol{b}$ 'nın maksimum değeridir ve maksimumu üreten katsayı vektörü $\boldsymbol{b}_1$ özvektörüne karşılık gelir. Dolayısıyla ortalamaları maksimum düzeyde ayıran diskriminant fonksiyonu $\boldsymbol{y}_1 = \boldsymbol{b}_1'\boldsymbol{X}$ 'dir, yani, $\boldsymbol{y}_1$ ortalamaları maksimum düzeyde ayıran boyutu ya da yönü temsil eder.

$\lambda_1, \lambda_2, ..., \lambda_s$ özdeğerlerine karşılık gelen $\boldsymbol{E}^{-1}\boldsymbol{H}$ 'nin *s* tane $\boldsymbol{b}_1, \boldsymbol{b}_2, ..., \boldsymbol{b}_s$ özvektörlerinden, $\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2, ..., \bar{\boldsymbol{x}}_k$ arasındaki farklılıkların boyutlarını ya da yönlerini gösteren *s*-tane diskriminant fonksiyonu, $\boldsymbol{y}_1 = \boldsymbol{b}_1'\boldsymbol{X}, \boldsymbol{y}_2 = \boldsymbol{b}_2'\boldsymbol{X}, ..., \boldsymbol{y}_s = \boldsymbol{b}_s'\boldsymbol{X}$ elde edilir. Bu diskriminant fonksiyonları ilişkisizdir, ancak *k*-1< *p* olduğundan $\boldsymbol{E}^{-1}\boldsymbol{H}$ matrisi simetrik olmadığı için bu diskriminant fonksiyonları ortogonal ($i \neq j$ için $\boldsymbol{b}_i'\boldsymbol{b}_j = 0$) değildirler. Dikkat edilmelidir ki özdeğerlere karşılık gelen $\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_s$ numaralanması, burada *k*-grup için yapılmamıştır.

Her bir diskriminant fonksiyonu $\boldsymbol{y}_i$ 'nin bağıl önemi, kendi öz değerinin özdeğerler toplamına oranı alınarak değerlendirilebilir:

$$\frac{\lambda_1}{\sum_{j=1}^{s} \lambda_j} \qquad (2.17)$$

Bu kritere göre, iki ya da üç diskriminant fonksiyonu grup farklılıklarını tanımlamak için genellikle yeterli olur. Küçük özdeğerlerle ilişkili diskriminant fonksiyonları ihmal edilebilir. Her bir diskriminant fonksiyonu için önemlilik testi de mevcuttur.

$\boldsymbol{E}^{-1}\boldsymbol{H}$ matrisi simetrik değildir. Böyle durumlarda özdeğerler ya karekök yöntemi ile ya da geliştirilmiş özel algoritmalar kullanılarak elde edilebilir (Tatlıdil, 2002). Özdeğerlerin ve özvektörlerin hesaplanması için pek çok algoritma sadece simetrik matrisler için kabul edilir. $\boldsymbol{E}$'nin Cholesky faktörizasyonunun $\boldsymbol{E} = \boldsymbol{U}'\boldsymbol{U}$ olması durumunda, $(\boldsymbol{U}^{-1})'\boldsymbol{H}\boldsymbol{U}^{-1}$ simetrik matrisinin özdeğerlerinin $\boldsymbol{E}^{-1}\boldsymbol{H}$'ınkiler ile aynı olduğu gösterilir. Ancak özvektörler için bir düzeltmeye ihtiyaç duyulur. $(\boldsymbol{U}^{-1})'\boldsymbol{H}\boldsymbol{U}^{-1}$'in bir özvektörü $\boldsymbol{c}$ ise, o zaman $\boldsymbol{b} = \boldsymbol{U}^{-1}\boldsymbol{c}$, $\boldsymbol{E}^{-1}\boldsymbol{H}$ 'in bir özvektörüdür.

Yukarıdaki tartışma eşit olmayan örneklem büyüklükleri $n_1, n_2, ..., n_k$ açısından ileri sürülmüştür. Bu durum, uygulamalarda yaygındır ve hiçbir zorlukla karşılaşılmadan ele alınabilir. İdeal olarak, en küçük n_i değişkenlerinin sayısı *p*'yi aşmalıdır. Bu matematiksel olarak gerekli değildir ancak daha durağan (stabil) diskriminant fonksiyonları sağlayacaktır (Rencher, 2002).

2.3.2. Diskriminant Fonksiyonları İçin Bir İlişki Ölçüsü: Fisher'in Kanonik Korelasyonu

Çoklu regresyon analizinde, $\boldsymbol{y}$ bağımlı değişkeni ile $x_1, x_2, ..., x_q$ bağımsız değişkenleri arasındaki ilişki çoklu korelasyon katsayısı R^2 ile verilir. Benzer olarak tek yönlü ANOVA'da bu ilişki Fisher'ın korelasyon oranı η^2 ile verilir. Bunlar sırasıyla

$$R^2 = \frac{\text{regresyon kareler toplamı}}{\text{genel kareler toplamı}}$$

ve

$$\eta^2 = \frac{\text{gruplar arası kareler toplamı}}{\text{genel kareler toplamı}}$$

olarak tanımlanmaktadır.

Bu ilişki ölçüleri kullanılarak, $\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_p$ bağımlı değişkenleri ve $\boldsymbol{\mu}_i$, $i = 1, 2, ..., k$ ile ilişkili bağımsız gruplama değişkeni *i* arasındaki ilişkinin ölçüleri de elde edilebilir. Bu ölçüler, "değişkenler gruplara ne kadar iyi ayrılabilir?" sorusuna cevap vermeye çalışır (Rencher, 2002).

(2.11) ve (2.12) eşitlikleri kullanılarak, birinci diskriminant fonksiyonu $\boldsymbol{y}_1 = \boldsymbol{b}_1'\boldsymbol{X}$ için grup arasındaki kareler toplamı ile toplam kareler arasındaki oranı:

$$\eta_\theta^2 = \theta = \frac{\lambda_1}{1+\lambda_1} = \frac{\text{SSH}(\boldsymbol{y}_1)}{\text{SSE}(\boldsymbol{y}_1) + \text{SSH}(\boldsymbol{y}_1)} \tag{2.18}$$

olduğundan, Roy'un θ istatistiğinin R^2 benzeri bir ilişki ölçüsü olarak görev yaptığı belirtilir. η_θ^2'nin başka bir yorumu, birinci diskriminant fonksiyonu ve *k*-1 (kukla) grup üye değişkenlerinin en iyi doğrusal kombinasyonu arasındaki maksimum R^2'dir. Maksimum korelasyona (birinci) kanonik korelasyon adı verilir. Karesel kanonik korelasyon her bir diskriminant fonksiyonu için

$$r_i^2 = \frac{\lambda_i}{1+\lambda_i}, \qquad i = 1, 2, ..., s \tag{2.19}$$

eşitliği yardımıyla hesaplanır.

Çok değişkenli ilişkinin bir ölçüsü olarak kullanılan ortalama karesel kanonik korelasyon;

$$A_p = \frac{\sum_{i=1}^{s} r_i^2}{s}, \qquad i = 1, 2, ..., s \tag{2.20}$$

ifadesi de ilişkinin bir ölçüsü olarak kullanılabilir.

2.4. Diskriminant Analizinde Anlamlılık Testleri

Diskriminant fonksiyonlarının kurulması için çok değişkenli normallik varsayımları açıkça gerekli değildir. Ancak hipotezleri test etmek için normallik varsayımına ihtiyaç duyulur.

Diskriminant analizi MANOVA'nın tersidir ve anlamsal olarak aralarında karışıklık mevcuttur. MANOVA'da bağımsız değişkenler gruplar ve bağımlı değişkenler tahmin edicilerken; diskriminant analizinde bağımsız değişkenler tahmin ediciler, bağımlı değişkenler ise gruplardır. Vurgulanan değişkenler farklılık göstermesine rağmen, matematiksel olarak ikisi de aynıdır. Bu nedenle diskriminant analizinde genel istatistiksel anlamlılığı değerlendirme kriterleri MANOVA'dakilerle benzerdir.

Diskriminant analizinde; Wilk's Lamda (Wilk's Λ), Roy en büyük karakteristik kökü (Roy's θ), Hotelling izi (Lawley-Hotelling $U^{(s)}$) ve Pillai kriteri (Pillai $V^{(s)}$) arasındaki seçim MANOVA ile aynı etkenlere sahiptir. Rencher'e göre, diskriminant fonksiyonları kapsamında Wilk's Λ özdeğerlerin bir alt kümesi üzerinde kullanılabildiği için; Roy's θ, Lawley-Hotelling $U^{(s)}$ ve Pillai $V^{(s)}$ olan diğer üç MANOVA testinden daha yararlıdır.

Diskriminant analizi adımlarını yönlendirmek ve grup üyeliğini tahmin etmek amacıyla bağımsız değişkenlerin güvenilirliğini değerlendirmek için uygun olan bir diğer istatistiksel kriter Mahalanobis uzaklığıdır.

2.4.1. İki Grup Durumu İçin Testler

(2.3) eşitliği aracılığıyla, diskriminant fonksiyonu $\boldsymbol{y} = \boldsymbol{b}'\boldsymbol{X}$ ile elde edilen dönüştürülmüş ortalamaların ayrışımının $\bar{\boldsymbol{x}}_1$ ve $\bar{\boldsymbol{x}}_2$ ortalama vektörleri arasındaki standartlaştırılmış uzaklığa (Mahalanobis D^2 uzaklığına) eşit olduğu görülmektedir. Çok değişkenli normal dağılımlı ve ortak varyans-kovaryanslı kitleden gelen örneklemler için D^2 değeri *F* dağılımına sahiptir.

$$F = \frac{n_1 n_2 (n_1 + n_2 - p - 1)}{p(n_1 + n_2)(n_1 + n_2 - 2)} D^2 \sim F_{\alpha, p, n_1+n_2-p-1} \tag{2.21}$$

(2.21) eşitliğinden yararlanarak, elde edilen diskriminant fonksiyonunun ayırıcı özelliğinin anlamlı olup olmadığının testinde *F* tablosunun uygun serbestlik derecelerindeki tablo değerleri kritik değer olarak kullanılır. Hesaplanan *F* değeri $F_{\alpha, p, n_1+n_2-p-1}$ tablo değeri ile karşılaştırılır. $F > F_{\alpha, p, n_1+n_2-p-1}$ olması durumunda, bulunan diskriminant fonksiyonun bireyleri (gözlemleri) gruplara ayırma özelliğinin iyi olduğu sonucuna ulaşılır. Aksi durumda diskriminant analizi sonuçlarına güvenilmeyeceği yorumu yapılır.

Bu standartlaştırılmış D^2 uzaklığı iki gruplu Hotelling T^2 ile de doğru orantılıdır.

$$T^2 = \left(\frac{n_1 n_2}{n_1 + n_2} \right) D^2 \tag{2.22}$$

olduğu ve T^2'nin *F*-dağılımı yaklaşımından faydalanarak

$$F = \frac{(n_1 + n_2 - p - 1)}{p(n_1 + n_2 - 2)} T^2 \sim F_{\alpha, p, n_1+n_2-p-1} \tag{2.23}$$

olduğu göz önüne alınırsa, *F*'nin anlamlılığı p ve $n_1 + n_2 - p - 1$ serbestlik dereceli *F*-dağılımının kritik değeri kullanılarak test edilir. $F > F_{\alpha, p, n_1+n_2-p-1}$ olması durumunda, dolayısıyla T^2 anlamlı ise diskriminant fonksiyonu katsayı vektörü $\boldsymbol{b}$, önemli ölçüde $\boldsymbol{0}$'dan farklıdır. Daha biçimsel olarak, kitle diskriminant fonksiyonu katsayı vektörü; $\boldsymbol{b} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ olarak ifade edilmektedir, böylece $H_0 : \boldsymbol{b} = \boldsymbol{0}$ hipotezi $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ hipotezine eşit olur.

Diskriminant fonksiyonu katsayılarının bir alt kümesinin anlamlılığını test etmek için, $\boldsymbol{X}$'lerin karşılık gelen altkümelerinin testi kullanılabilir. Belirli bir $\boldsymbol{b}_0'\boldsymbol{X}$ formuna sahip olan kitle diskriminant fonksiyonunun hipotezleri de test edilebilir (Rencher, 1998).

2.4.2. İkiden Fazla Grup Durumu İçin Testler

İkiden fazla grup olması durumda elde edilen diskriminant fonksiyonlarının anlamlılık kontrolünde Wilk's Λ kriteri kullanılır. Bu kriter Wilk's *U* olarak ve genelleştirilmiş varyans oranı olarak da bilinmektedir (Rencher,2002). Λ kriteri ile diskriminant fonksiyonlarının diskriminant değerleri arasında cebirsel bir ilişki vardır. Λ değerinin küçük bulunması gruplar arası farklılığın önemli olduğunu gösteren bir işarettir. Bu değer, aynı zamanda MANOVA'nında temelini teşkil etmektedir (Cangül, 2006).

Wilk's Λ testi olarak bilinen, $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = ... = \boldsymbol{\mu}_k$ hipotezinin olabilirlik oran testi

$$\Lambda = \frac{|\boldsymbol{E}|}{|\boldsymbol{E}+\boldsymbol{H}|}$$

ile verilir. $\Lambda \le \Lambda_{\alpha,p,\upsilon_H,\upsilon_E}$ ise sıfır hipotezi reddedilir. Burada *p:* değişken sayısı (boyut), $\upsilon_H = k-1$: hipotez için serbestlik derecesi ve $\upsilon_E = N-k$: hata için serbestlik derecesi olarak tanımlanır.

Wilk's Λ, $\boldsymbol{E}^{-1}\boldsymbol{H}$ 'in $\lambda_1, \lambda_2, ..., \lambda_s$ özdeğerleri bakımından aşağıdaki şekilde de ifade edilir:

$$\Lambda = \prod_{i=1}^{s} \frac{1}{1+\lambda_1}$$

$\lambda = \boldsymbol{b}'\boldsymbol{H}\boldsymbol{b} / \boldsymbol{b}'\boldsymbol{E}\boldsymbol{b}$ diskriminant kriterinin, $\boldsymbol{E}^{-1}\boldsymbol{H}$ 'in en büyük özdeğeri λ_1 ile maksimize edildiği ve kalan $\lambda_2, ..., \lambda_s$ özdeğerlerinin diğer diskriminant boyutlarına karşılık geldiği ifade edilmişti. Bu özdeğerler ortalama vektörler arasındaki anlamlı farklılıklar için Wilk's Λ testindekiler ile aynıdır.

Dengeli olmayan bir tasarım (her bir grupta farklı örneklem) için $N = \sum_i n_i$ ya da dengeli (her bir grupta eşit örneklem) durumda $N = kn$ iken

$$\Lambda_1 = \prod_{i=1}^{s} \frac{1}{1+\lambda_1}, \qquad (2.24)$$

$\Lambda_{p,k-1,N-k}$ olarak dağılmaktadır. Λ_1 küçük olduğundan dolayı, bir ya da daha fazla λ_1 büyük olsa bile özdeğerlerin anlamlılığı için ve bu yüzden diskriminant fonksiyonları için Wilks' Λ - testi yapılır.

s özdeğer, $\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2, ..., \bar{\boldsymbol{x}}_k$ ortalama vektörlerinin ayrımının s boyutunu ifade eder. Eğer varsa bu boyutların hangisinin anlamlı olduğu ile ilgilenilir.

Λ için kullanılan kritik değerler tablosundan yararlanılarak sağlanan anlamlılık testine ek olarak, Λ_1 için χ^2 yaklaşımı kullanılabilir. $\upsilon_E = N - k = \sum_i n_i - k$ ve $\upsilon_H = k - 1$ ile

$$\begin{aligned} V_1 &= -[\upsilon_E - \tfrac{1}{2}(p - \upsilon_H + 1)] \ln \Lambda_1 \\ &= -[N - 1 - \tfrac{1}{2}(p + k)] \ln \prod_{i=1}^{s} \frac{1}{1+\lambda_i} \\ &= [N - 1 - \tfrac{1}{2}(p + k)] \sum_{i=1}^{s} \ln(1+\lambda_i), \end{aligned} \qquad (2.25)$$

yaklaşık olarak $p(k-1)$ serbestlik dereceli χ^2'dir. Λ_1 test istatistiği ve onun (2.25) yaklaşımı $\lambda_1, \lambda_2, ..., \lambda_s$'in tamamının anlamlılığını test eder. Test, H_0'ın reddedilmesine neden oluyorsa, λ'ların en az birinin anlamlı olarak sıfırdan farklı olduğu sonucu çıkarılır ve bundan dolayı ortalama vektörlerin ayrışımının en az bir boyutu vardır. λ_1 en büyük olduğu için, $\boldsymbol{y}_1 = \boldsymbol{b}_1' \boldsymbol{X}$ ile birlikte sadece onun anlamlılığından emin olunur.

$\lambda_2, \lambda_3 ..., \lambda_s$'nin anlamlılığını test etmek için, Wilk's Λ'dan λ_1 silinir ve ilişkili χ^2-yaklaşımı aşağıdaki gibi elde edilir:

$$\Lambda_2 = \prod_{i=2}^{s} \frac{1}{1+\lambda_i}, \qquad (2.26)$$

$$V_2 = -[N-1-\tfrac{1}{2}(p+k)]\ln\Lambda_2$$

$$= [N-1-\tfrac{1}{2}(p+k)]\sum_{i=2}^{s}\ln(1+\lambda_i) \qquad (2.27)$$

Burada V_2, yaklaşık $(p-1)(k-2)$ serbestlik dereceli χ^2 'dir. Bu test H_0 'ın reddedilmesine neden oluyorsa, en azından $\boldsymbol{y}_2 = \boldsymbol{b}_2'\boldsymbol{X}$ diskriminant fonksiyonu ile ilişkili λ_2 'nin anlamlı olduğu sonucu çıkarılır. Bu teste her bir λ_i için H_0 reddedilmeyene kadar bu şekilde devam edilebilir.

m_inci adımdaki test istatistiği

$$\Lambda_m = \prod_{i=m}^{s}\frac{1}{1+\lambda_i} \qquad (2.28)$$

'dir, $\Lambda_{p-m+1,k-m,N-k-m+1}$ ile dağılmaktadır.

$$V_m = -[N-1-\tfrac{1}{2}(p+k)]\ln\Lambda_m$$

$$= [N-1-\tfrac{1}{2}(p+k)]\sum_{i=m}^{s}\ln(1+\lambda_i) \qquad (2.29)$$

istatistiği, yaklaşık $(p-m+1)(k-m)$ serbestlik dereceli χ^2 dağılımına sahiptir. Bazı durumlarda, araştırmacının pratik olarak öneme sahip olduğunu düşünmesinden dolayı daha fazla λ istatistiksel olarak anlamlı olacaktır. $\lambda_i / \sum_j \lambda_j$ küçük ise, ilişkili diskriminant fonksiyonu anlamlı olsa bile ilgi çekici olmayabilir.

Her bir Λ_i için bir *F*- yaklaşımı da kullanılabilir.

$$\Lambda_1 = \prod_{i=1}^{s} \frac{1}{1+\lambda_i}$$

için, $\upsilon_E = N-k$ ve $\upsilon_H = k-1$ ile kullanılarak,

$$F = \frac{1-\Lambda_1^{1/t}}{\Lambda_1^{1/t}} \frac{\mathrm{df}_2}{\mathrm{df}_1} \tag{2.30}$$

elde edilir, burada

$$t = \sqrt{\frac{p^2(k-1)^2-4}{p^2+(k-1)^2-5}}, \qquad w = N-1-\tfrac{1}{2}(p+k),$$

$$\mathrm{df}_1 = p(k-1), \qquad \mathrm{df}_2 = wt-\tfrac{1}{2}[p(k-1)-2]$$

'dir.

$$\Lambda_m = \prod_{i=m}^{s} \frac{1}{1+\lambda_i}, \qquad m = 2,3,\ldots,s$$

için,

$$F = \frac{1-\Lambda_m^{1/t}}{\Lambda_m^{1/t}} \frac{\mathrm{df}_2}{\mathrm{df}_1} \tag{2.31}$$

kullanılır, burada *p* ve *k*-1 yerine *p*-*m*+1 ve *k*-*m* ile:

$$t = \sqrt{\frac{(p-m+1)^2(k-m)^2 - 4}{(p-m+1)^2 + (k-m)^2 - 5}},$$

$$w = N - 1 - \tfrac{1}{2}(p+k),$$

$$\mathrm{df}_1 = (p-m+1)(k-m),$$

$$\mathrm{df}_2 = wt - \tfrac{1}{2}[(p-m+1)(k-m) - 2]$$

şeklindedir.

2.5. Diskriminant Fonksiyonlarının Yorumu

Diskriminant fonksiyonlarını yorumlama ve her bir değişkenin katkısını belirleme arasında yakın bir ilişki vardır ve her zaman bir ayrım yapılamamaktadır. Yorumda katsayıların işareti dikkate alınır; katkısını belirlemede işaretler görmezden gelinir ve katsayılar mutlak değerce sıralanır. Daha yaygın olarak fonksiyonları yorumlamadan ziyade değişkenlerin katkısını değerlendirme ile ilgilenilir.

Grupları ayırmak için diğer değişkenlerin varlığında her bir değişkenin katkısını değerlendirmek için yaygın üç yaklaşım vardır. Bu üç yaklaşım:

1. Standartlaştırılmış diskriminant fonksiyonu katsayılarını belirlemek,
2. Her bir değişken için kısmi F-testi hesaplamak,
3. Her bir değişken ve diskriminant fonksiyonu arasında bir ilişki hesaplamaktır.

2.5.1. Standartlaştırılmış Diskriminant Fonksiyonu Katsayıları

Diskriminant fonksiyonlarındaki değişkenlerin ayırmaya etkilerinin ya da fonksiyona ne kadar katkıda bulunduklarının bilinmesi özellikle yorum aşamasında önemli olduğundan, böyle bir karşılaştırmanın yapılabilmesi için bulunan katsayıların standartlaştırılması gerekir.

İki grup durumunda, iki grubun ayrılmasında *x*'lerin bağıl katkısı $\boldsymbol{y} = \boldsymbol{b}'\boldsymbol{X} = b_1 x_1 + b_2 x_2 + ... + b_p x_p$ diskriminant fonksiyonundaki b_r, $r = 1, 2, ..., p$ katsayıları karşılaştırılarak değerlendirilir. İkiden fazla grubun ayrılmasında *x*'lerin katkısını değerlendirmek için diskriminant fonksiyonlarının kullanımı ile ilgili yine benzer yorumlar yapılır. Ancak bu tür karşılaştırmalar sadece *x*'ler orantılı ise, yani aynı ölçek üzerinde ölçülmüş ve karşılaştırılabilir varyanslı ise bilgilendiricidir. *x*'ler orantılı değilse, standartlaştırılmış değişkenlere uygulanabilen b_r^* katsayılarına ihtiyaç duyulur.

İki gruplu durumda, Grup 1 ya da 2'deki *i*_inci gözlem vektörü $\boldsymbol{x}_{1i}$ ya da $\boldsymbol{x}_{2i}$ için, standartlaştırılmış değişkenler açısından diskriminant fonksiyonu

$$y_{1i} = b_1^* \frac{x_{1i1} - \overline{x}_{11}}{s_1} + b_2^* \frac{x_{1i2} - \overline{x}_{12}}{s_2} + ... + b_p^* \frac{x_{1ip} - \overline{x}_{1p}}{s_p}, \qquad (2.32)$$

$$i = 1, 2, ..., n_1,$$

$$y_{2i} = b_1^* \frac{x_{2i1} - \overline{x}_{21}}{s_1} + b_2^* \frac{x_{2i2} - \overline{x}_{22}}{s_2} + ... + b_p^* \frac{x_{2ip} - \overline{x}_{2p}}{s_p},$$

$$i = 1, 2, ..., n_2$$

şeklinde ifade edilir. Burada $\bar{\boldsymbol{x}}_1' = (\bar{x}_{11}, \bar{x}_{12}, ...\bar{x}_{1p})$ ve $\bar{\boldsymbol{x}}_2' = (\bar{x}_{21}, \bar{x}_{22}, ...\bar{x}_{2p})$ iki grup için ortalama vektörleridir ve s_r, $\boldsymbol{S}$ 'in *r*_inci köşegen elemanının karekökü olarak elde edilen *r*_inci değişkeninin örneklem içi standart sapmasıdır. Açık bir şekilde, bu standartlaştırılmış katsayılar

$$b_r^* = s_r b_r, \qquad r = 1, 2, ..., p \tag{2.33}$$

formunda olmalıdır. Vektör formu ise

$$\boldsymbol{b}^* = (\text{diag } \boldsymbol{S})^{1/2} \boldsymbol{b} \tag{2.34}$$

şeklinde ifade edilir.

İkiden fazla grup durumu için, benzer bir biçimdeki diskriminant fonksiyonları standartlaştırılabilir. b_{mr}, $m = 1, 2, ..., s$; $r = 1, 2, ..., p$ ile *m*_inci diskriminant fonksiyonundaki *r*_inci katsayı ifade edilirse, o zaman

$$b_{mr}^* = s_r b_{mr} \tag{2.35}$$

standartlaştırılmış formdur, burada $\upsilon_E = N - k = \sum_{i=1}^{k} n_i - k$ hata için serbestlik derecesi olmak üzere s_r, $\boldsymbol{S} = \boldsymbol{E} / \upsilon_E$ 'nin köşegeninden elde edilen grup içi standart sapmadır. İki grup için bir diskriminant fonksiyonu olduğundan (2.3)'deki b_r^* 'ın sadece tek bir alt simgeye sahip olması gibi, burada birçok diskriminant fonksiyonu olduğundan b_{mr}^* 'ın iki alt simgeye sahip olduğuna dikkat edilmelidir.

Alternatif olarak, m_inci özvektör sadece bir skaler ile çarpılana kadar tek olduğu için,

$$b_{mr}^{*} = \sqrt{e_{rr}}\, b_{mr}, \qquad r = 1, 2, ..., p \tag{2.36}$$

kullanılarak standartlaştırma daha da kolaylaştırılabilir. Burada e_{rr}, $\boldsymbol{E}$ 'nin r_inci köşegen elemanıdır.

Değişkenlerin grup ayrılmasına bağıl katkısını ölçmek için, diğer bir ifadeyle değişkenler arasında farklı olan ölçekleri dengelemek için standartlaştırılmış diskriminant fonksiyon katsayıları kullanılır. Örneğin, iki gruptan birincisindeki gözlemler için (2.32) ile

$$y_{1i} = b_1^{*} \frac{x_{1i1} - \overline{x}_{11}}{s_1} + b_2^{*} \frac{x_{1i2} - \overline{x}_{12}}{s_2} + ... + b_p^{*} \frac{x_{1ip} - \overline{x}_{1p}}{s_p}, \qquad i = 1, 2, ..., n_1$$

fonksiyonu mevcuttur. $(x_{1ir} - \overline{x}_{1r}) / s_r$ standartlaştırılmış değişkenleri ölçekten bağımsızdır ve bundan dolayı $b_r^{*} = s_r b_r$, $r = 1, 2, ..., p$ standartlaştırılmış katsayıları, grupları maksimum olarak ayırırken diskriminant fonksiyonu $\boldsymbol{y}$ 'ye değişkenlerin ortak katkısını tam olarak yansıtır. İkiden fazla grup durumu için, $\boldsymbol{b} = (b_1, b_2, ..., b_p)'$ her bir diskriminant fonksiyonu katsayı vektörü, $\boldsymbol{E}^{-1}\boldsymbol{H}$ 'nin bir özvektörüdür ve bu şekilde diğerlerinin varlığında her bir değişkenin etkisinin yanı sıra değişkenler arasındaki örneklem korelasyonlarını da hesaba katar.

Bu standartlaştırma s-tane diskriminant fonksiyonunun her biri için uygulanır. Genel olarak, her biri farklı bir yoruma sahip olacaktır; yani, b_r^{*} katsayılarının yapısı bir fonksiyondan diğerine çeşitlilik gösterecektir. Gruplara ayırmak amacıyla katkılarının sırasına göre değişkenleri sıralamak için,

katsayıların mutlak değerleri kullanılabilir. Daha ileri gitmek ve yorumlamak ya da bir diskriminant fonksiyonuna "isim" vermek istersek, işaretler dikkate alınabilir. Bundan dolayı, örneğin; $y_1 = .8x_1 - .9x_2 + .5x_3$, $y_2 = .8x_1 + .9x_2 + .5x_3$'den farklı anlama gelmektedir. Çünkü y_1; x_1 ve x_2 arasındaki farka bağlı iken y_2; x_1 ve x_2 'nin toplamı ile ilişkilidir.

Diskriminant fonksiyonu bir regresyon denklemi gibi diğer lineer kombinasyonlar ile aynı sınırlamalara tabidir. Örneğin:

1. Özellikle değişkenler eklenir ya da silinirse bir değişken için katsayı değişebilir.
2. Örneklem büyüklüğü değişkenlerin sayısına göre çok büyük olmazsa, katsayılar örneklemden örnekleme durağan olmayabilir.

Birinci sınırlama, eldeki belirli değişkenlerin varlığında katsayıların her bir değişkenin katkısını yansıttığına dikkat çeker. Gerçekte bu durum, katsayılara uygulanmak istenen şeydir. İkinci sınırlamaya göre, katsayıların tahminleri işleme alınan değişkenlerden bağımsız değildir. Daha durağan tahminler, 2 değişkenli 50 gözlemden elde edilenden çok 20 değişkenli 50 gözlemden elde edilir. Diğer bir ifadeyle, N / p çok küçük ise, bir örneklemde yüksek sıralamaya (ranka) sahip değişkenler diğer örneklemde daha az önemli olarak ortaya çıkabilir (Rencher, 2002).

2.5.2. Kısmi *F*-Değerleri

Herhangi bir x_r değişkeni için, diğer değişkenlere x_r'nin eklenmesiyle sağlanan ayrışımı görmek için, yani x_r'nin önemini görmek için kısmi *F*-testi yapılır. Her bir değişken için kısmi *F* hesaplandıktan sonra, değişkenler sıralanır.

İki grup durumunda, kısmi *F* değeri

$$F = (v - p + 1)\frac{T_p^2 - T_{p-1}^2}{v + T_{p-1}^2} \tag{2.37}$$

olarak verilmektedir, burada T_p^2; tüm *p* değişken için iki örneklem Hotelling T^2 'dir, T_{p-1}^2; x_r hariç tüm değişkenler için T^2-istatistiği ve $v = n_1 + n_2 - 2$ 'dir. (2.37)'deki *F*-istatistiği $F_{1,v-p+1}$ olarak dağılmaktadır.

İkiden fazla grup durumu için, diğer *p*-1 değişkene göre düzeltilmiş x_r için kısmi Λ

$$\Lambda(x_r \setminus x_1, ..., x_{r-1}, x_{r+1}, ..., x_p) = \frac{\Lambda_p}{\Lambda_{p\text{-}1}} \tag{2.38}$$

olarak verilmektedir, burada Λ_p; tüm *p* değişken için Wilks' Λ 'dır ve Λ_{p-1}; x_r hariç tüm değişkenleri içerir. Karşılık gelen kısmi *F*

$$F = \frac{1-\Lambda}{\Lambda}\frac{v_E - p + 1}{v_H} \tag{2.39}$$

olarak verilmektedir, burada $v_E = N - k$, $v_H = k - 1$ ve Λ (2.38)'da tanımlanmıştır. (2.38)'daki kısmi Λ-istatistiği, Λ_{1,v_H,v_E-p+1} olarak dağılmaktadır ve (2.39)'deki kısmı *F*, F_{v_H,v_E-p+1} olarak dağılmaktadır.

(2.38) ve (2.39)'deki kısmi *F*-değerleri tek boyutlu grup ayrışımı ile ilişkili değildir çünkü bunlar standartlaştırılmış diskriminant fonksiyon katsayılarıdır. Örneğin; x_2, *s* diskriminant fonksiyonunun her birinde farklı bir katkıya sahip

olur, fakat x_2 için kısmi *F*- değeri tüm boyutlar göz önüne alınarak grup ayrışımına x_2'nin katkısının genel bir gösterimini oluşturur. Ancak kısmi *F*-değerleri, birinci diskriminant fonksiyonları için özellikle birinci fonksiyonun mevcut ayrışmanın çoğunu açıklamasından dolayı $\lambda_1 / \sum_j \lambda_j$ çok büyükse, standartlaştırılmış katsayılara göre değişkenleri çoğu zaman aynı sırada sıralayacaktır.

Tüm *x*'ler için yapılan $\eta_\Lambda^2 = 1 - \Lambda$ genel ilişki ölçüsüne benzer olarak, x_r için ilişkinin kısmi bir dizini,

$$R_r^2 = 1 - \Lambda_r, \qquad r = 1, 2, ..., p \tag{2.40}$$

olarak tanımlanabilir, burada Λ_r ; x_r için (2.38)'daki kısmi Λ 'dır. Bu kısmi R^2, gruplanan değişkenler ve diğer $(p-1)$ tane *x*'e göre düzeltildikten sonraki x_i arasındaki ilişkinin bir ölçüsüdür.

2.5.3.Değişkenler ile Diskriminant Fonksiyonları Arasındaki İlişkiler

Birçok kitabın ve araştırma yazılarının değerlendirdiğine göre değişken öneminin en iyi ölçümü $r_{x_i y_j}$, her bir değişken ve diskriminant fonksiyonu arasındaki ilişkidir (korelasyondur). İddia edildiğine göre bu ilişkiler değişkenlerin diskriminant fonksiyonlarına ortak katkısına göre standartlaştırılmış katsayılardan daha bilgi vericidir. Bu ilişkiler çoğu zaman yük ve yapı katsayıları olarak adlandırılır ve rutin bir şekilde birçok temel programlarda hesaplanmaktadır. Fakat, Rencher (1988; 1992b; 1998) söz konusu bu korelasyonların çok değişkenli durumda değil de tek değişkenli durumda her bir değişkenin katkısını gösterdiğini belirtmektedir. Bu korelasyonlar aslında her bir değişken için *t* yi ve *F* i yeniden üretirler ve devamında diğer değişkenlerin varlığını görmezden gelerek sadece her

bir değişkenin kendi başına grupları nasıl ayırdığını gösterirler. Bu yüzden bu korelasyonlar değişkenlerin grupların ayrılmasına ortak katkısı hakkında herhangi bir bilgi vermez. Eğer diskriminant fonksiyonlarının yorumu için kullanılırlarsa yanıltıcı olurlar.

Çok değişkenli bilgi sağlamada korelasyonların başarısızlığı grafiksel incelemeler yapılarak da tahmin edilebilirdi. Standartlaştırılmış katsayılara olan itiraz, bazı değişkenlerin silinip ve diğerlerinin eklenmesi durumunda değiştiklerinden dolayı onların "durağan olmaması" görüşü üzerine dayalıdır. Aslında bu yolla onların davranışları ve hatta değişkenlerin birbirleri üzerindeki ortak etkisinin yansımaları bilinmek istenir. Çok değişkenli analizde, ilgi eldeki değişken kümesinin ortak performansı üzerine odaklanır. Diğer tüm değişkenlerden bağımsız olarak her bir değişkenin katkısını istemek, diğer değişkenleri görmezden gelen tek değişkenli bir dizin istemektir. $r_{x_i y_j}$ ilişkileri sabittir ve değişkenlerin eklenip silinmesinden etkilenmezler. Bu onların doğası gereği tek değişkenli olduklarının açık bir göstergesidir (Rencher, 2002).

2.6. Diskriminant Analizinde Saçılım Grafikleri

Diskriminant analizi ile bireyler tam ölçüm uzayından uygun olan bir alt uzaya indirgenir. Bu nedenle diskriminant analizi, asıl amacı dışında boyut indirgeme tekniği biçiminde de düşünülebilir. Diskriminant analizi tarafından etkilenmiş boyut indirgemenin bir faydası grafik çizme potansiyelidir. $\boldsymbol{E}^{-1}\boldsymbol{H}$'ın büyük özdeğerlerinin sayısı ortalama vektörler tarafından oluşturulan uzayın boyutunu yansıtır. Birçok veri setinde, ilk iki diskriminant fonksiyonu $\lambda_1 + \lambda_2 + ... + \lambda_s$'nin çoğunu açıklar ve ardından ortalama vektörlerinin şekli etkili bir şekilde iki boyutlu grafikte gösterilebilir. Eğer gerekli boyut ikiden büyükse, iki boyutlu grafikteki gruplar arası görünümde bazı bozulmalar meydana gelebilir, yani iki boyutta çakışan bazı grupları üçüncü boyutta daha iyi ayırabilir.

Bireysel gözlem vektörleri $\boldsymbol{x}_{ij}$'lere göre ilk iki diskriminant fonksiyonunu çizmek için $i=1,2,...,k$; $j=1,2,...,n_i$ olmak üzere $y_{1ij}=\boldsymbol{b}_1'\boldsymbol{x}_{ij}$ ve $y_{2ij}=\boldsymbol{b}_2'\boldsymbol{x}_{ij}$ hesaplanır ve

$$\boldsymbol{y}_{ij}=\begin{pmatrix} y_{1ij} \\ y_{2ij} \end{pmatrix}=\begin{pmatrix} \boldsymbol{b}_1'\boldsymbol{x}_{ij} \\ \boldsymbol{b}_2'\boldsymbol{x}_{ij} \end{pmatrix}=\begin{pmatrix} \boldsymbol{b}_1' \\ \boldsymbol{b}_2' \end{pmatrix}\boldsymbol{x}_{ij}=\boldsymbol{B}\boldsymbol{x}_{ij} \tag{2.41}$$

ifadesinin $N=\sum_i n_i$ değerinin saçılım grafiği çizilir.

Dönüştürülmüş ortalama vektörleri,

$$\bar{\boldsymbol{y}}_i=\begin{pmatrix} \bar{y}_{1i} \\ \bar{y}_{2i} \end{pmatrix}=\begin{pmatrix} \boldsymbol{b}_1' \\ \boldsymbol{b}_2' \end{pmatrix}\bar{\boldsymbol{x}}_i=\boldsymbol{B}\bar{\boldsymbol{x}}_i, \quad i=1,2,...,k \tag{2.42}$$

$\boldsymbol{y}_{ij}$ bireysel değerleri ile çizilmiş olmalıdır. Bazı durumlarda, grafik sadece $\bar{\boldsymbol{y}}_1,\bar{\boldsymbol{y}}_2,...,\bar{\boldsymbol{y}}_k$ dönüştürülmüş ortalama vektörleri ile gösterilebilir.

$\boldsymbol{E}^{-1}\boldsymbol{H}$'in özdeğerleri bireysel noktalarının değil, ortalama vektörlerinin boyutunu ortaya çıkarmaktadır. Gerekli boyut daha az olmalıyken değişkenler ilişkili olduğu için, bireysel gözlemlerin boyutu *p*'dir. Örneğin *s*=2 ise, bu şekilde ortalama vektörleri sadece iki boyutu kaplar, bireysel gözlem vektörleri normal olarak ikiden daha fazla boyuta uzanırlar ve onların grafiğe dahil edilmesi ortalama vektörlerin iki boyutlu düzlem üzerinde bir izdüşümünü oluşturur.

Diskriminant fonksiyonlarının ilişkisiz ancak ortogonal olmadıklarından dolayı, $\boldsymbol{b}_1$ ve $\boldsymbol{b}_2$ arasındaki açı 90^0 olamaz (yani, $\boldsymbol{b}_1'\boldsymbol{b}_2\neq 0$). Yine de uygulamada alışılmış olan yöntem, dikdörtgensel bir koordinat sistemi üzerine diskriminant fonksiyonlarını çizmektir. Oluşan bozulma genellikle ciddi değildir.

2.7. Diskriminant Analizinde Değişkenlerin Adımsal Seçimi

Birçok uygulamada çok sayıda bağımlı değişken mevcuttur ve araştırmacı grupları ayırmak için diğer değişkenlerin varlığında gereksiz olan değişkenleri atmak ister. Araştırmacının diğerlerine oranla bazı bağımsız değişkenlere daha fazla öncelik vermek için bir nedeni yoksa ön incelemede giriş sırasını belirlemek için istatistiksel kriterler kullanılır. Diğer bir deyişle daha az bağımsız değişken isteniyor ama bunlar arasında bir tercih sebebi yoksa daha az tahmin edici üretmek için adımsal (stepwise) diskriminant analizi kullanılır. Burada bağımlı değişkenler (*y*'ler) seçilerek vurgulanır ve bundan dolayı temel model (tek-yönlü MANOVA) değişmez. Diğer yandan, regresyondaki alt küme seçiminde modelin değişmesine bağlı olarak bağımsız değişkenler seçilmektedir.

İleri doğru seçim yönteminde; kendisi aracılığıyla grupları maksimum şekilde ayıran bir tek değişken ile başlanır. Daha sonra her bir adımda, Wilk's Λ 'ya dayanarak kısmi *F*-istatistiğini maksimum yapan değişken dahil edilir. Böylece üzerinde ve ötesinde ayrışım gruplarına göre maksimum ek ayrışımın elde edilmesi, zaten diğer değişkenler tarafından sağlanmış olur.

Geriye doğru eleme yönteminde; tüm değişkenlerle başlanır ve daha sonra her bir adımda kısmi *F* değeri ile gösterilen en az katkı sağlayan değişken silinir.

Adımsal seçim ileriye ve geriye doğru yaklaşımların bir kombinasyonudur. Her seferinde değişkenlerden biri eklenir ve her bir adımda değişkenler, önceden girilen herhangi bir değişkenin son eklenen değişkenlerin varlığında gereksiz hale gelip gelmediğini görmek için yeniden incelenir. Bu işlem, giriş için mevcut değişkenler arasındaki en büyük kısmi *F* önceden belirlenmiş bir eşik değeri aşamadığında sonlandırılır. Adımsal seçim yöntemi, pratikte uygulanmasından dolayı uzun zamandır popülerdir.

Uygulanan bu yöntemlerin tamamı genel olarak adımsal diskriminant analizi olarak adlandırılmaktadır. Ancak, aslında adımsal MANOVA yapılmaktadır. Hiçbir diskriminant fonksiyonu seçim sırasında hesaplanamaz. Seçilmiş

değişkenler için diskriminant fonksiyonları alt küme seçimi tamamlandıktan sonra hesaplanabilir. Ayrıca, bu seçilen değişkenler sınıflandırma analizinde kullanılabilecektir.

2.8. Sınıflandırma Analizi: Gözlemlerin Gruplara Atanması

Diskriminant fonksiyonları tarafından karakterize edilen grup ayrışımında diskriminant analizinin tanımlayıcı yönü ele alınır. Gözlemlerin gruplara atanması olayı ise diskriminant analizinin öngörü yönüdür. Bu durumu, diskriminant analizini tanımlayıcı yönünden açık bir şekilde ayırmak için 'sınıflandırma analizi' olarak adlandırmak tercih edilmiştir. Fakat sınıflandırma sık sık diskriminant analizi olarak da ifade edilmektedir.

Sınıflandırmada, grup üyeliği bilinmeyen bir örnekleme birimi (konu veya nesne), birimle ilişkili ölçülmüş değeri $\boldsymbol{X}$ olan *p*-boyutlu vektörler üzerinde bir gruba atanır. Birimi sınıflandırmak için her bir gruptan gözlem vektörlerinin önceden elde edilmiş bir örneğinin bulunması gereklidir. Daha sonraki adımda $\boldsymbol{X}$ ile *k* örneklemin $\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2, \ldots \bar{\boldsymbol{x}}_k$ ortalama vektörleri karşılaştırılır ve birimi $\boldsymbol{X}$ 'e en yakın olan $\bar{\boldsymbol{x}}_i$ 'nın grubuna atanır. Burada, gruplar terimi alındıkları *k* örneklemi ya da *k* kitleyi ifade etmektedir.

Bu sınıflandırma tekniğini açıklamak için bazı örnekler verilebilir:

1. Bir üniversite kabul komitesi başarılı olma olasılığı ya da başarısız olma olasılığı olarak adayları sınıflamak isteyebilir. Eldeki değişkenler çeşitli alanlarındaki lise puanları, standartlaştırılmış test skorları, lisenin derecesi, ileri seviye derslerin sayısı, vs. olabilecektir.
2. Afrikalı, ya da ' katil' olarak adlandırılan arılar görsel olarak sıradan yerli bal arılarından ayrılamazlar. Onları kolaylıkla tanımlamak için kromatografi peak özelliğine dayanan on değişken kullanılabilmektedir (Lavine and Carlson, 1987).

3. Soygunları çözülebilir ya da çözülemez olarak sınıflandırılmak için parmak izinin varlığı, görgü tanıklarının varlığı ve polis varana kadarki zaman gibi değişkenler kullanılabilmektedir.
4. Bir takım değişkenler beş farklı hava istasyonunda ölçülebilir. Bu değişkenlere bağlı olarak özel bir havaalanında iki saat içindeki yükseklik sınırını tahmin etmek istenebilir. Yükseklik için değişkenler; kapalı, alçak enstrüman, yüksek enstrüman, alçak açık, yüksek açık şeklinde kategorilere ayrılabilmektedir.(Lachenbruch, 1975).

Gözlemleri gruplara atamak üzere her grup için bir sınıflandırma fonksiyonu oluşturulur. Yeni gözlemlerin gruplara atanması için de bu sınıflandırma fonksiyonları kullanılır. Sınıflandırma fonksiyonları, yeni gözlemler için grup üyeliğini tahmin etmek ve aynı örneklemde çapraz geçerlilik (holdout) yöntemi ile gözlemleri sınıflandırmanın yeterliliğini kontrol etmek için kullanılırlar.

2.8.1. İki Gruba Sınıflandırma

İki kitle olması durumunda, bir örnekleme birimi iki kitleden birine sınıflandırılabilir. Mevcut olan bilgi örnekleme birimi üzerinde ölçülen $\boldsymbol{X}$ gözlem vektöründeki *p* değişkeni içermektedir. 2.8. bölümünde verilen ilk örnekte, $\boldsymbol{X}$ 'de kaydedilmiş lise puanları ve farklı test skorları ile bir aday vardır. Bu adayın üniversitede başarılı olup olmayacağı bilinmemektedir, fakat başarılı olup olmadıkları bilinen üniversitedeki önceki öğrencilerin verileri mevcuttur. Başarılı olanlar için $\bar{\boldsymbol{x}}_1$ ve başarısız olanlar için $\bar{\boldsymbol{x}}_2$ ile $\boldsymbol{X}$ karşılaştırılarak, bu adayın sonunda hangi gruba ait olacağı tahmin etmeye çalışılır.

İki kitle olduğu zaman Fisher 'in (1936) lineer sınıflandırma yöntemi kullanılabilir. Fisher 'in yöntemi için normallik varsayımı gerekli değildir; ancak temel varsayım bu iki kitlenin aynı kovaryans $(\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2)$ matrisine sahip

olmasıdır. Her iki kitleden birer örneklem elde edilir ve $\bar{\boldsymbol{x}}_1$, $\bar{\boldsymbol{x}}_2$ ve $\boldsymbol{S}$ hesaplanır. Sınıflandırma için basit bir yöntem diskriminant fonksiyonuna dayalı olarak yapılabilir:

$$\boldsymbol{y} = \boldsymbol{b}'\boldsymbol{X} = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\boldsymbol{S}^{-1}\boldsymbol{X} . \tag{2.43}$$

Burada $\boldsymbol{X}$, iki gruptan (kitleden) birine sınıflamak istenen yeni örnekleme birimi üzerindeki ölçümlerin vektörüdür. $\boldsymbol{X}$'nin $\bar{\boldsymbol{x}}_1$'ya mı veya $\bar{\boldsymbol{x}}_2$'ya mı yakın olup olmadığını belirlemek için (2.43)'deki $\boldsymbol{y}$'nin $\bar{\boldsymbol{y}}_1$ dönüştürülmüş ortalamasına mı yoksa $\bar{\boldsymbol{y}}_2$ dönüştürülmüş ortalamasına mı daha yakın olduğu kontrol edilmelidir. İlk örneklemden her bir $\boldsymbol{x}_{1i}$ gözlemi için (2.43) hesaplanır ve $y_{11}, y_{12}, ..., y_{1n_1}$ elde edilir. Bu değerler kullanılarak $\bar{\boldsymbol{y}}_1 = \sum_{i=1}^{n_1} y_{1i} / n_1 = \boldsymbol{b}'\bar{\boldsymbol{x}}_1 = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\boldsymbol{S}^{-1}\bar{\boldsymbol{x}}_1$ ve benzer şekilde $\bar{\boldsymbol{y}}_2 = \boldsymbol{b}'\bar{\boldsymbol{x}}_2$ elde edilir. İki grup G_1 ve G_2 olarak gösterilsin. Fisher'in lineer sınıflandırma yöntemi eğer $\boldsymbol{y} = \boldsymbol{b}'\boldsymbol{X}$, $\bar{\boldsymbol{y}}_1$'ya $\bar{\boldsymbol{y}}_2$'dan daha yakınsa $\boldsymbol{X}$'i G_1 grubuna atar; $\boldsymbol{y}$, $\bar{\boldsymbol{y}}_2$'ya daha yakınsa $\boldsymbol{X}$'i G_2 grubuna atar. Bu durum Şekil 2.3.'te örneklendirilmiştir.

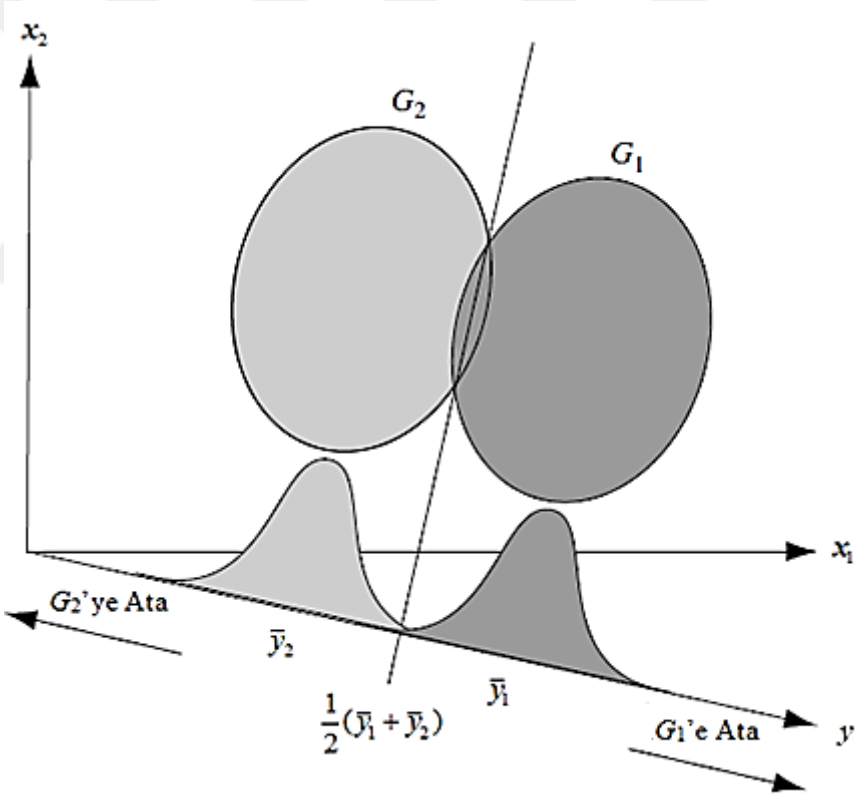
Şekil 2.3'deki yapı için,

$$\boldsymbol{y} > \frac{1}{2}(\bar{\boldsymbol{y}}_1 + \bar{\boldsymbol{y}}_2) \tag{2.44}$$

ise $\boldsymbol{y}$'nin $\bar{\boldsymbol{y}}_1$'ya daha yakın olduğu anlaşılır. Bu genelde doğrudur, çünkü $\bar{\boldsymbol{y}}_1$ her zaman $\bar{\boldsymbol{y}}_2$'dan daha büyüktür. Bu durum kolaylıkla aşağıdaki gibi gösterilebilir:

$$\bar{y}_1 - \bar{y}_2 = \boldsymbol{b}'(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) > 0, \quad (2.45)$$

çünkü $\boldsymbol{S}^{-1}$ pozitif tanımlıdır. Bu nedenle $\bar{y}_1 > \bar{y}_2$'dir ($\boldsymbol{b}$, $\boldsymbol{b}' = (\bar{\boldsymbol{x}}_2 - \bar{\boldsymbol{x}}_1)'\boldsymbol{S}^{-1}$ formunda olsaydı, o zaman $\bar{y}_2 - \bar{y}_1$ pozitif olurdu). $\frac{1}{2}(\bar{y}_1 + \bar{y}_2)$ orta nokta olduğundan, $y > \frac{1}{2}(\bar{y}_1 + \bar{y}_2)$ ifadesi $\boldsymbol{y}$'nin $\bar{y}_1$'ya daha yakın olduğunu gösterir. (2.45) ile $\bar{y}_1$'dan $\bar{y}_2$'ya olan mesafe $\bar{\boldsymbol{x}}_1$'dan $\bar{\boldsymbol{x}}_2$'ya olan mesafeyle aynıdır.



Şekil 2.3. İki gruba sınıflandırma için Fisher'in prosedürü (Rencher, 2002)

Sınıflandırma kuralını $\boldsymbol{X}$ açısından ifade etmek için, önce $\frac{1}{2}(\bar{y}_1 + \bar{y}_2)$

$$\frac{1}{2}(\bar{y}_1 + \bar{y}_2) = \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) \tag{2.46}$$

formunda yazılır. Bu durumda sınıflandırma kuralı şu hale gelir:

$$\boldsymbol{b}'\boldsymbol{X} = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\boldsymbol{S}^{-1}\boldsymbol{X} > \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) \tag{2.47}$$

ise, $\boldsymbol{X}$ G_1 grubuna atanır ve

$$\boldsymbol{b}'\boldsymbol{X} = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\boldsymbol{S}^{-1}\boldsymbol{X} < \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) \tag{2.48}$$

ise, $\boldsymbol{X}$ G_2 grubuna atanır.

Bu lineer sınıflandırma kuralı grupların tanımlayıcı ayrışımı ile bağlantılı olarak kullanılan aynı $\boldsymbol{y} = \boldsymbol{b}'\boldsymbol{X}$ diskriminant fonksiyonunu kullanır. Bu yüzden iki-grup durumundaki diskriminant fonksiyonu lineer sınıflandırma fonksiyonu olarak da görev yapar. Bu lineer fonksiyon aracılığıyla yapılan sınıflandırma analizi lineer diskriminant analizi (LDA) olarak adlandırılır. Ancak ikiden fazla grup durumunda, tanımlayıcı diskriminant fonksiyonlarından farklı olan sınıflandırma fonksiyonları kullanılır.

Dağılımsal varsayımlar yapılmadığı için Fisher'in (2.47) ve (2.48)'yı kullanan yaklaşımı aslında parametrik değildir. Fakat iki kitle eşit kovaryans matrisleriyle normal ise, o zaman bu yöntem asimptotik olarak en uygun olanıdır. Yani, hatalı sınıflandırma olasılığı minimize edilmiştir.

π_1 ve π_2 önsel olasılıkları iki kitle için biliniyorsa, sınıflandırma kuralı bu ek bilgiden yararlanmak için Bayes sınıflandırma kuralına dönüştürülebilir.

Çok değişkenli normal dağılım, bireysel olarak her bir tahmin edicinin her bir çift tahmin edici arasındaki bir miktar korelasyon ile tek boyutlu normal dağılıma uyduğunu varsayar. *p*-boyutlu $\boldsymbol{X}$ rasgele değişkeninin çok değişkenli normal dağılımına sahip olduğunu göstermek için, $\boldsymbol{X} \sim \mathrm{N_p}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ yazılır. Burada $\mathrm{E}(\boldsymbol{X}) = \boldsymbol{\mu}$, $\boldsymbol{X}$ (*p*-boyutlu bir vektör)'in ortalaması ve $\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}$, $\boldsymbol{X}$ 'in $p \times p$ kovaryans matrisidir.

Çok değişkenli normal yoğunluk fonksiyonu,

$$f(\mathrm{x}) = \frac{1}{(2\pi)^{p/2}\left|\boldsymbol{\Sigma}\right|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})\right) \tag{2.49}$$

olarak tanımlanır.

p tahmin edici durumunda LDA sınıflandırıcısı, *k*_ıncı gruptaki gözlemlerin çok değişkenli bir normal dağılım $N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$'dan çekildiğini varsayar. Burada $\boldsymbol{\mu}_k$ gruba özgü ortalama vektörü ve $\boldsymbol{\Sigma}$, tüm *k* grup için ortak olan kovaryans matrisidir.

π_k, *k*_ıncı gruptan gelen rasgele seçilen bir gözlemin genel ya da önsel olasılığını ve $f_k(\mathrm{x}) = \mathrm{P}(\boldsymbol{X} = \mathrm{x} \mid \boldsymbol{y} = k)$, *k*_ıncı gruptan gelen bir gözlem için *x*'in yoğunluk fonksiyonunu göstersin.

Bu durumda Bayes teoremi

$$\mathrm{P}(\boldsymbol{y} = k \mid \boldsymbol{X} = \mathrm{x}) = \frac{\pi_k f_k(\mathrm{x})}{\sum_{l=1}^{k} \pi_l f_l(\mathrm{x})} \tag{2.50}$$

olarak ifade edilir. $p_k(\mathrm{x}) = \mathrm{P}(\boldsymbol{y} = k \mid \boldsymbol{X} = \mathrm{x})$ kısaltması yapılarak, k_ıncı grup için yoğunluk fonksiyonu $f_k(\mathrm{x})$ (2.50)'de yerine konulursa Bayes teoremi

$$p_k(x) = \frac{\pi_k \dfrac{1}{(2\pi)^{p/2}\left|\boldsymbol{\Sigma}\right|^{1/2}} \exp\left(-\dfrac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu}_k)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_k)\right)}{\sum_{l=1}^{k} \pi_l \dfrac{1}{(2\pi)^{p/2}\left|\boldsymbol{\Sigma}\right|^{1/2}} \exp\left(-\dfrac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu}_l)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_l)\right)} \qquad (2.51)$$

olarak elde edilir. Bayes sınıflandırıcısı, $\boldsymbol{X}$ gözlemini (2.51) değeri en büyük olan gruba atamayı gerektirir. Denk olarak Bayes sınıflandırıcısı $\boldsymbol{X}$ gözlemini, (2.51)'in logaritmasını alarak ve terimleri yeniden düzenleyerek elde edilen

$$\delta_k(\mathrm{x}) = \boldsymbol{\mu}_k'\boldsymbol{\Sigma}^{-1}\boldsymbol{X} - \frac{1}{2}\boldsymbol{\mu}_k'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \ln \pi_k \qquad (2.52)$$

değeri en büyük olan gruba atamaktadır.

Önsel olasılıklar π_1; G_1'deki gözlemlerin oranıdır, π_2; G_2'deki gözlemlerin oranıdır ve burada $\pi_2 = 1-\pi_1$'dir. Örneğin, belli bir üniversiteye yeni girenlerin 80%'inin mezun olduğu varsayıldığında, $\pi_1 = 0.8$ ve $\pi_2 = 0.2$ olur. Önsel olasılıkları kullanmak için, iki kitlenin $f(\boldsymbol{X} = \mathrm{x} \mid \boldsymbol{y} = G_1) = f_1(\mathrm{x})$ ve $f(\boldsymbol{X} = \mathrm{x} \mid \boldsymbol{y} = G_2) = f_2(\mathrm{x})$ yoğunluk fonksiyonları da bilinmelidir. O zaman Welch (1939)'e göre hatalı sınıflandırma olasılığını en aza indiren en uygun (optimum) sınıflandırma kuralı şöyledir:

$$\pi_1 f(\boldsymbol{X} = \mathrm{x} \mid \boldsymbol{y} = G_1) > \pi_2 f(\boldsymbol{X} = \mathrm{x} \mid \boldsymbol{y} = G_2), \qquad (2.53)$$

diğer bir ifadeyle

$$\pi_1 f_1(\mathrm{x}) > \pi_2 f_2(\mathrm{x})$$

ise $\boldsymbol{X}$ G_1 grubuna atanır ve tersi ise $\boldsymbol{X}$ G_2 grubuna atanır. G_1 tarafından temsil edilen kitleden örneklem alındığında, $f_1(\mathrm{x})$ yoğunluk için uygun bir gösterimdir ve alışılmış anlamda koşullu bir dağılımı temsil etmemektedir.

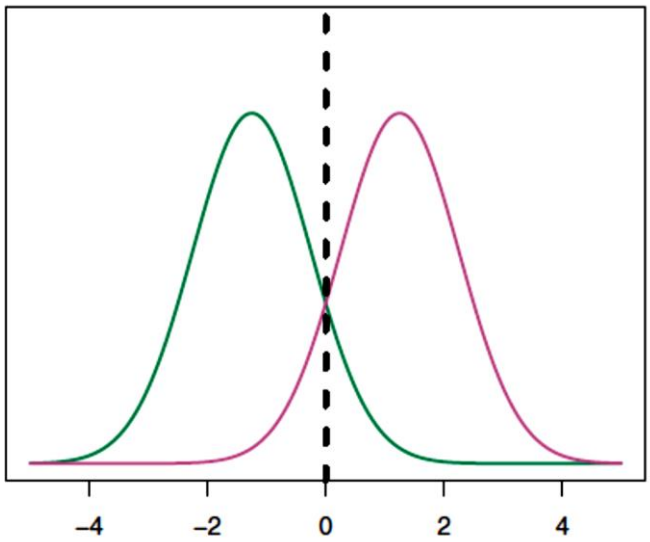
İki yoğunluğun eşit kovaryans matrisli çok değişkenli normal olduğu varsayılarak, yani $f(\boldsymbol{X} = \mathrm{x} \mid \boldsymbol{y} = G_1) = N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ ve $f(\boldsymbol{X} = \mathrm{x} \mid \boldsymbol{y} = G_2) = N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ ise o zaman (2.50)'den ($\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ ve $\boldsymbol{\Sigma}$ yerine tahminleri ile) aşağıdaki kural elde edilir:

$$\boldsymbol{b}'\boldsymbol{X} = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\boldsymbol{S}^{-1}\boldsymbol{X} > \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) + \ln\left(\frac{\pi_2}{\pi_1}\right) \qquad (2.54)$$

ise $\boldsymbol{X}$ G_1 grubuna atanır ve tersi ise $\boldsymbol{X}$ G_2 grubuna atanır. Parametrelerin yerini tahminleri kullanıldığından, (2.54)'deki kural (2.53)'deki gibi artık optimal değildir. Ancak bu kural asimptotik olarak optimaldir, yani örneklem büyüklüğü arttıkça optimalliğe yaklaşır.

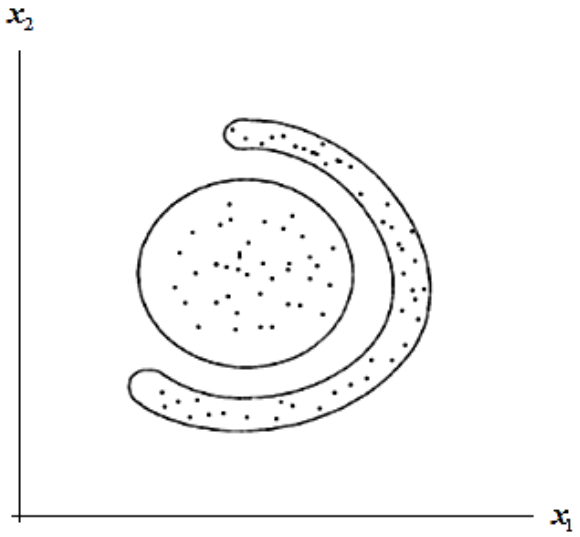
$p = 1$ boyutlu iki normal yoğunluk fonksiyonu $f_1(\mathrm{x})$ ve $f_2(\mathrm{x})$, iki farklı grubu temsil etsin. İki yoğunluk fonksiyonu için ortalama ve varyans parametreleri $\mu_1 = -1.25$, $\mu_2 = 1.25$ ve $\sigma_1^2 = \sigma_2^2 = 1$ hesaplanmış olsun. İki yoğunluk çakışır (üst üste gelir) ve bu yüzden x verilmişken gözlemlerin ait olduğu grup hakkında bazı belirsizlikler söz konusudur. Bir gözlemin her iki gruptan eşit olasılıkla geldiği varsayılırsa, yani $\pi_1 = \pi_2 = 0.5$ alınırsa, Bayes sınıflandırıcısı $\boldsymbol{b}'\boldsymbol{X} < 0$ ise gözlemi G_1 grubuna ve diğer durumda G_2 grubuna atar. Bu örnek Şekil 2.4.'de

gösterilmektedir. Bu durumda, $\boldsymbol{X}$'in her bir grup içinde normal dağılımından çekildiği ve ilişkili tüm parametreler bilindiği zaman Bayes sınıflandırıcısı hesaplanabilir. Ancak bir gerçek yaşam durumunda, Bayes sınıflandırıcısı hesaplanamaz.



Şekil 2.4. İki tek-boyutlu normal yoğunluk fonksiyon. Kesikli dikey çizgi Bayes karar sınırını temsil etmektedir (James, Witten, Hastie ve Tibshirani, 2015)

$\pi_1 = \pi_2$ ise, (2.51)'deki normalliğe dayanan Bayes karar kuralı (2.47) ve (2.48)'da verilen Fisher'in yöntemi ile aynı olur. Bu nedenle normallik varsayımına dayanmayan Fisher'in kuralı, veri $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ ve $\pi_1 = \pi_2$ ile çok değişkenli normal kitleden geldiğinde en uygun özelliklere sahiptir. Bundan dolayı Fisher'in yöntemi parametrik olmamasına rağmen, normal olarak dağılmış kitleler ya da lineer eğilimli diğer kitleler için daha iyi çalışmaktadır.



Şekil 2.5. Lineer olmayan ayrışımlı iki kitle (Rencher,2002)

Şekil 2.5.'te lineer eğilimli olmayan iki kitlenin ayrışımı gösterilmektedir. Bu örnekteki gibi %95 kontüre sahip iki kitle varsayıldığında, noktalar düz bir çizgi üzerinde herhangi bir yönde yansırsa neredeyse tamamı çakışacaktır. Böyle bir durumda lineer bir diskriminant yöntemi iki kitleyi başarılı bir şekilde ayıramayacaktır.

2.8.2. İkiden Fazla Gruba Sınıflandırma

İki grup durumundaki gibi, $\bar{x}_1, \bar{x}_2, ..., \bar{x}_k$ örneklem ortalama vektörlerini bulmak için *k* grubun her birinden bir örneklem kullanılır. Grup üyeliği bilinmeyen $\boldsymbol{X}$ vektörü için bir yaklaşım $\boldsymbol{X}$ 'e en yakın olan ortalama vektörü bulmak için bir uzaklık fonksiyonu kullanmaktır ve karşılık gelen gruba $\boldsymbol{X}$ atanır. İkiden fazla grup olması durumunda tanımlayıcı diskriminant fonksiyonlarından farklı olan sınıflandırma fonksiyonları kullanılır. Verilerin normal dağıldığı ancak, grupların kitle varyans-kovaryans matrislerinin eşit ya da farklı olmaları durumuna göre kullanılan sınıflandırma fonksiyonları ele alınır.

2.8.2.1. Lineer Sınıflandırma Fonksiyonları

Birimlerin *k* tane normal dağılımlı ve ortak varyans-kovaryans matrise sahip kitleden gelmesi durumu için geliştirilen yöntem, Fisher'in iki grubun ayrımında kullanılan yönteminin devamı biçimindedir. Anderson (1962) tarafından geliştirilen bu yöntem *k* grubun birbiri ile karşılaştırılması esasına dayanır.

$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \cdots = \boldsymbol{\Sigma}_k$ eşitliği varsayılır. Birleştirilmiş örneklem kovaryans matrisi ile ortak kitle kovaryans matrisi

$$\boldsymbol{S} = \frac{1}{N-k}\sum_{i=1}^{k}(n_i - 1)\boldsymbol{S}_i = \frac{\boldsymbol{E}}{N-k} \tag{2.55}$$

tahmin edilir. Burada n_i ve $\boldsymbol{S}_i$, sırasıyla örneklem büyüklüğü ve *i*_inci grubun kovaryans matrisidir, $\boldsymbol{E}$ tek-yönlü MANOVA'dan elde edilen hata matrisi ve $N = \sum_i n_i$ 'dir. Her bir $\bar{\boldsymbol{x}}_i$, $i = 1, 2, ..., k$ için

$$D_i^2(\boldsymbol{X}) = (\boldsymbol{X} - \bar{\boldsymbol{x}}_i)'\boldsymbol{S}^{-1}(\boldsymbol{X} - \bar{\boldsymbol{x}}_i) \tag{2.56}$$

uzaklık fonksiyonu ile $\boldsymbol{X}$ karşılaştırılır ve $\boldsymbol{X}$, $D_i^2(\boldsymbol{X})$ 'nin en küçük olduğu gruba atanır.

Genişletilen (2.54) ile, LDA için genişletilmiş bir lineer bir sınıflandırma kuralı elde edilir:

$$\begin{aligned} D_i^2(\boldsymbol{X}) &= \boldsymbol{X}'\boldsymbol{S}^{-1}\boldsymbol{X} - \boldsymbol{X}'\boldsymbol{S}^{-1}\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}}_i'\boldsymbol{S}^{-1}\boldsymbol{X} + \bar{\boldsymbol{x}}_i'\boldsymbol{S}^{-1}\bar{\boldsymbol{x}}_i \\ &= \boldsymbol{X}'\boldsymbol{S}^{-1}\boldsymbol{X} - 2\bar{\boldsymbol{x}}_i'\boldsymbol{S}^{-1}\boldsymbol{X} + \bar{\boldsymbol{x}}_i'\boldsymbol{S}^{-1}\bar{\boldsymbol{x}}_i . \end{aligned} \tag{2.57}$$

(2.57)'in sağ kısımdaki $\boldsymbol{X}'\boldsymbol{S}^{-1}\boldsymbol{X}$ terimi, *i*'nin bir fonksiyonu olmadığı ve bu nedenle gruptan gruba değişmeyeceği için ihmal edilebilmektedir. İkinci terim $\boldsymbol{X}$ 'in lineer bir fonksiyonudur ve üçüncü terim $\boldsymbol{X}$ 'i içermemektedir. Bu nedenle $\boldsymbol{X}'\boldsymbol{S}^{-1}\boldsymbol{X}$ silinir ve $L_i(\boldsymbol{X})$ ile gösterilen lineer sınıflandırma fonksiyonu

$$L_i(\boldsymbol{X}) = -2\bar{\boldsymbol{x}}_i'\boldsymbol{S}^{-1}\boldsymbol{X} + \bar{\boldsymbol{x}}_i'\boldsymbol{S}^{-1}\bar{\boldsymbol{x}}_i$$

elde edilir. Bu eşitlik, normal dağılım ve önsel olasılıklara dayanan Bayes kuralını kullanabilmek için $-\frac{1}{2}$ ile çarpılır, bu durumda lineer sınıflandırma kuralı şu şekilde olur: $\boldsymbol{X}$,

$$L_i(\boldsymbol{X}) = \bar{\boldsymbol{x}}_i'\boldsymbol{S}^{-1}\boldsymbol{X} - \frac{1}{2}\bar{\boldsymbol{x}}_i'\boldsymbol{S}^{-1}\bar{\boldsymbol{x}}_i, \qquad i = 1,2,...,k \tag{2.58}$$

değeri maksimum olan gruba atanır. $\boldsymbol{X}$ 'nin bir fonksiyonu olarak (2.58) eşitliğinin lineerliğini vurgulamak için

$$L_i(\boldsymbol{X}) = \boldsymbol{c}_i'\boldsymbol{X} + c_{i0} = c_{i1}\,\mathrm{x}_1 + c_{i2}\,\mathrm{x}_2 + \cdots + c_{ip}\,\mathrm{x}_p + c_{i0}$$

olarak ifade edilebilir. Burada $\boldsymbol{c}_i' = \bar{\boldsymbol{x}}_i'\boldsymbol{S}^{-1}$ ve $c_{i0} = -\frac{1}{2}\bar{\boldsymbol{x}}_i'\boldsymbol{S}^{-1}\bar{\boldsymbol{x}}_i$ 'dir. Bu yöntemi kullanarak $\boldsymbol{X}$ 'i bir gruba atamak için *k* grubun her biri için $\boldsymbol{c}_i$, c_{i0} ve $L_i(\boldsymbol{x})$, $i = 1,2,...,k$ hesaplanır ve $\boldsymbol{X}$, $L_i(\boldsymbol{X})$ 'si en büyük olan gruba atanır. Bu (2.56)'deki $D_i^2(\boldsymbol{X})$ 'si en küçük olan, yani $\boldsymbol{X}$ 'ye en yakın olan $\bar{\boldsymbol{x}}_i$ ortalama vektörünün grubu ile aynı olacaktır.

İkiden fazla grup durumuna göre (2.53)'deki optimum sınıflandırma kuralı genişletildiğinde $\boldsymbol{X}$, $\pi_i f(\boldsymbol{X} = \mathrm{x} \mid y = G_i)$'yi maksimum yapan gruba atanır. Bu kural ile yanlış sınıflandırma olasılığı minimize edilir. Eşit kovaryans matrisleri ve grup üyelerinin $\pi_1, \pi_2, ..., \pi_k$ önsel olasılıkları ile normallik varsayımı yapılırsa, o zaman $f(\boldsymbol{X} = \mathrm{x} \mid y = G_i) = N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ olur ve parametrelerin yerine tahminleri kullanılarak genişletilen kural

$$L_i'(\boldsymbol{X}) = \ln \pi_i + \bar{\boldsymbol{x}}_i' \boldsymbol{S}^{-1} \boldsymbol{X} - \frac{1}{2} \bar{\boldsymbol{x}}_i' \boldsymbol{S}^{-1} \bar{\boldsymbol{x}}_i, \qquad i = 1, 2, ..., k \tag{2.59}$$

olarak elde edilir ve $\boldsymbol{X}$ $L_i'(\boldsymbol{X})$'in maksimum değerli grubuna atanır. $\pi_1 = \pi_2 = ... = \pi_k$ ise, bu durumda normal dağılım için sınıflandırma oranını optimize eden (2.59), $\boldsymbol{X}$'in $\bar{\boldsymbol{x}}_i$'ya uzaklığını minimize eden sezgisel yaklaşıma dayalı olan (2.58)'a indirgenir.

Literatürde, (2.58)'da tanımlanmış olan $L_i(\boldsymbol{X})$ lineer sınıflandırma fonksiyonları lineer diskriminant fonksiyonları (LDF) olarak da tanımlanır. Ancak onlar, katsayıları $\boldsymbol{E}^{-1}\boldsymbol{H}$'ın özvektörleri olan lineer diskriminant fonksiyonlarından farklıdır. Aslında, k-tane sınıflandırma fonksiyonu ve $s = \min(p, k\text{-}1)$-tane diskriminant fonksiyonu olacaktır. Burada k grup sayısı ve p değişken sayısıdır. Pek çok durumda grup farklılıklarını etkili bir şekilde tanımlamak için tüm s diskriminant fonksiyonlarına ihtiyaç duyulmaz, buna karşılık tüm k sınıflandırma fonksiyonlarının gözlemleri gruplara atamada kullanılmaları gerekir (Rencher,2002).

2.8.2.2. Karesel Sınıflandırma Fonksiyonları

İkiden fazla grup durumunda lineer sınıflandırma fonksiyonları, her bir grup içindeki gözlemlerin gruba ait ortalama vektörü ve tüm *k* sınıf için ortak olan bir kovaryans matrisi ile çok değişkenli normal dağılımdan çekildiğini varsaymaktadır. Ancak, uygulamada çoğu zaman meydana gelen sınıflandırma fonksiyonları kovaryans matrislerinin heterojenliğine karşı duyarlıdır. Gözlemler sıklıkla kovaryans matrislerindeki köşegen üzerinde daha büyük varyanslara sahip olan gruplara sınıflandırılma eğilimindedir. Bu nedenle eğer şüphelenmek için bir neden varsa kitle kovaryans matrisleri eşit varsayılmamalıdır.

$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \cdots = \boldsymbol{\Sigma}_k$ eşitliği kabul edilmezse, sınıflandırma kuralları sınıflandırma oranlarının optimalliğini korumak için kolaylıkla değiştirilebilir. (2.56)'nın yerine,

$$D_i^2(\boldsymbol{X}) = (\boldsymbol{X} - \bar{\boldsymbol{x}}_i)' \boldsymbol{S}_i^{-1} (\boldsymbol{X} - \bar{\boldsymbol{x}}_i), \qquad i = 1, 2, ..., k \tag{2.60}$$

kullanılabilir. Burada $\boldsymbol{S}_i$, *i*_inci grup için örneklem kovaryans matrisidir. Önceden olduğu gibi, $\boldsymbol{X}$, $D_i^2(\boldsymbol{X})$'si en küçük olan gruba atanır. $\boldsymbol{S}$ yerine $\boldsymbol{S}_i$ ile (2.60), (2.58)'deki gibi $\boldsymbol{X}$'nin lineer bir fonksiyonuna indirgenemez ancak karesel bir fonksiyon olarak kalır. Bu nedenle $\boldsymbol{S}_i$'ye dayanan fonksiyonlar karesel sınıflandırma fonksiyonları (QDF) olarak adlandırılır. Bu sınıflandırma fonksiyonları ile yapılan analiz ise karesel diskriminant analizi (QDA) olarak adlandırılır.

Eşit olmayan kovaryans matrisleri ve $\pi_1, \pi_2, ..., \pi_k$ önsel olasılıkları ile normallik varsayımı yapılırsa, o zaman $f(\boldsymbol{X} = \mathrm{x} \mid \boldsymbol{y} = G_i) = N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ ve $\pi_i f(\boldsymbol{X} = \mathrm{x} \mid \boldsymbol{y} = G_i)$'ye dayanan optimal karesel sınıflandırma kuralı,

$$Q_i(\boldsymbol{X}) = \ln \pi_i - \frac{1}{2}\ln|\boldsymbol{S}_i| - \frac{1}{2}(\boldsymbol{X} - \bar{\boldsymbol{x}}_i)'\boldsymbol{S}_i^{-1}(\boldsymbol{X} - \bar{\boldsymbol{x}}_i) \quad (2.61)$$

olarak elde edilir ve $\boldsymbol{X}$, $Q_i(\boldsymbol{X})$ değeri maksimum olan gruba atanır. $\pi_1 = \pi_2 = ... = \pi_k$ ise ya da π_i'ler bilinmiyorsa, $\ln \pi_i$ terimi silinir.

$\boldsymbol{S}_i$'ye dayanan bir karesel sınıflandırma kuralı kullanmak için, her bir n_i, p'den daha büyük olmalıdır ki böylece $\boldsymbol{S}_i^{-1}$ mevcut olsun. Bu kısıtlama $\boldsymbol{S}$'ye dayanan lineer sınıflandırma kuralları için geçerli değildir. Karesel sınıflandırma fonksiyonları ile daha fazla parametre tahmin edildiği için, n_i'nin daha büyük değerleri tahminlerin durağanlığı için gereklidir.

Neden k grubun ortak bir kovaryans matrisi paylaştığını varsayıp saymamamız önemlidir? Diğer bir deyişle, neden LDA yerine QDA, ya da tam tersi tercih edilir? Bu soruların cevabı yanlı-varyans durumunda yatmaktadır. p tahmin edici olduğunda, bir kovaryans matrisi tahmin etmek $p(p+1)/2$ parametre tahmin edilmesini gerektirir. QDA sınıflandırıcısı her bir gruptaki toplam $kp(p+1)/2$ parametre için, ayrı bir kovaryans matrisi tahmin eder. Örneğin, 50 tahmin edici ile bu sayı 1275'in herhangi bir katı kadar çok fazla parametre olacaktır. k grubun ortak bir kovaryans matrisli olduğu varsayımının kullanıldığı yerde LDA sınıflandırma modeli x'e göre lineer olur, bunun anlamı tahmin etmek için kp tane lineer katsayı olduğudur. Sonuç olarak LDA sınıflandırıcısı, QDA sınıflandırıcısına göre daha az esnek bir sınıflandırıcıdır ve bu nedenle daha küçük bir varyansa sahiptir. Bu durum aslında potansiyel olarak tahminin performansının gelişmesine neden olabilir. Fakat LDA sınıflandırıcısının k grubun ortak bir kovaryans matrisi paylaştığı varsayımı sıkıntılı ise, bu durumda LDA yüksek yanlılıktan dolayı yanlış bir tercih olabilir. Örneğin, nispeten birkaç eğitim gözlemi varsa ve bu nedenle azalan varyans önemli ise, LDA QDA'dan

daha iyi bir tercih olma eğilimindedir. Bunun tersine, eğitim seti çok büyükse bu nedenle sınıflandırıcının varyansı önemli bir endişe değilse ya da *k* grup için ortak kovaryans matrisi varsayımı açıkça sağlanmıyor ise QDA tavsiye edilir (James; Witten; Hastie; Tibshirani, 2015).

2.9. Tahmin Edilen Hatalı Sınıflandırma Oranları

Bilinmeyen bir gözlem herhangi bir kurala göre sınıflandırıldığında, her zaman o gözlemi hatalı sınıflandırma olasılığı vardır. Bu durum herhangi bir istatistiksel yöntemin doğasında olan belirsizliğin bir parçasıdır. Diskriminant analizi fonksiyonlarının grup ayrışımındaki etkisi, önem testi kullanarak ya da $\lambda_i / \sum_j \lambda_j$ yi inceleyerek değerlendirilir. Grup üyeliği tahmin etmede kullanılan sınıflandırma yöntemlerinin yeterliliğini değerlendirmek için de genel olarak hatalı sınıflandırma oranı kullanılır. Bu oran literatürde genel olarak yanlış sınıflandırma olasılığı olarak da kullanılır. Ayrıca onun tamamlayıcısı olan doğru sınıflandırma oranı da kullanılabilmektedir.

Hatalı sınıflandırma oranının basit bir tahmini, sınıflandırma fonksiyonlarının hesaplanmasında kullanılan aynı veri seti üzerinde sınıflandırma yönteminin denenmesiyle elde edilebilir. Bu yöntem genellikle tekrar yerine koyma metodu olarak bilinir. Her bir gözlem vektörü $\boldsymbol{x}_{ij}$, sınıflandırma fonksiyonlarını temsil eder ve bir gruba atanır. Bu durumda doğru sınıflandırmaların sayısı ve yanlış sınıflandırmaların sayısı hesaplanır. Tekrar yerine koymadan kaynaklanan yanlış sınıflandırma oranı hata oranı olarak adlandırılır. Sonuçlar sınıflandırma tablosunda veya hata (confusion) matrisinde gösterilir. Sınıflandırma tablosu, diskriminant fonksiyonunun veri setindeki her bir gözlemi nasıl sınıflandıracağını açıklar. Sınıflandırma tablolarından doğru bir şekilde sınıflandırılan gözlemlerin yüzdesi ile sınıflandırma hatalarının niteliği ve sayısı bulunur.

Her grupta eşit sayıda gözlem bulunduğunda, sınıflandırma süreci tarafından doğru bir şekilde sınıflandırılan gözlem oranı ile şans eseri doğru sınıflandırılması

gereken gözlem oranını karşılaştırmak kolaydır. İki adet eşit büyüklükte grup varsa, gözlemlerin sadece %50'si doğru bir şekilde sınıflandırılır. Bu durumda gözlemler gruplara yansız atanır ve her gruptaki atamaların yarısı doğru olur. Üç eşit büyüklükte grup varsa, doğru sınıflandırma %33 olur ve oranlar grup sayısına göre bu şekilde devam eder. Ancak gruplardaki gözlem sayıları eşit olmadığında, doğru sınıflandırılması gereken gözlem oranlarının hesaplanması biraz daha karmaşıktır.

İki grup için sınıflandırma tablosu aşağıdaki formda olur:

Tablo 2.1. İki grup için sınıflandırma tablosu

Gerçek Grup	Gözlem Sayısı	Tahmin Edilen Grup 1	Tahmin Edilen Grup 2
1	n_1	n_{11}	n_{12}
2	n_2	n_{21}	n_{22}

Tablo 2.1.'e göre; Grup 1'deki n_1 gözlemleri arasında, $n_1 = n_{11} + n_{12}$ iken, n_{11} tanesi Grup 1'e doğru bir şekilde sınıflandırılmıştır ve n_{12} tanesi Grup 2'ye yanlış sınıflandırılmıştır. Benzer şekilde Grup 2'deki n_2 gözlemin, $n_2 = n_{21} + n_{22}$ iken, n_{21} tanesi Grup 1'e yanlış sınıflandırılmıştır ve n_{22} tanesi Grup 2'ye doğru bir şekilde sınıflandırılmıştır. Böylece,

$$\text{hata oranı} = \frac{n_{12} + n_{21}}{n_1 + n_2}$$

$$= \frac{n_{12} + n_{21}}{n_{11} + n_{12} + n_{21} + n_{22}} = \frac{n_{12} + n_{21}}{n}$$

ile tahmin edilir. Benzer olarak,

$$\text{doğru sınıflandırma oranı} = \frac{n_{11} + n_{22}}{n}$$

şeklinde ifade edilir.

İkiden fazla grup için sınıflandırma tablosu aşağıdaki formda olur:

Tablo 2.2. İkiden fazla grup için sınıflandırma tablosu

Gerçek Grup	Sınıflandırılan Grup 1	2	...	k	Toplam
1	n_{11}	n_{12}	...	n_{1k}	$n_{1.}$
2	n_{21}	n_{22}	...	n_{2k}	$n_{2.}$
⋮	⋮	⋮	⋮	⋮	⋮
k	n_{k1}	n_{k2}	...	n_{kk}	$n_{k.}$
Toplam	$n_{.1}$	$n_{.2}$	...	$n_{.k}$	$n_{..}$

Tablo 2.2.'ye göre, satırlar 1'den *k*'ye kadar, gözlemlerin gerçekten ait olduğu *k* grubu gösterir. Sütunlar *k* gruba yapılan tahmini sınıflandırmaları gösterir. n_{11}, 1. gruptaki doğru sınıflandırılmış gözlem sayısıdır. Ancak n_{12}, yanlışlıkla 2. Gruba sınıflandırılmış gözlem sayısını gösterir. Bu durumda $n_{ij} = i_$inci gruba ait olup $j_$inci gruba sınıflandırılanların sayısıdır. İdeal koşullarda bu hata matrisi, köşegen bir matris olur; uygulamada köşegen-dışı elemanların çok küçük sayılar olarak alınması beklenir.

Satır toplamları veri setindeki her bir kitleye ya da gruba ait olan gözlemlerin sayısını, sütun toplamları ise bu grupların her birine sınıflandırılmış olan gözlemlerin sayısını verir. Veri setindeki toplam gözlem sayısı $n_{..}$'dır. Buradaki nokta gösterimi, satır toplamlarındaki ikinci alt indis üzerindeki toplamı

göstermek için ve sütun toplamlarındaki birinci alt indisi üzerindeki toplamı göstermek için kullanılır.

$\boldsymbol{y}_j$ grubundaki bir birimin $\boldsymbol{y}_i$ kitlesine sınıflandırılması olasılığı

$$p(i \mid j)$$

ile gösterilsin. Bu hatalı sınıflandırma olasılıkları, *i* grubuna hatalı sınıflandırılan *j* grubundan alınan gözlem sayısının, *j* grubundan alınan örneklemdeki toplam gözlem sayısına bölünmesiyle

$$\hat{p}(i \mid j) = \frac{n_{ij}}{n_{j.}}$$

ile tahmin edilir. Bu tahmin hatalı sınıflandırma oranlarını verir.

Hatalı sınıflandırma oranı kolayca elde edilebilmekte ve birçok sınıflandırma yazılım programları tarafından rutin olarak verilmektedir. Bu oran gelecekteki bir gözlemi yanlış sınıflandıracak olan mevcut örnekleme dayalı sınıflandırma fonksiyonlarımızın bir olasılık tahminidir. Bu olasılığa gerçek hata oranı denir. Ne yazık ki hatalı sınıflandırma oranı gerçek hata oranını eksik tahmin eder çünkü sınıflandırma fonksiyonlarının hesaplanmasında kullanılan veri seti aynı zamanda onların değerlendirilmesinde de kullanılır. Hatalı sınıflandırma oranındaki bu yanlılığı azaltmaya yönelik geliştirilen bazı tahmin yaklaşımları mevcuttur (Rencher, 2002.)

2.10. Hata Oranlarının Geliştirilmiş Tahminleri

Büyük örneklemler için tahmin edilen hata oranı, gerçek hata oranı için sadece küçük bir miktarda yanlılığa sahiptir ve hata oranı küçük bir şüphe ile kullanılabilir. Ancak küçük örneklemler için bu oran yanlılığa neden olur. Hata

oranındaki yanlılığı azaltmak için, yani hata oranını daha gerçekçi bir seviyeye çıkarmak için iki yöntem ele alınır.

2.10.1. Örneklemi Parçalama

Yanlılığı önlemek amacıyla örneklem, bir sınıflandırma kuralı oluşturmak için kullanılan deneme örneklemi ve onu değerlendirmek için kullanılan bir kontrol örneklemi olarak rasgele iki parçaya ayrılır. Deneme örneklemi ile lineer ve karesel sınıflandırma fonksiyonları hesaplanır. Bu işlem sonunda, deneme örnekleminden elde edilen sınıflandırma fonksiyonları ile kontrol örneklemindeki her bir gözlem vektörü temsil edilmiş olur. Deneme örneklemindeki gözlemlerden hatalı sınıflandırılan gözlemlerin oranı hesaplanabilir.

Bu gözlemler sınıflandırma fonksiyonlarının hesaplanmasında kullanılmadığı için elde edilen hata oranı yansızdır. Ancak bu yöntem, verinin en iyi şekilde kullanılmasını sağlamaz ve bu nedenle tahmin edilen hatalı sınıflandırma olasılıkları genellikle kesin değildir.

Rencher'e (2002) göre örneklemi parçalamanın en az iki dezavantajı vardır. Birincisi; bu yöntem elde olmayan büyük örneklemler gerektirir. İkincisi; uygulamada kullanılan sınıflandırma fonksiyonlarını değerlendirmez. Örneklemin yarısına dayalı olan hata tahmini bütün örnekleme dayalı olana göre önemli ölçüde farklı olabilir. Hata oranı tahmininin varyansını minimize etmek amacıyla verilerin tamamını ya da tamamına yakınını sınıflandırma fonksiyonları oluşturmak için kullanmak tercih edilir.

2.10.2. Holdout Yöntemi: Çapraz Doğrulama

Holdout yöntemi örneklemi parçalama yönteminin gelişmiş bir halidir. Bu yöntem bir-tanesini-dışarda-bırakma yöntemi ya da çapraz doğrulama olarak da bilinmektedir. Holdout yöntemi hata oranlarını tahmin etmek için kullanılır. Bu yöntemde, biri hariç bütün gözlemler kullanılarak sınıflandırma kuralı hesaplanır

ve bu kural daha sonra ihmal edilen gözlemi sınıflandırmak için kullanılır. Her bir gözlem için bu yöntem tekrarlanır ve hatalı sınıflandırılan gözlemlerin oranını hesaplanır.

Holdout yönteminde $N = \sum_i n_i$ örneklem büyüklüğündeki her bir gözlem diğer $N-1$ gözleme dayalı fonksiyon tarafından sınıflandırılır. Bu yöntemde N farklı sınıflandırma yöntemi oluşturulması gerektiği için hesaplama yükü artar. Ancak, gelecekteki yeni gözlemler için gerçek sınıflandırma kuralı bütün N gözleme dayalı olmalıdır.

3. LOJİSTİK REGRESYON ANALİZİ

Lojistik regresyon analizi, gözlemlerin gruplara atanmasında sık kullanılan üç yöntemden (diğerleri kümeleme analizi ve diskriminant analizi) birisidir. Kümeleme analizinde gözlemlerin atanacağı grup sayısı tam bilinmezken, diskriminant ve lojistik regresyon analizinde grup sayısı bilinmekte, mevcut veriler kullanılarak bir sınıflandırma modeli elde edilmekte ve kurulan bu model yardımıyla veri kümesine eklenen yeni gözlemlerin gruplara atanması mümkün olabilmektedir (Tatlıdil, 2002).

Lojistik regresyonda yanıt değişken, doğrusal regresyon analizinde olduğu gibi, bazı değişken değerlerine dayanarak tahmin edilmeye çalışılır. Bağımlı değişken 0,1 gibi iki (binary) ya da ikiden çok kategori içeren kesikli (multinominal) değişken olduğunda normallik varsayımı bozulmakta ve doğrusal regresyon analizi uygulanamamaktadır. Böyle durumlarda lojistik regresyon analizi önerilir.

Değişkenlerin normal dağıldığı ve grupların ortak kovaryans matrisine sahip olduğu varsayımlarına dayanan diskriminant analizi yönteminde bazı değişkenlerin kesikli olması, yöntemin yanlı tahminler vermesine neden olur. Bu durumda da lojistik regresyon analizi daha etkili ve güçlü (robust) kestirimler verdiğinden diskriminant analizine alternatif olarak uygulanmaktadır (Başarır, 1990).

Lojistik regresyonda bağımsız değişkenlerin normal dağılım göstermesi, bağımlı değişken ile doğrusal ilişki içinde olması ya da bağımlı değişkenlerin (grupların) eşit varyansa sahip olması zorunluluğu yoktur (Tabachnick ve Fidell, 2015). Lojistik regresyon kategorik verilerin modellemesinde sıklıkla kullanılan bir istatistiksel yöntemdir. Bu yöntem yanıt değişkenin iki gruplu olduğu durumdan ikiden fazla gruplu olduğu duruma genelleştirilebilir. Bu analiz, yanıt değişkeninin üç veya daha fazla grup içerdiği durumlarda; yanıt değişkeni ile açıklayıcı değişkenler arasındaki ilişkiyi belirlemede kullanılan yöntemlerden birisidir.

Lojistik regresyon modelleri bağımlı değişkenin yapısına göre üçe ayrılmaktadır. Bu modellerden ikili lojistik regresyon modeli; kategorik bağımlı değişkenin iki durumlu (örneğin: var-yok) olduğu durumda kullanılır. multinominal lojistik (MNL) regresyon modeli; kategorik bağımlı değişkenin çok kategorili ve nominal (örneğin: medeni durum, evli- bekar- boşanmış) olduğu durumlarda kullanılır. Çok kategorili ve sıralı bir yapı söz konusu ise (örneğin: likert tipi ölçekler, az-orta-çok) sıralı (ordinal) lojistik regresyon modelleri kullanılır.

3.1. İki Gruplu Lojistik Regresyon Modelleri

Lojistik regresyon analizinde, yanıt değişken doğrudan modellenememektedir. Daha doğru bir yaklaşımla, lojistik regresyon analizi, $\boldsymbol{y}$ yanıt değişkeninin değerinin birleştirilmiş olasılığı üzerine kurulmuştur. Varsayalım ki model,

$$\boldsymbol{y}_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i \tag{3.1}$$

biçimine sahiptir. Burada, $\boldsymbol{\beta}' = [\beta_0, \beta_1, \beta_2 ..., \beta_p]$ ve $i = 1, 2, ..., n$'dir ve yanıt değişkeni $\boldsymbol{y}_i$'de p açıklayıcı değişken ya 0 ya da 1 olmak üzere iki grup değerini alır. Yanıt değişkeni $\boldsymbol{y}_i$'nin, aşağıdaki olasılık dağılımı ile Bernoulli rasgele değişkeni olduğu varsayılır:

$\boldsymbol{y}_i$	Olasılık
1	$P(\boldsymbol{y}_i = 1) = p_i$
0	$P(\boldsymbol{y}_i = 0) = 1 - p_i$

Burada $P(\boldsymbol{y}_i)$ i_inci deneğin yanıt değişkenin kategorilerinin birisinde yer almasına ilişkin kestirilen olasılıktır. $E(\varepsilon_i)=0$ olduğundan yanıt değişkenin beklenen değeri, $E(\boldsymbol{y}_i)=\boldsymbol{x}_i'\boldsymbol{\beta}=p_i$ olarak yazılır. Bu eşitlik, $E(\boldsymbol{y}_i)=\boldsymbol{x}_i'\boldsymbol{\beta}$ yanıt fonksiyonu ile verilen beklenen yanıtın, sadece yanıt değişkeninin 1 değerini aldığı $p_i=\mathrm{P}(\boldsymbol{y}_i=1|\boldsymbol{X}=\boldsymbol{x}_i)$ koşullu olasılık anlamına gelir. Lojistik regresyonda,

$$p_i=\frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta})} \tag{3.2}$$

diğer bir ifadeyle,

$$p_i=\frac{1}{1+\exp[-(\boldsymbol{x}_i'\boldsymbol{\beta})]} \tag{3.3}$$

ikili (binary) lojistik fonksiyonu kullanılır. (3.2)'deki fonksiyon üzerinde matematiksel işlemler yapılarak

$$\frac{p_i}{1-p_i}=\exp(\boldsymbol{x}_i'\boldsymbol{\beta}) \tag{3.4}$$

elde edilir. $\frac{p_i}{1-p_i}$ değeri, odds (olabilirlik) oranı olarak adlandırılır ve (0, ∞) aralığında herhangi bir değer alabilir. 0 ve ∞' a yakın odds değerleri yanıt değişkeninin 1 değerini almasının, sırasıyla çok düşük ve çok yüksek olasılıklarını gösterir.

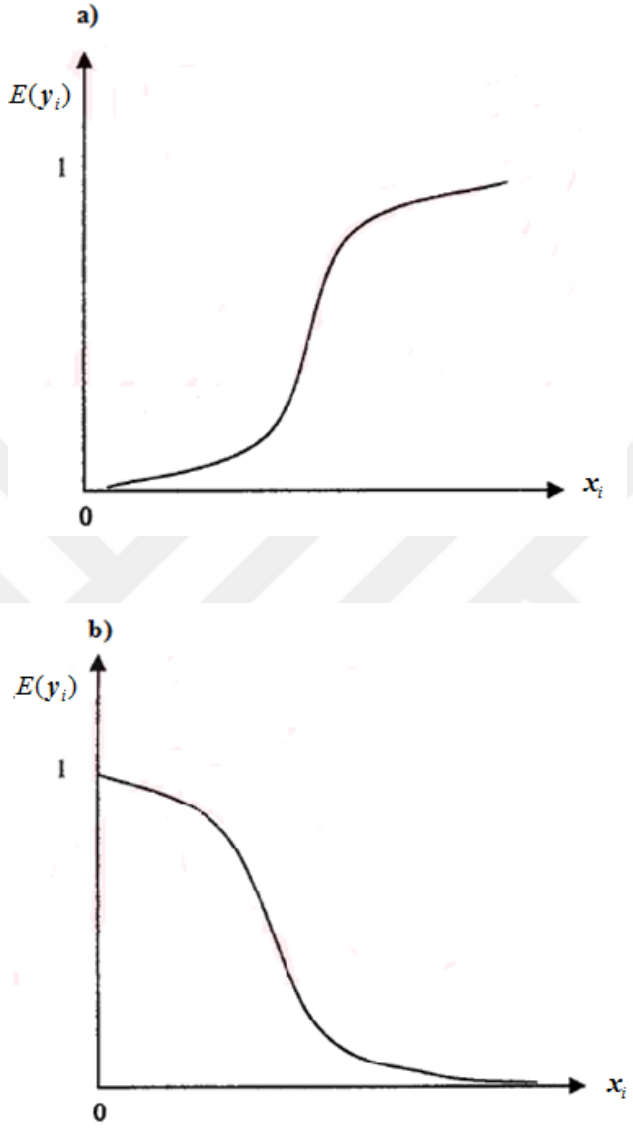
(3.4)'ün her iki tarafının logaritması alınarak,

$$\ln\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{x}_i'\boldsymbol{\beta} \tag{3.5}$$

elde edilir. Burada $\ln\left(\frac{p_i}{1-p_i}\right)$, p_i olasılığının lojit dönüşümü (log-odds oranı) olarak adlandırılır.

Lojistik regresyonda, yanıt değişkenin açıklayıcı değişkenlere karşı saçılım grafiğinden, yanıt değişkeninin kesikli olmasından dolayı ilişkiler açıkça görülemez. Bu nedenle yanıt değişken yerine p_i olasılık değerine karşı grafik çizilir. Lojistik fonksiyon çizimi adını alan bu grafik S şeklindedir ve fonksiyon sürekli olup 0-1 arasında değerler alır.

Lineer regresyon modelinde $\boldsymbol{\beta}$, $\boldsymbol{X}$'deki bir-birim artışla ilişkili $\boldsymbol{y}$'deki ortalama değişimi verir. Bunun tersine, lojistik regresyon modelinde $\boldsymbol{X}$ bir birim artarken $\boldsymbol{\beta}$ tarafından log odds oranı değişir. Ancak, (3.2)'deki p_i ve $\boldsymbol{x}_i$ arasındaki ilişki düz bir doğru olmadığından, $\boldsymbol{\beta}$ $\boldsymbol{x}_i$'deki bir birim artışla ilişkili olan p_i'deki değişime karşılık gelmez. $\boldsymbol{x}_i$'deki bir birim artış nedeniyle değişen p_i'nin miktarı $\boldsymbol{x}_i$'nin mevcut değerine bağlı olacaktır. Fakat $\boldsymbol{x}_i$ değerine bakılmaksızın, $\boldsymbol{\beta}$ pozitif ise, artan $\boldsymbol{x}_i$, artan p_i ile ilişkili olur ve $\boldsymbol{\beta}$ negatif ise, bu durumda artan $\boldsymbol{x}_i$, azalan p_i ile ilişkili olur. Bu durumda p_i ve $\boldsymbol{x}_i$ arasındaki ilişki doğrusal değildir. Lojistik regresyon fonksiyonunda, $\boldsymbol{x}_i$'nin mevcut değerlerine bağlı olarak $\boldsymbol{x}_i$'deki birim değişim başına $E(\boldsymbol{y}_i) = p_i$'deki değişim Şekil 3.1'de gösterilmektedir.



Şekil 3.1. İki gruplu lojistik regresyon fonksiyonu a) artan, b)azalan (Vupa,2004)

İki grup lojistik regresyon modeline ilişkin varsayımlar kısaca şöyledir:

- $y_i \in \{0,1\}$
- $P(y_i=1|X=x_i)=p_i$

- $y_1, y_2, ..., y_n$ değerleri istatistiksel olarak bağımsızdır,
- Açıklayıcı değişkenler (x_p) birbirinden bağımsızdır.

Lojistik regresyon modelinin yanıt değişkeninin sınırlarını genişletmek amacıyla uygulanan $\ln\left(\frac{p_i}{1-p_i}\right)$ lojit dönüşümünün bazı özellikleri de şöyle sıralanır:

- p_i arttıkça $\text{lojit}(p_i)$ 'de artar.
- p_i, 0-1 arasında iken $\text{lojit}(p_i)$ tüm gerçel değerleri alır.
- $p_i < 0.5$ ise $\text{lojit}(p_i) < 0$ ve $p_i > 0.5$ ise $\text{lojit}(p_i) > 0$'dır.

Bu özelliklerden üçüncüsü gözlemlerin sınıflara atanmasında kullanıldığı için çok önemlidir.

3.2. İkiden Fazla Gruplu Lojistik Regresyon Modelleri

Yanıt değişkeninin ikiden fazla gruplu kesikli değişken olması durumunda bu değişkenin açıklayıcı değişkenler üzerindeki regresyon modeli, çoklu grup (polychotomous) lojistik model adını almaktadır. Uygulama da bağımlı değişkenin yapısına göre iki farklı analiz uygulanır. Bağımlı değişken ikiden çok kategorili ve nominal bir değişken ise "çok gruplu (multinominal) lojistik regresyon analizi" kullanılır. bağımlı değişken sıralama ölçeğiyle elde edilmiş ise, bu durumda da "sıralı (ordinal) lojistik regresyon analizi" kullanılır. örneğin üç farklı akademik programda öğrenim görmekte olan öğrencilerden oluşan bir bağımlı değişkenin olması durumunda, MNL regresyon uygulanır. Öğrencilerin öğrenim gördükleri akademik programdaki başarılarının "düşük", "orta" ve "yüksek" olarak gruplandığı durumda ise sıralı lojistik regresyon uygulanır.

Sıralı ölçekli yanıt değişken, en az üç kategoride gözlenen değerler içermelidir. Sıralı ölçekli veriler kodlanırken ya da isimsel olarak kategorileri belirlendiğinde yanıtların doğal sıralama yapısında olması gerekir. Örneğin hastalık şiddeti söz konusu ise, hafif<orta<ağır olarak kategoriler belirlenmelidir. Hasta bireyin hastalık şiddeti bu kategori yapısı içinde doğru olarak değerlendirilmelidir.

MNL regresyon analizi yönteminde yanıt değişkeninin herhangi bir grubu, referans grup olarak alınır ve diğer gruplar bu referans gruba göre analiz edilir. *k* gruptan oluşan yanıt değişkeni için, yanıt değişkeni ile açıklayıcı değişkenler arasındaki ilişkinin incelenmesinde referans grubu ile her bir grubun tek tek incelendiği *k*–1 tane denklemin hesaplanması gerekmektedir (Kleinbaum ve Klein, 2002).

Başlangıçta yanıt değişkeninin 0, 1 ve 2 gibi üç seviyeli olduğu durumu ele alınsın. Bu durumda iki tane farklı iki grup lojistik model söz konusudur. Biri $\boldsymbol{y}_i = 1$'e karşı $\boldsymbol{y}_i = 0$ için, diğeri ise $\boldsymbol{y}_i = 2$'ye karşı $\boldsymbol{y}_i = 0$ içindir. Böylece $\boldsymbol{y}_i = 0$ grubu referans gruptur. $\boldsymbol{y}_i = 2$'ye karşı $\boldsymbol{y}_i = 1$'i karşılaştıran lojistik fonksiyon yukarıda tanımlı iki karşılaştırmaya ilişkin lojistik fonksiyonların farklarından elde edilmektedir.

Her bir gruptaki koşullu olasılıklar

$$\mathrm{P}(\boldsymbol{y}_i=0|\boldsymbol{x}_i)=\frac{1}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_1)+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_2)}$$

$$\mathrm{P}(\boldsymbol{y}_i=1|\boldsymbol{x}_i)=\frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_1)}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_1)+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_2)}$$

$$\mathrm{P}(\boldsymbol{y}_i=2|\boldsymbol{x}_i)=\frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_2)}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_1)+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_2)}$$

olarak elde edilir, burada $\boldsymbol{\beta}'_j = [\beta_{j0}, \beta_{j1}, \beta_{j2}, ..., \beta_{jp}]$, $j = 0,1,2$'dir. Bu koşullu olasılıklar kullanılarak ikili karşılaştırmaya ilişkin lojistik fonksiyonlar;

$$\ln\left(\frac{\mathrm{P}(\boldsymbol{y}_i=1|\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{y}_i=0|\boldsymbol{x}_i)}\right) = \beta_{10} + \beta_{11}\boldsymbol{x}_1 + \beta_{12}\boldsymbol{x}_2 + ... + \beta_{1p}\boldsymbol{x}_p$$

$$\ln\left(\frac{\mathrm{P}(\boldsymbol{y}_i=2|\boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{y}_i=0|\boldsymbol{x}_i)}\right) = \beta_{20} + \beta_{21}\boldsymbol{x}_1 + \beta_{22}\boldsymbol{x}_2 + ... + \beta_{2p}\boldsymbol{x}_p$$

şeklinde tanımlanır.

Yanıt değişken *k* kategorili olduğunda grupları ikişerli karşılaştıran (*k*-1) tane lojistik modele ihtiyaç duyulmaktadır. Bu durumda koşullu olasılıklar $\boldsymbol{y}_i$'nin *k* grubu için

$$\mathrm{P}(\boldsymbol{y}_i=j \mid \boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)}{\sum_{j=0}^{k-1}\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)}$$

ya da eşdeğer olarak

$$p_{ji} = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)}{\sum_{j=1}^{k}\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)} \tag{3.6}$$

ve lojistik fonksiyonlar

$$\ln\left(\frac{\mathrm{P}(\boldsymbol{y}_i=j \mid \boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{y}_i=0 \mid \boldsymbol{x}_i)}\right)=\boldsymbol{x}_i'\boldsymbol{\beta}_j \tag{3.7}$$

genelleştirilmiş formları ile elde edilir, burada $\boldsymbol{\beta}_j'=[\beta_{j0},\beta_{j1},\beta_{j2},...,\beta_{jp}]$, $j=1,2,...,k$ 'dır.

Diskriminant analizinde olduğu gibi lojistik regresyon analizinde de *k* grubu birbirinden ayırmak için (*k*-1) tane lojistik fonksiyon gereklidir.

3.3. Lojistik Model Parametrelerinin Tahmini

Lojistik modelde parametrelerin tahmin edilmesi için çeşitli yöntemler önerilmiştir. Bu yöntemler; en çok olabilirlik (EÇO), yeniden ağırlıklandırılmış iteratif EKK ile tekrarlı veri durumunda kullanılan minimum lojit ki-kare yöntemleridir (Tatlıdil, 2002). Bunlardan en yaygın kullanılanı EÇO tahmin yöntemidir.

Genel olarak EÇO yöntemi, gözlenen veri kümesini elde etmenin olasılığını maksimum yapan bilinmeyen parametrelerin değerlerini tahmin etmeyi amaçlar. Bu yöntemi uygulamak için öncelikle EÇO fonksiyonunun oluşturulması gerekmektedir. Bu fonksiyon gözlenen verilerin olasılıklarını, bilinmeyen parametrelerin bir fonksiyonu olarak verir. Bu parametrelerin EÇO tahmin edicileri, fonksiyonu maksimum yapan değerleri bulacak şekilde seçilir. Bu nedenle sonuçta elde edilen tahminler, gözlenen verilerle çok yakın değerlere sahiptir (Kaşko, 2007).

İki gruplu lojistik regresyon modeli olması durumunda $\boldsymbol{y}_i$ Bernoulli rasgele değişkeni için, başarı olasılığı $p_i=\mathrm{P}(\boldsymbol{y}_i=1|\boldsymbol{X}=\boldsymbol{x}_i)$ başarısızlık olasılığı $1-p_i$ olduğunda *i*_inci gözlem için genel olasılık $i=1,2,...,n$ için,

$$P(\boldsymbol{y}_i \mid \boldsymbol{x}_i)=p_i^{\,y_i}(1-p_i)^{1-y_i} \tag{3.8}$$

şeklinde yazılabilir. Gözlemlerin birbirinden bağımsız oldukları varsayıldığı için, n gözlem için bu olasılık terimlerinin çarpılmasıyla olabilirlik (likelihood) fonksiyon

$$l(\boldsymbol{y} \mid \boldsymbol{x}) = P(\boldsymbol{y} \mid \boldsymbol{x}) = \prod_{i=1}^{n} p_i^{\,y_i}(1-p_i)^{1-y_i} \tag{3.9}$$

şeklinde elde edilir.

EÇO yönteminin kuralı gereği, $\boldsymbol{y}$ yanıt değişkeninin gözlenme olabilirliğini maksimum kılacak biçimde, p açıklayıcı değişkene ilişkin $\boldsymbol{\beta}$ parametrelerinin tahminini bulmak amaçlanır. Bu durumda amaç, $l(\boldsymbol{y} \mid \boldsymbol{x})$ olabilirlik fonksiyonunu maksimum yapacak $\hat{\boldsymbol{\beta}}$ katsayılar vektörünü belirlemektir. Lojistik modelin olabilirlik fonksiyonu, (3.8) eşitliğinde p_i yerine (3.2) ile verilen açık ifadesi yerine konularak tekrar elde edilir. Elde edilen lojistik olabilirlik fonksiyonu, logaritması alındığında

$$L(\boldsymbol{\beta}) = \ln[l(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\beta})] = \sum_{i=1}^{n} \left[\boldsymbol{y}_i \ln(p_i) + (1-\boldsymbol{y}_i)\ln(1-p_i) \right] \tag{3.10}$$

şeklinde olup, bu fonksiyon log likelihood (LL) fonksiyonu olarak adlandırılır. LL fonksiyonunun $\boldsymbol{\beta}$'ya göre birinci türevi,

$$\sum_{i=1}^{n} (\boldsymbol{y}_i - p_i) = 0 \tag{3.11}$$

$$\sum_{i=1}^{n}(\boldsymbol{y}_i - p_i)\boldsymbol{x}_i = 0 \qquad (3.12)$$

olabilirlik denklemlerini vermektedir. Burada (3.11) ve (3.12) denklemleri sırasıyla β_0'a ve $z = 1, 2, ..., p$ için β_z'ye göre türevi alınıp sıfıra eşitlenerek elde edilen olabilirlik denklemleridir. Bu denklemlerin çözümü ile $\hat{\boldsymbol{\beta}}$ tahmin değerleri elde edilir.

Standart çoklu regresyon analizinde $\boldsymbol{\beta}$'ya göre türev alınarak elde edilen olabilirlik eşitlikleri bilinmeyen parametreleri içeren doğrusal ifadelerdir, bu nedenle kolayca çözümlenebilir. Lojistik regresyon analizinde p_i'nin (3.2) ile tanımlandığı gibi üstel olması nedeniyle denklem $\boldsymbol{\beta}$'ya göre doğrusal değildir. Bu nedenle EÇO yöntemi ile tek adımda kesin çözüme gidilemez, iteratif çözümleme gerekmektedir (Aldrich ve Nelson, 1986: Başarır 1990'dan).

İteratif çözümlemede $\boldsymbol{\beta}$'lara herhangi bir başlangıç değerleri verilerek elde edilen ilk tahminlerinden, her adımda δ kadar küçük miktarda eksiltme ya da arttırma yapılıp türevler alınır ve EÇO tahminleri bulunur. İteratif işlemler yakınsama sağlanıncaya dek devam eder. Yakınsama ancak δ düzeltme terimlerinin iterasyon değerlerini değiştirmediği noktada sağlanır ve süreç durur.

Bu yöntemin en önemli dezavantajı iterasyon sayısının çok olabilmesidir. İterasyon sayısını azaltmanın en önemli yolu ise başlangıç değerinin doğru seçimidir. Bu değerin belirlenmesinde diskriminant katsayıları ve grafiksel gösteriminden yararlanılarak gözle tahmin en yaygın kullanılan iki yoldur (Tatlıdil, 2002).

Hosmer ve Lemeshow (1989) olabilirlik denklemlerinin çözümlerinin, iteratif bir yöntem olan genelleştirilmiş ağırlıklı EKK yöntemi ile de elde edilebileceğini belirtmişlerdir.

İkiden fazla gruplu lojistik regresyon modellerinde kullanılan tahmin yöntemleri, iki grup durumunda uygulanan yöntemlerin genel halidir.

Çoklu grup lojistik regresyon için (3.6) ile elde edilen koşullu olasılıkların genel formu,

$$P(\boldsymbol{y}_i=j \mid \boldsymbol{x}_i)=P_j(\boldsymbol{x}_i) \tag{3.13}$$

ile verilebilir.

Olabilirlik fonksiyonu oluşturmak için grup üyeliğini belirlemede üç tane iki değer alan değişkenden yararlanılır. $\boldsymbol{y}=j$, $j=0,1,2$ için 3 gruba $\boldsymbol{y}_j$ olarak kodlama yapılır. Bu değişkenler, $\boldsymbol{y}=0$ için $\boldsymbol{y}_0=1$, $\boldsymbol{y}_1=0$ ve $\boldsymbol{y}_2=0$; $\boldsymbol{y}=1$ için $\boldsymbol{y}_0=0$, $\boldsymbol{y}_1=1$ve $\boldsymbol{y}_2=0$; $\boldsymbol{y}=2$ için $\boldsymbol{y}_0=0$, $\boldsymbol{y}_1=0$ ve $\boldsymbol{y}_2=1$ olmak üzere gösterge (dummy) değişkenler biçiminde kodlanmaktadır. $\boldsymbol{y}$ 'nin tüm değerleri için $\sum_j \boldsymbol{y}_j=1$ 'dir. n bağımsız gözlemli örneklem için koşullu olabilirlik fonksiyonu,

$$l(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\beta}) = \prod_{i=1}^{n} P_0(\boldsymbol{x}_i)^{y_{0i}} P_1(\boldsymbol{x}_i)^{y_{1i}} P_2(\boldsymbol{x}_i)^{y_{2i}} \tag{3.14}$$

şeklindedir.

Diskriminant analizinde olduğu gibi lojistik regresyon analizinde de *k* grubu birbirinden ayırmak için (*k*-1) tane lojistik fonksiyon gerektiğinden, toplam (k-1)(p+1) tane katsayı tahmin değeri elde edilir. Tüm sınıfların referans grupla karşılaştırılmasında $s=1,2,...,k-1$ için karşılık gelen koşullu olasılık

$$p_{si} = \mathrm{P}(\boldsymbol{y}_i = \mathrm{s} \mid \boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_s)}{\sum_{j=1}^{k} \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)} \tag{3.15}$$

ile elde edilir. n gözlem için bu olasılık terimlerinin çarpılmasıyla olabilirlik fonksiyonu

$$\begin{aligned} l(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}) &= \prod_{i=1}^{n} \mathrm{P}(\boldsymbol{y}_i = \mathrm{s} \mid \boldsymbol{x}_i) \\ &= \prod_{i=1}^{n} \left[\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_s) / \sum_{j=1}^{k} \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j) \right] \end{aligned} \tag{3.16}$$

şeklinde elde edilir. Yanıt değişkeninin *k* seviyeli olduğu çoklu grup lojistik modelde açıklayıcı değişkenlerin oluşturduğu katsayılar matrisi $\boldsymbol{X}$, $n(k-1)\times(k-1)(p+1)$ boyutludur ve $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n)$ 'dir. $\hat{p}_{si}$, (3.15) eşitliğinden p_{si} olasılık değerinin i_inci gözlem için tanımlanmış tahmini, $r' = (r_1', r_2', ..., r_n')$ gözlenen ve beklenen değerlerin farkından elde edilen rezidü (artık) vektörü ve $s = 1, 2, ..., k-1$ olmak üzere,

$$r_i = \mathrm{y}_{si} - \hat{p}_{si}$$

şeklinde tanımlanır. EÇO tahmini, (3.16) eşitliğinin logaritması alınarak elde edilen log olabilirlik (LL) fonksiyonunun $\boldsymbol{\beta}$ 'ya göre türevi alınarak

$$\frac{\partial \ln[l(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta})]}{\partial \boldsymbol{\beta}} = \boldsymbol{X}r' \tag{3.17}$$

şeklinde bulunan vektörün sıfıra eşitlenerek çözümünden elde edilir.

3.4. Lojistik Regresyon Parametrelerinin Anlamlılık Testleri

Tahmin yöntemlerinden herhangi birinin kullanımı ile bulunan $\boldsymbol{\beta}$ katsayılarının tahmininden sonra elde edilen modeldeki parametrelere ait anlamlılık testleri yapılır. Bağımlı değişken hakkında, modelde yer alan bağımsız değişkeni içeren modelin, o değişkeni içermeyen modelden daha çok bilgi içerip içermediği söz konusu bağımsız değişkene ait katsayının anlamlılık testi yapıldığında belirlenmiş olur. Buna göre bağımlı değişkene ait gözlenen değerler, her iki modelden elde edilen tahmin değerleriyle karşılaştırılır. Eğer herhangi bir bağımsız değişken içeren modelin, tahmin değerleri adı geçen değişkeni içermeyen modelden daha iyi ise o değişkenin anlamlı olduğu sonucuna varılır. Ayrıca, tahmin edilen değerlerin, gözlenen değerleri ne kadar doğru yansıttığı da dikkate alınması gereken bir husustur (Hosmer ve Lemeshow, 2000).

3.4.1. Olabilirlik oranı testi (Likelihood-ratio test)

Lojistik regresyonda gözlenen ve beklenen değerlerin karşılaştırılması log olabilirlik (likelihood) fonksiyonu ile yapılır. Olabilirliğin logaritması anlamına gelen log likelihood (LL) değeri, modelden elde edilen değişkenleri içeren olabilirlik fonksiyonu yardımıyla hesaplanır. LL değerinin -2 katı χ^2 dağılımına uyar ve lojistik regresyonda $-2LL$ istatistik değeri, ölçeklendirilmiş sapma istatistiği ya da χ^2 sapması olarak adlandırılır. $D = -2LL$ olarak ifade edilir.

Olabilirlik oran testi, hipotez testi amacıyla $D = -2LL$ sapma değerini kullanır. Doğrusal regresyondaki hata kareler toplamına karşılık gelen bu D istatistiği (Deviance istatistiği) uyum iyiliğine karar vermede önemli bir rol oynamaktadır. Bu eşitliğe göre değişkenli ve değişkensiz modele göre ayrı ayrı D

istatistiği bulunduktan sonra, G istatistiği ile bağımsız değişkenin anlamlı olup olmadığı konusunda bilgi edinilmektedir (Hosmer and Lemeshow 2000).

$$G = D(Değişkensiz\ Model) - D(Değişkenli\ Model) \tag{3.18}$$

Olabilirlik fonksiyonunun yazılması ile,

$$D = \sum_{i=1}^{n} d_i^2$$

$$= -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\hat{p}_i}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\hat{p}_i}{1-y_i}\right)\right] \tag{3.19}$$

biçimine dönüşen sapma ölçütü, p modeldeki parametre sayısını göstermek üzere, $(n-p)$ serbestlik dereceli χ^2 tablo değeri ile karşılaştırılır.

G'yi hesaplamak için farkı alınacak D değerlerinin her ikisi için de doymuş modelin olabilirlikleri ortak olduğundan G istatistiği

$$G = -2\ln\left[\frac{\text{Değişkensiz Modelin Olabilirliği}}{\text{Değişkenli Modelin Olabilirliği}}\right] \tag{3.20}$$

şeklini almaktadır. Sapma değeri minimum olan model en iyi model olarak düşünülür.

3.4.2. Wald Testi

Lojistik regresyon katsayılarının anlamlı olup olmadığını test etmede kullanılabilecek ikinci test Wald testidir. Wald testine ait test istatistiğinin dağılımı

standart normal dağılıma yaklaşır. Her değişken için standart hatalar kullanılarak Z testi yapılır. Lojistik regresyon katsayılarının, standart hatasına oranı;

$$W = Z = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} \tag{3.21}$$

Wald istatistiği olarak adlandırılır. Bu istatistik standart normal dağılımın kritik değerleri ile karşılaştırılarak anlamlılığı belirlenir. Aynı zamanda Wald istatistiği

$$W = Z^2 = \left(\frac{\boldsymbol{\beta}}{SE(\boldsymbol{\beta})} \right)^2 \tag{3.22}$$

şeklinde de elde edilebilmektedir. Bu durumda ise Wald test istatistiği 1 serbestlik dereceli χ^2 dağılımı gösterir ve 1 serbestlik dereceli χ^2 dağılımının kritik değerleri ile karşılaştırılarak anlamlılığı belirlenir.

Menard (1995), büyük katsayılarda standart hatanın büyümesi Wald istatistiği (χ^2) değerinin küçüldüğü konusunda uyarmıştır. Agresti (1986), küçük örnek genişliği için olabilirlik oran testinin Wald istatistiğinden daha uygun olduğunu bildirmiştir. Wald testi, örnek hacminin büyük olması durumunda anlam kazanmaktadır.

3.4.3. Skor Testi

Bağımlı değişken ile bağımsız değişken arasında ilişki olup olmadığı konusunda fikir veren bu istatistiğin temeli olabilirlik fonksiyonuna dayanmamaktadır. Bu istatistik, asimptotik olarak *p* serbestlik dereceli χ^2 dağılımı göstermektedir (Sharma, 1996).

Tek parametreli iki gruplu basit lojistik regresyon modeli,

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \tag{3.23}$$

olmak üzere, skor testi (ST) için test istatistiği,

$$ST = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sqrt{\bar{y}(1-\bar{y}) \sum_{i=1}^{n} (x_i - \bar{x})^2}} \tag{3.24}$$

ile elde edilir. Burada $\bar{y} = n_1 / n$ ve $\bar{y}(1-\bar{y}) \sum_{i=1}^{n} (x_i - \bar{x})^2$ varyans tahminidir.

Skor istatistiği yukarıdaki eşitlik ile lojistik regresyondaki herhangi bir değişkene ait katsayının anlamlılık testi için geçerli bir yöntemdir (Hosmer and Lemeshow, 2000). Katsayıların anlamlılık kontrolü yapıldıktan sonra katsayıların yorumlanması odds oranları kullanılarak yapılmaktadır (Christensen, 1990).

3.5. Lojistik Regresyon Katsayılarının Yorumlanması

Lojistik yanıt fonksiyonunda tahmin edilen $\boldsymbol{\beta}$ regresyon katsayısının yorumlanması doğrusal bir regresyon modelindeki kadar kolay değildir. Lojistik model karmaşık ve lineer olmayan bir forma sahip olduğu için $\boldsymbol{X}$ değişkenindeki bir birimlik artışın etkisini ölçmek zordur. Lineer lojit modeldeki $\boldsymbol{\beta}$ 'nın işareti lojistik fonksiyon eğrisinin yönünü belirler. Lojistik regresyon eğrisi için lineer yaklaşım bu parametrelerin tahminin kolaylaştırmasına rağmen, yorumlanmaları daha fazla dikkat gerektirir.

Açıklayıcı değişkenlerdeki birim başına değişim, logit ya da odds'ların doğal logaritmasında artan ya da azalan etkiye sahiptir. Ancak, bu etkiler bilinmeyen $\boldsymbol{\beta}$ parametrelerinin büyüklüğü ile doğrudan ölçülememektedir.

β_i katsayısı yorumlanırken; $\boldsymbol{x}$ 'deki her birimlik artış için $\frac{p_i}{1-p_i}$ odds oranı tahmini ile elde edilen lojistik yanıt fonksiyonundan yararlanılır.

$$\text{odds} = \frac{p_i}{1-p_i} = \exp(\beta_0 + \beta_i \boldsymbol{x}) = e^{\beta_0}(e^{\beta_i})^x \tag{3.25}$$

(3.25), β_i katsayısı için kolay bir yorumlama sağlar. Bu durumda $\boldsymbol{x}$ değişkenindeki her birimlik artış, olayın gerçekleşme olasılığının (odds'un) logaritmasını β_i kat arttırır. Diğer bir ifadeyle, $\boldsymbol{x}$ değişkenindeki her birimlik artış olasılığın logaritmasını β_i kat arttırıyor ise olasılığı e^{β_i} kat arttırır şeklinde yorumlama yapılır.

Bağımsız değişkenlerin ölçek tiplerine göre katsayıların yorumlanması değişmektedir. Lojistik regresyon modelindeki tahmin edilen katsayının alternatif olarak uygun yorumu, iki lojit arasındaki farka dayandırılır.

Belli bir özelliğe sahip olma durumuna göre $\boldsymbol{x}$ değişkeninin 1 ve 0 olarak ikili değer aldığı ölçek tipi için, $p_i = p_i(x)$ 'nin ve $1-p_i$ 'nin iki farklı değeri olur. $\boldsymbol{x}=1$ değerli gözlemler arasındaki odds değeri $p(1)/[1-p(1)]$, benzer olarak $\boldsymbol{x}=0$ değerli gözlemler arasındaki odds değeri $p(0)/[1-p(0)]$ olarak tanımlanır. Odds oranı (OR),

$$\text{OR} = \frac{\text{odds}(\boldsymbol{x}=1)}{\text{odds}(\boldsymbol{x}=0)} = \frac{p(1)/[1-p(1)]}{p(0)/[1-p(0)]}$$

$$= \frac{\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right)\left(\frac{1}{1+e^{\beta_0}}\right)}{\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right)\left(\frac{1}{1+e^{\beta_0+\beta_1}}\right)} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

olarak elde edilir.

İkili bağımsız değişkenli lojistik regresyon için lojit fark aracılığıyla,

$$\ln[\mathrm{OR}(1,0)] = \ln\left[\mathrm{odds}(\boldsymbol{x}=1)\right] - \ln\left[\mathrm{odds}(\boldsymbol{x}=0)\right]$$
$$= \ln(e^{\beta_1}) = \beta_1$$

elde edilir.

OR'nin dağılımı yeterince büyük veriler için normaldir. Bu durumda OR'ya göre yorum yapılabilir: OR>1 ise özelliğe sahip olmak, özelliğe sahip olmamaya göre yanıt değişken üzerindeki olasılığı arttırır. OR<1 ise özelliğe sahip olmamak, özelliğe sahip olmaya göre olasılığı arttırır. Her zaman büyük örneklem ihtiyacı sağlanmayabilir. Böyle durumlarda dahi odds oranının logaritmik dönüşümü $\ln(\mathrm{OR})$'ın dağılımı normaldir.

4. PROBİT REGRESYON ANALİZİ

Kategorik verilerin analizlerinde kullanılan doğrusal olmayan modelleme yöntemleri üzerine yapılan çalışmalardan biri de probit regresyon analizidir. Probit modelin literatüre ilk geçişi Berkson (1944) tarafından önerilen lojistik modelden 10 yıl önce 1934' te; İngilizcesi probability unit'in (olasılık birimi) kısaltması olan probit terimini ortaya atan Bliss tarafından gerçekleşmiştir.

Bliss'in yaptığı çalışma probit modelin ilk olarak uygulandığı çalışmadır. Bliss bu çalışmada; doz-yanıt (dosage-response) ilişkisini araştırmıştır. Biyolojiksel mikroorganizma kitlesine ait bu araştırmada; kademeli olarak arttırılan doz düzeylerine maruz kalan organizmaların verdikleri iki sınıflı ölüm ya da sağ kalım yanıtlarının doz düzeyleri ile olan ilişkisini incelemek için probit model kullanılmıştır.

İlerleyen yıllarda, iki gruplu yanıt değişkenlerin analizlerinin ardından, nominal ya da sıralı yapıdaki ikiden fazla gruplu yanıt değişken analizlerini incelemek için sırasıyla çok guruplu ve sıralı probit regresyon modelleri üzerinde çalışılmıştır.

4.1. İki Gruplu Probit Model

Probit modelin altında yatan varsayım, yanıt fonksiyonunun

$$\boldsymbol{y}_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i$$

formunda olmasıdır. Burada $\boldsymbol{x}_i$ gözlemlenebilen fakat $\boldsymbol{y}_i$ gözlemlenemeyen değişkendir. Probit modelde, temel yanıt değişkenin iki düzeyli hale getirilmemiş iken normal dağıldığı sayılır. Yanıt değişken

$$\boldsymbol{y}_i = \begin{cases} 1, \boldsymbol{y}_i > c \\ 0, \boldsymbol{y}_i < c \end{cases}$$

şeklinde iki düzeyli hale getirilir. $\boldsymbol{y}_i$ yanıt değişkeninin sonucu atanırken, eşik değer olarak kullanılan c değeri genellikle 0 olarak alınmakta olup, sıfır yerine başka sayı değeri de kullanılabilir.

x değişkeni, μ ortalamalı σ^2 varyanslı normal dağılıma sahip ise x değişkenine ait olasılık yoğunluk ve kümülatif dağılım fonksiyonları şu şekilde olmaktadır:

$$f(\mathrm{x}) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(\mathrm{x}-\mu)^2/2\sigma^2}, -\infty < \mathrm{x} < \infty \tag{4.1}$$

$$F(\mathrm{x}) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma}} e^{-(\mathrm{x}-\mu)^2/2\sigma^2} \tag{4.2}$$

Normal standart z değişkeni için, $\Phi(z)$ kümülatif normal dağılım fonksiyonu $\Phi(z) = P(Z \le z)$ olarak tanımlanırsa, hata terimlerinin bağımsız ve standart normal dağılıma sahip olduğu varsayımı ile probit model

$$\begin{aligned} p_i = P(\boldsymbol{y}_i = 1 \mid \boldsymbol{X}) &= P(\boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i > 0 \mid \boldsymbol{X}) \\ &= P(\varepsilon_i > -\boldsymbol{x}_i'\boldsymbol{\beta} \mid \boldsymbol{X}) \\ &= 1 - F(-\boldsymbol{x}_i'\boldsymbol{\beta}) \\ &= 1 - \Phi\left(\frac{-\boldsymbol{x}_i'\boldsymbol{\beta}}{\Sigma}\right), \Sigma = \mathrm{I} \\ &= \Phi(\boldsymbol{x}_i'\boldsymbol{\beta}) \end{aligned} \tag{4.3}$$

şeklinde tanımlanır. Ayrıca

$$\begin{aligned} P(\boldsymbol{y}_i = 0 \mid \boldsymbol{X}) &= P(\boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i \le 0 \mid \boldsymbol{X}) \\ &= P(\varepsilon_i \le -\boldsymbol{x}_i'\boldsymbol{\beta} \mid \boldsymbol{X}) \\ &= F(-\boldsymbol{x}_i'\boldsymbol{\beta}) \\ &= \Phi(-\boldsymbol{x}_i'\boldsymbol{\beta}) \end{aligned} \tag{4.4}$$

olarak ifade edilebilir. Burada Φ standart normal olasılık dağılımıdır. $\boldsymbol{x}_i'\boldsymbol{\beta}$ probit skoru ya da indeksi olarak adlandırılır ve normal dağılıma sahiptir.

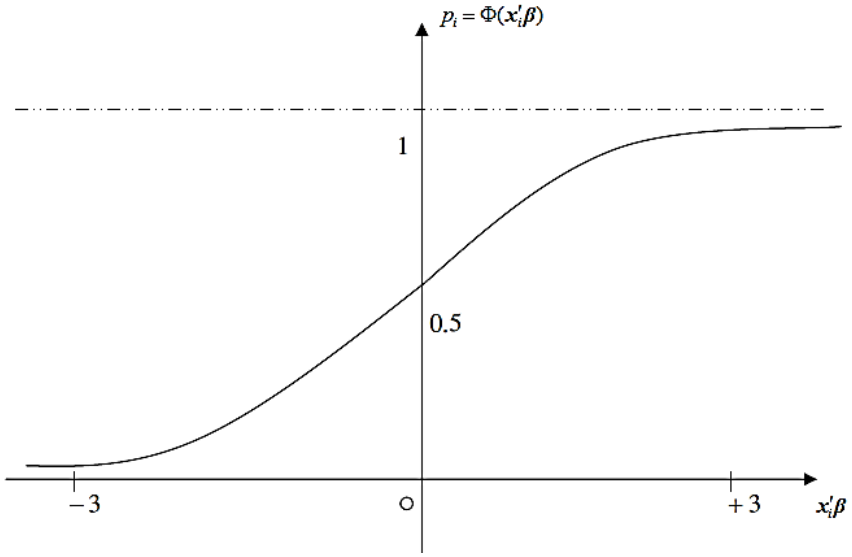
Bir olayın gerçekleşme olasılığını ifade eden p_i, 0 ve 1 aralığı içinde yer alır. Bu olasılık standart normal eğrinin $-\infty$ ile $\boldsymbol{x}_i'\boldsymbol{\beta}$ arasındaki bölgenin alanına eşit olup $\boldsymbol{x}_i'\boldsymbol{\beta}$ indeksinin büyük değerleri olayın gerçekleşme olasılığının yüksek olduğunu ifade etmektedir.

Kümülatif dağılım fonksiyonu monoton bir fonksiyondur ve fonksiyon monoton olduğu sürece tersi vardır. Probit model için bu fonksiyon monoton artandır ve probit regresyon denklemini elde etmek için (4.3) eşitliğinin tersi alınabilir:

$$\Phi^{-1}(p_i) = \boldsymbol{x}_i'\boldsymbol{\beta} \tag{4.5}$$

burada Φ^{-1} kümülatif normal dağılım fonksiyonunun tersidir ve (4.3) eşitliğindeki kümülatif normal dağılım fonksiyonunun tersi alınarak probit fonksiyonunun doğrusallaştırılması sağlanır. (4.5) modeli de genelleştirilmiş lineer modeller formunda probit regresyon modeli olarak ifade edilir (Özarıcı, 1996).

Probit katsayısı $\boldsymbol{\beta}$, tahmindeki bir birimlik artışın probit indeksinde yapacağı $\boldsymbol{\beta}$ standart sapmalık yükselmeyi ifade eder. Diğer bir ifadeyle, probit katsayısı bağımsız değişkenin bağımlı değişkene ait standart z değerinde yapacağı etkiyi ölçer. Bu katsayıların sayısal büyüklüklerinin bir önemi ve özel bir yorumu yoktur, sadece ilişkinin yönü ve derecesini belirler (Arı ve Önder, 2012).



Şekil 4.1. Probit Model

4.2. Probit Model Parametrelerinin Tahmini

İki düzeyli bağımlı değişken modellerinden probit regresyon modelindeki parametreleri tahmin etmede lojistik model için uygulanan benzer yöntemler kullanılır. Genellikle ağırlıklı EKK, EÇO, minimum ki-kare ve iteratif olarak yeniden ağırlıklandırılmış EKK teknikleri kullanılmaktadır (Finney, 1971). Lojit modelde olduğu gibi bireysel gözlemler kullanıldığında probit modelde de en uygun tahmin yöntemi EÇO yöntemi olmaktadır (İçyüz, 2010).

Regresyon modellerinin tahmininde sıradan EKK dan sonra en çok kullanılan ve sıradan EKK tekniğinden daha güçlü teorik özelliklere sahip nokta tahmin yöntemi EÇO yöntemidir.

Kümülatif normal dönüştürme yaklaşımı doğrusal olmadığından probit modelin tahmini için ağırlıklı EKK yöntemi uygun değildir. Bu nedenle, EÇO yöntemi ile probit modelin parametrelerinin daha tutarlı tahminleri elde edilebilir.

$\boldsymbol{y}_i$ Bernoulli rasgele değişkeni için, bir olayın meydana gelme olasılığı $p_i = \mathrm{P}(\boldsymbol{y}_i = 1 | \boldsymbol{X} = \boldsymbol{x}_i)$, $i = 1, 2, ..., n$ ve meydana gelmeme olasılığı $1 - p_i$ olduğunda i_inci gözlem için genel olasılık,

$$\mathrm{P}(\boldsymbol{y}_i \mid \boldsymbol{x}_i) = p_i^{\,y_i} (1 - p_i)^{1 - y_i} \tag{4.6}$$

şeklinde yazılabilir. Gözlemlerin birbirinden bağımsız oldukları varsayıldığı için, n gözlem için bu olasılık terimlerinin çarpılmasıyla olabilirlik (likelihood) fonksiyonu

$$l(\boldsymbol{y} \mid \boldsymbol{x}) = P(\boldsymbol{y} \mid \boldsymbol{x}) = \prod_{i=1}^{n} p_i^{\,y_i} (1 - p_i)^{1 - y_i}$$

veya

$$l(\boldsymbol{y} \mid \boldsymbol{x}) = p_1 ... p_{n_1} (1 - p_{n_1 + 1}) ... (1 - p_n)$$

şeklindedir, burada $n_1 :\ \boldsymbol{y}_i = 1$ olan gözlem sayısı, $n_2 :\ \boldsymbol{y}_i = 0$ olan gözlem sayısı olmak üzere $n_1 + n_2 = n$ 'dir.

EÇO yönteminde, probit olabilirlik değerini mümkün olduğunca maksimize edecek model parametre değerleri bulunmaya çalışılmaktadır. Bu nedenle elde

edilen probit olabilirlik fonksiyonu, p_i yerine (4.3) ile verilen açık ifadesi konularak logaritması alındığında

$$L(\boldsymbol{\beta}) = \ln[l(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\beta})] = \sum_{i=1}^{n} \left[\boldsymbol{y}_i \ln(\Phi(\boldsymbol{x}_i'\boldsymbol{\beta})) + (1-\boldsymbol{y}_i) \ln(1-\Phi(\boldsymbol{x}_i'\boldsymbol{\beta})) \right] \quad (4.7)$$

şeklinde olup, bu fonksiyon log olabilirlik (log likelihood, LL) fonksiyonu olarak adlandırılır. LL fonksiyonunun $\boldsymbol{\beta}$ 'ya göre birinci türevi alınıp sıfıra eşitlenir.

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left[\frac{\boldsymbol{y}_i \phi_i}{\Phi_i} + (1-\boldsymbol{y}_i) \frac{-\phi_i}{(1-\Phi_i)} \right] \boldsymbol{x}_i = 0 \quad (4.8)$$

Burada ϕ_i: standart normal dağılım için olasılık yoğunluk fonksiyonu, Φ_i: kümülatif dağılım fonksiyonudur.

Probit için olabilirlik denklemleri

$$\sum_{i=1}^{n} \left[\frac{\left[\boldsymbol{y}_i - \Phi(\boldsymbol{x}'\hat{\boldsymbol{\beta}}) \right] \phi(\boldsymbol{x}'\hat{\boldsymbol{\beta}})}{\Phi(\boldsymbol{x}'\hat{\boldsymbol{\beta}}) \left[1-\Phi(\boldsymbol{x}'\hat{\boldsymbol{\beta}}) \right]} \right] \boldsymbol{x}_{ij} = 0 \text{ , } j = 1, 2, ..., p \text{ parametre sayısı} \quad (4.9)$$

olarak elde edilir.

Elde edilen *p*-tane denklemin çözülmesiyle model parametrelerinin tahmin değerleri elde edilir. Ancak olabilirlik fonksiyonundaki β_j parametreleri doğrusal değildir. Lojistik modelde açıklandığı gibi probit modelde de EÇO yöntemi ile doğrusal olmayan parametreler tahmin edilebileceğinden EÇO yöntemi ile tek adımda kesin çözüme gidilemez, iteratif çözümleme gerekmektedir.

İteratif çözüm yardımıyla, bir başlangıç tahmininden başlanarak, log olabilirlik değerini değiştiren parametrelerdeki değişimin büyüklüğü ve yönü belirlenir. Daha fazla değişim ile olabilirlik fonksiyonunda yeterli bir iyileşme olmadığı zaman, iterasyon durur ve algoritmanın dönüştüğü söylenir. Bir dizi farklı dönüşüm kriterleri kullanılabilir. Log-olabilirlikteki (yüzde) değişim ya da parametrelerdeki (yüzde) değişim gibi farklı algoritmalar, bir iterasyondan diğerine, katsayılardaki değişimi ölçen farklı teknikler kullanır (Aldrich ve Nelson, 1984).

EÇO tahmin edicilerini elde etmek için iteratif yöntemlerden genellikle Newton-Raphson ve skorlama yöntemi kullanılmaktadır (İçyüz,2010). Newton-Raphson ile benzeri algoritmaları kullanan ve probit analizinin çözümlenmesinde yararlanılan istatistik programlarından bir tanesi olan SPSS programı, probit regresyon modellerin parametre tahminlerini hesaplarken Newton-Raphson algoritmasından yararlanır. Benzer istatistik programları; Quasi-Newton, Simplex, Simplex ve Quasi-Newton, Hooke-Jeeves Pattern Mover, Rosenbrock Pattern Search, Rosenbrock ve Quasi-Newton gibi farklı yöntemler kullanarak da işlem yapabilir (Özarıcı, 1996).

4.3. İkiden Fazla Gruplu (Multinominal) Probit Model

İki düzeyli bağımlı değişkenlerin analizlerinin ardından, nominal yapıdaki bağımlı değişken analizlerini incelemek için ikiden fazla düzeyli (*multinominal*) yapıdaki modellemeler üzerinde çalışılmıştır. Yanıt düzeylerinin çok gruplu dağılım göstermesi durumunda, kategoriler arasındaki bağımlılık yapısını göz önüne alan ve kategoriler arasında farklılık gösteren değişkenleri de modele katma olanağı sağlayan Multinominal Probit (MNP) modeli bu koşullarda diğer tüm modellerden daha kapsamlı olmaktadır.

MNP model ilk olarak 1927'de Thurstone tarafından önerilmiştir. Ardından uzun bir süre matematiksel olarak hesaplama yöntemleri karışık olan bu modelin

hesaplama yöntemleri üzerinde çalışılmıştır. Çünkü MNP modelde kullanılan bağ fonksiyonu çok katlı integral hesaplamalarını gerektirmektedir. Bu sıkıntıdan dolayı bir dönem bu yöntem, uygulamalarda yaygın olarak kullanılamamıştır. Araştırmacılar bu dönem boyunca, yöntemi kolay uygulanabilecek hale getirmek için çalışmışlardır. MNP modelinin ilk olarak formülleştirilmesi ise literatürde Hausman ve Wise tarafından 1978'de olmuştur. İlerleyen yıllarda MNP modelin geliştirilmesi devam etmiş ve bilgisayar uygulamalarında ortaya çıkan hesaplama zorluğu aşılmıştır. Böylelikle MNP model bilgisayar uygulamalarına daha kolay aktarılabilmiş ve geliştirilebilmiştir. Bu kolaylıkla birlikte modelin yaygın biçimde kullanımını sağlamıştır.

MNP modelin temelinde MNL modelde olduğu gibi; bireyin o kategori içerisinde yer alması olasılığının modellenmesi yatar. MNP modelde bağ fonksiyonu olarak; kümülatif standart normal dağılım fonksiyonunun tersini kullanır. Dolayısıyla incelenen fonksiyon yapısı kümülatif standart normal dağılır. Gerekli varsayım ise hataların normal dağıldığı varsayımıdır. Hataların varyans kovaryans matrisinin hesaplanabildiği model, seçimler arasındaki ilişkiyi inceleyebilme olanağı sağlamaktadır.

MNL modeller hataların her birinin bağımsız dağılabileceğini varsayar. Bu bağımsızlık varsayımı olabilirlik fonksiyonunu daha kolay hesaplama avantajına sahip olmasına rağmen, çoğu durumda gerçek olmayan tahminlere neden olur. MNP modeller yoğun bir hesaplama gerektirir, ancak bağımsızlık varsayımı gerektirmez ve pek çok araştırmacı bu sebeple MNP modelin daha iyi bir model olduğunu varsayar (Kropko, 2008).

MNP modelin hataların bağımsızlık varsayımını gerektirmemesi; bu varsayımın hem gerçellendiği hem de gerçellenmediği koşulda kullanılabilmesini sağlar. Bağımsızlık durumu varyans-kovaryans matrisinin birim matrisi ($\Sigma = I$) olduğu durumda ortaya çıkar ve Hausman test istatistiği ile test edilebilir. Varyans-kovaryans matrisinin birim matrise eşit olmaması durumunda ($\Sigma \neq I$), yani

bağımsızlık varsayımının gerçekleşmediği durumda MNP model, MNL modellere göre parametre kestirimlerinde üstün olmaktadır (Altınışık,2007).

Yanıt değişken *k* kategorili olduğunda varsayılan model

$$\boldsymbol{y}_{ij} = \boldsymbol{x}_i'\boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_j$$

şeklindedir, burada $\boldsymbol{\varepsilon}_j \sim N(0,\Sigma)$, $\boldsymbol{x}_i' = [1, \mathrm{x}_{i1}, \mathrm{x}_{i2}, ..., \mathrm{x}_{ip}]$, $\boldsymbol{\beta}_j' = [\beta_{j0}, \beta_{j1}, \beta_{j2}, ..., \beta_{jp}]$, $i = 1,2,...,n$ ve $j = 1,2,...,k$'dır. Bu durumda *i*_inci bireyin *j*_inci kategoriye seçilme olasılığı

$$\begin{aligned} p_{ij} &= P(\boldsymbol{y}_i = j \mid \boldsymbol{x}_i) \\ &= P(\boldsymbol{y}_{ij} > \boldsymbol{y}_{i1}, ..., \boldsymbol{y}_{ij} > \boldsymbol{y}_{i(j-1)}, \boldsymbol{y}_{ij} > \boldsymbol{y}_{i(j+1)}, ..., \boldsymbol{y}_{ij} > \boldsymbol{y}_{ik}) \\ &= P[(\varepsilon_{ij} - \varepsilon_{i1}) > \boldsymbol{x}_i'(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_j), ..., (\varepsilon_{ij} - \varepsilon_{i(j-1)}) > \boldsymbol{x}_i'(\boldsymbol{\beta}_{(j-1)} - \boldsymbol{\beta}_j), \\ &\qquad (\varepsilon_{ij} - \varepsilon_{i(j+1)}) > \boldsymbol{x}_i'(\boldsymbol{\beta}_{(j+1)} - \boldsymbol{\beta}_j), ..., (\varepsilon_{ij} - \varepsilon_{ik}) > \boldsymbol{x}_i'(\boldsymbol{\beta}_k - \boldsymbol{\beta}_j)] \end{aligned} \quad (4.10)$$

olarak elde edilir. Bu olasılığa bakarak sadece $\boldsymbol{y}_{ij}$'ler arasındaki farklılıklar belirlenir bu nedenle lojit model durumundaki gibi bir referans kategorisi atanır. Bunun bir sonucu olarak da kovaryans matrisi $k \times k$ dan $(k-1)\times(k-1)$ boyutuna indirgenir (Hujer, 2004). MNL modelde olduğu gibi MNP modelde de kategori sayısının bir eksiği $(k-1)$ kadar model tahmin edilir.

$l = 1, ..., (j-1), (j+1), ..., k$ olmak üzere $\tilde{\varepsilon}_{il} = \varepsilon_{ij} - \varepsilon_{il}$ ve $u_{il} = \boldsymbol{x}_i'(\boldsymbol{\beta}_l - \boldsymbol{\beta}_j)$ tanımlanırsa, bu durumda $P(\boldsymbol{y}_i = j \mid \boldsymbol{x}_i)$ olasılığı

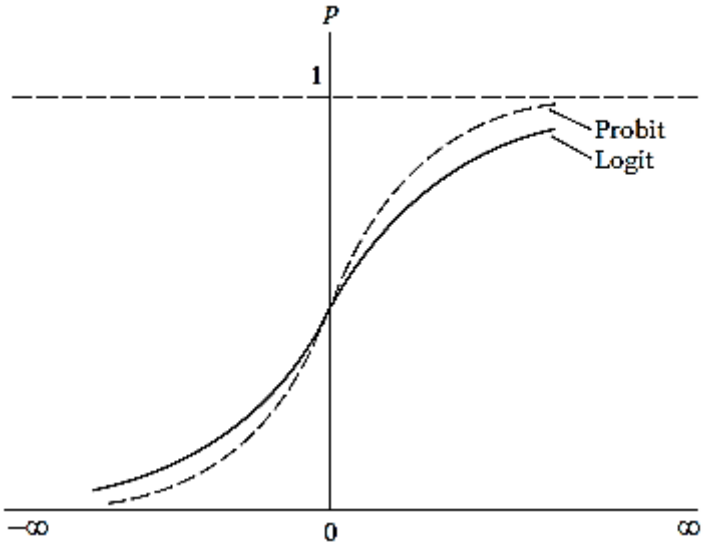
$$\int_{u_{i1}}^{\infty} \dots \int_{u_{i(j-1)}}^{\infty} \int_{u_{i(j+1)}}^{\infty} \dots \int_{u_{ik}}^{\infty} \phi(\tilde{\varepsilon}_{i1}, \dots, \tilde{\varepsilon}_{i(j-1)}, \tilde{\varepsilon}_{i(j+1)}, \dots, \tilde{\varepsilon}_{ik}) d\tilde{\varepsilon}_{i1} \dots d\tilde{\varepsilon}_{i(j-1)} d\tilde{\varepsilon}_{i(j+1)} \dots d\tilde{\varepsilon}_{ik} \tag{4.11}$$

ile verilir. MNP modeller ile burada pratik olarak çözüm elde etmek zordur. Böyle yüksek boyutlu integraller kapalı formda değildirler, yani integraller sınırlandırılmamıştır. Bu nedenle $k \geq 3$ için bu integraller Monte-Carlo simülasyon teknikleri kullanılarak hesaplanabilmektedir.

4.4. Probit Regresyon ile Lojistik Regresyon Analizinin Karşılaştırması

Probit regresyon analizi lojistik regresyon ile oldukça ilişkilidir ve probit model, lojistik regresyon modeline alternatif olarak kullanılan bir modeldir. Her iki regresyon modeli de yanıt değişkeninin iki ya da daha fazla gruptaki gözlemlerin olasılığı üzerine yoğunlaşır. İki analiz de belirli bir dizi değişkenden yanıt değişkenin 1'e eşit olmasının olasılık tahminlerini üretir.

Lojistik regresyon ile probit regresyon analizi arasındaki farklılık, yanıt değişkeni oluşturan olasılıklara uygulanan dönüşümden kaynaklanır. Lojistik regresyon olasılığın lojit dönüşümünü kullanırken; probit regresyon her bir gözlenen olasılığın, gözlenen olasılığın bulunduğu normal dağılımın altındaki değerlerle değiştirildiği probit dönüşümü kullanır (Tabachnick ve Fidell, 2015). Dolayısıyla lojistik modelde lojistik dağılım fonksiyonu kullanılırken, probit modelde kümülatif standart normal dağılım fonksiyonu kullanılmaktadır. Ayrıca probit ve lojit modelin aralarındaki diğer bir fark, lojistik birikimli dağılım fonksiyonunun uç bölgelerinin probit modele göre daha geniş olmasıdır; Şekil 4.2'deki birikimli dağılım fonksiyonları incelendiğinde bu durum görülebilmektedir.



Şekil 4.2. Probit ve lojit modelin kümülatif dağılımları (Gujarati, 2003)

Lojistik dağılım ve kümülatif normal dağılımın birbirlerine benzeyen dağılımlar olmasından dolayı örneklem çok büyük olmadıkça, iki modele ait eşitliklerden elde edilen sonuçlar çok farklı olmayacaktır (Sakal, 2014).

Olasılığın 0.5 olduğu durumlarda her iki dönüşümde sıfır değerini üretirler. Çünkü probit için normal dağılımda gözlemlerin yarısı z=0 değerinin altında kalacaktır. Lojit dönüşüm için ise durum

$$\ln\left(\frac{0.5}{1-0.5}\right)=0$$

şeklindedir. Ancak, uç noktalarda değerler farklılaşmaktadır. Örneğin; 0.95 olasılığında probit(z) değeri $\Phi^{-1}(0.95)=1.65$ iken lojit(0.95)=2.94'tür. Fakat lojit ve probit dağılımlarının şekilleri, olasılıklar çok aşırı yüksek olmadıkça oldukça benzerdir, aynı zamanda iki analizin sonuçları da oldukça benzerdir.

Probit modeldeki yanıt değişken için normal dağılım varsayımı, probit regresyon analizini lojistik regresyon analize göre daha sınırlandırıcı yapmaktadır. Bu nedenle, çok düşük veya çok yüksek değerlere sahip birçok gözlem değerinin bulunduğu veri setlerinde normal dağılımla ilgili sıkıntı olacağından, lojistik regresyon probit regresyon analizine göre daha iyi bir seçenek olarak düşünülür (Tabachnick ve Fidell, 2015).

5. UYGULAMA VE BULGULAR

Bu çalışmada ilk olarak omurga hastalıklarının teşhis ve sınıflandırılmasını sağlamak amacıyla, 3 sınıflı ve her bir sınıfta farklı büyüklükte örneklem bulunan vertebral kolon (omurga) veri seti kullanılacaktır. Bu veri üzerine Diskriminant Analizi, Lojistik Regresyon Analizi ve Probit Regresyon Analizi uygulanarak optimum sınıflandırma modeli geliştirilip bu yöntemlerin sınıflandırma doğrulukları karşılaştırılacaktır. Bu uygulamada kullanılan vertebral kolon veri seti UCI Machine Learning Repository (Frank ve Asuncion, 2010) internet sitesinden elde edilmiştir.

İkinci uygulama olarak bir simülasyon çalışması ile 3 sınıflı ve her bir sınıfta eşit büyüklükte örneklem bulunan granite veri seti üretilmiştir. Üretilen bu simülasyon verisi üzerine araştırılan bu 3 analiz için benzer adımlar uygulanarak, analiz yöntemlerinin sınıflandırma doğrulukları karşılaştırılacaktır.

Çalışmadaki hesaplamalar SPSS 21 ve STATA 13 programlarından elde edilmiştir. Üretilen simülasyon verisi ise MATLAB programı aracılığıyla elde edilmiştir.

5.1. Gerçek Veri Seti Üzerine Uygulama

5.1.1. Veri Setine İlişkin Bilgiler

Vertebral kolon (omurga) insan vücudunun ayrılmaz bir parçasıdır. Genellikle otuz üç omurgadan oluşan, yirmi dört adet eklemli, dokuzu kaynaşmış omurgadan oluşan bir yapıdır. Gövdenin dorsal (arka) kısmında bulunur ve omurlar arası (intervertebral) disklerin sayısı ile ayrılmıştır. Bu otuz üç vertebra (omurga), servikal (boyuna ait) eğride yedi omur, torasik (göğse ait) eğride on iki omur, lomber (bele ait) eğride beş omur ve sakrum (kuyruk sokumu) eğrisinde dokuz omur içeren beş farklı gruba ayrılmıştır (Bailey ve Love, 2008).

Bu çalışmadaki vertebral kolon veri setinde iki farklı omurga bozukluğu olan disk hernisi (bel fıtığı) ve spondilolistezi (bel kayması) yer almaktadır.

- Disk Hernisi (Bel fıtığı); omurganın en hareketli bölümlerine etki eden fleksiyon (esneme) kuvvetinin etkisi ile oluşan omurlar arası disk çıkıntısıdır.
- Spondilolistezi (Bel kayması); omurga kemiklerinin birbirleri üzerinde öne ya da arkaya doğru kaymış olmasıdır.

Bu çalışmada kullanılan veri seti, Dr. Henrique da Mota tarafından Fransa, Lyon'da bulunan Massues Mesleki Eğitim ve Araştırma Merkezi'ndeki omurga ameliyatlarında tıbbi bir bölgede toplanmıştır. Bu veri tabanı omurganın sagital panoramik radyografilerinden elde edilen 310 hastayla ilgili verileri içermektedir. Verideki 100 hasta, omurgalarında herhangi bir patoloji olmayan gönüllü (normal) hastalardır. Geri kalan veriler, disk hernisi (60 hasta) veya spondilolistezi (150 hasta) nedeni ile ameliyat edilen hastalardan elde edilmiştir. Bu nedenle, veri tabanı 210 anormal hastadan oluşmaktadır.

Veri tabanındaki her hasta, leğen kemiğine (pelvis) ve bel kemiğine (lomber) ait omurganın şekli ve hizalamasından kaynaklanan 6 biyomekanik özellik ile aşağıdaki gibi karakterize edilmiştir.

- Pelvis/leğen kemiği insidansı (pelvic incidence, PI),
- Pelvis eğimi (pelvic tilt, PT),
- Lordoz/bel kemiği açısı (lumbar lordosis, LuL),
- Sakral/kuyruk sokumu eğim (sacral slope, SS),
- Pelvis yarıçapı (pelvic radius, PR)
- Spondilolistezi /kayma derecesi (degree spondylolisthesis, DS)

Çalışmada kullanılan örneklem EK-1'de verilmiştir.

5.1.2. Diskriminant Analizi Uygulaması

Diskriminant analizi uygulanmasından önce analize ait tüm varsayımlar (normallik, aykırı değer ve çoklu ilişki) test edilmiştir.

Diskriminant Analizinin varsayımlarından biri olan değişkenlerin normalliğinin sınanması için; tanımsal istatistikler hesaplanmış, normallik grafikleri ve histogramlar elde edilmiştir ve değişkenlere tek tek Kolmogorov-Smirnov testi uygulanmıştır.

1.tip hata katsayısı %5 olmak üzere, Kolmogorov-Smirnov testi sonuçları ve tanımsal istatistikler Tablo 5.1'de verilmiştir. Normallik testi sonucu tüm değişkenlerin anlamlılık düzeyi 0,05 ten küçük çıkmıştır. Büyük örneklemlerde veri normal dağılsa bile bu değerler anlamlı çıkmamaktadır. Bu nedenle çok değişkenli analiz için sürekli değişkenlerin normalliğe sahip olup olmadığını test etmenin bir başka yolu olan çarpıklık ve basıklık değerlerine göre değerlendirme yapılmıştır. Çarpıklık değerleri -1,5 ve 1,5 arasında ve basıklık değerleri -2 ve 2 arasında olan verinin normal dağıldığı varsayımı yapılır (George ve Mallery, 2010).

Tablo 5.1.Vertebral Kolon Verisinin Tanımsal İstatistikleri

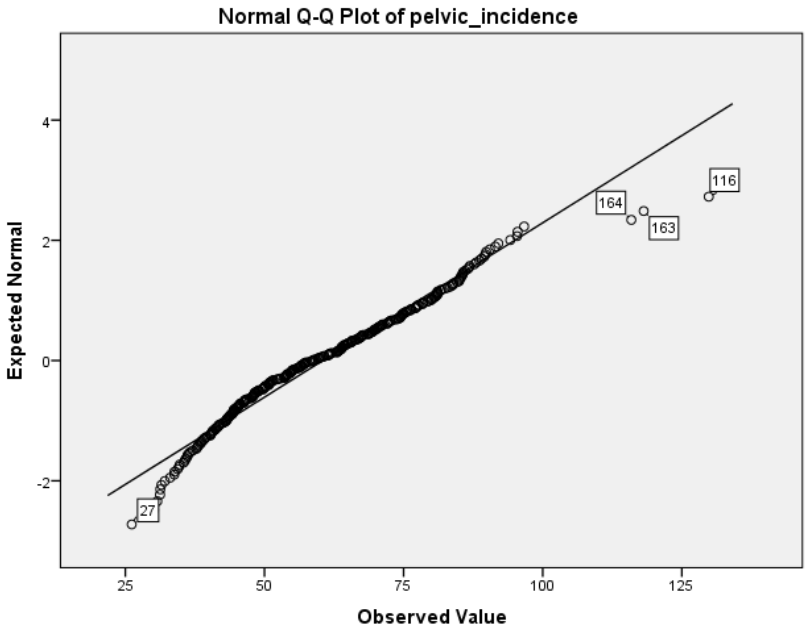
Değişken	K.S. İst.*	K.S. Anlamlılık**	Ortalama	Standart Sapma	Çarpıklık Katsayısı	Basıklık Katsayısı
PI	0,071	0,001	60,496	17,236	0,520	0,224
PT	0,083	0,000	17,543	10,008	0,677	0,676
LuL	0,069	0,001	51,931	18,554	0,599	0,162
SS	0,058	0,014	42,954	13,423	0,793	**3,008**
PR	0,059	0,011	117,92	13,318	-0,177	0,935
DS	0,173	0,000	26,297	37,559	**4,318**	**38,068**

*Kolmogorov-Smirnov Test İstatistiği

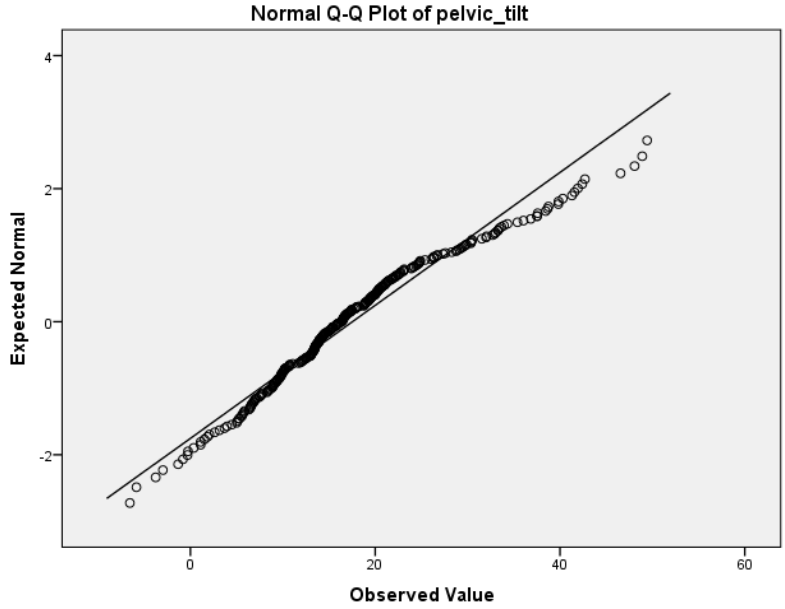
**Kolmogorov-Smirnov Anlamlılık Değeri

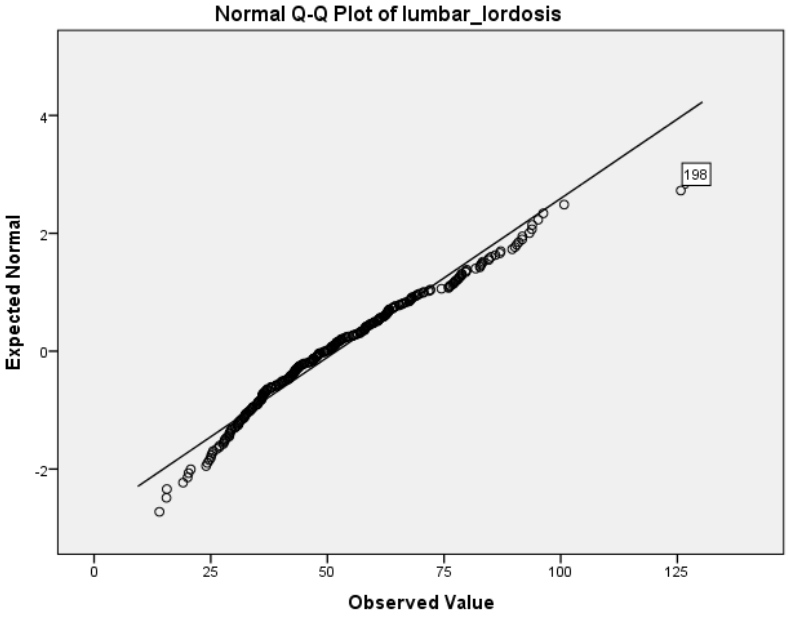
Analiz sonuçlarına göre sakral/kuyruk sokumu eğimi (sacral slope) ve spondilolistezi / kayma derecesi değişkenlerinde 2'ten büyük çarpıklık ve basıklık değerlerine rastlanmıştır.

Ayrıca değişkenlerin doğrusallıkları Normal Q-Q grafikleri ile incelenmiş ve 116. verinin pelvis insidansı, sakral eğim ve spondilolistezi derecesi değişkenlerinde ortak aykırı gözleme sahip olduğu saptanmıştır.

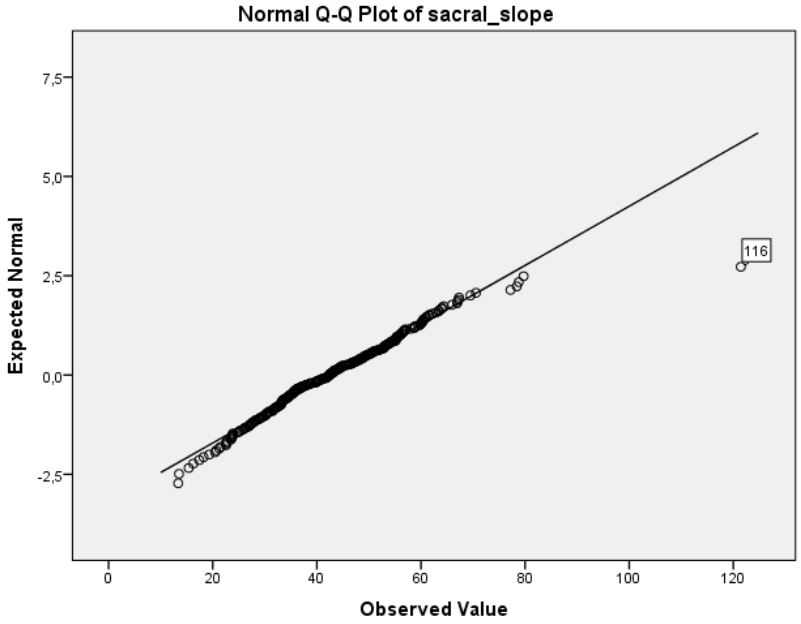


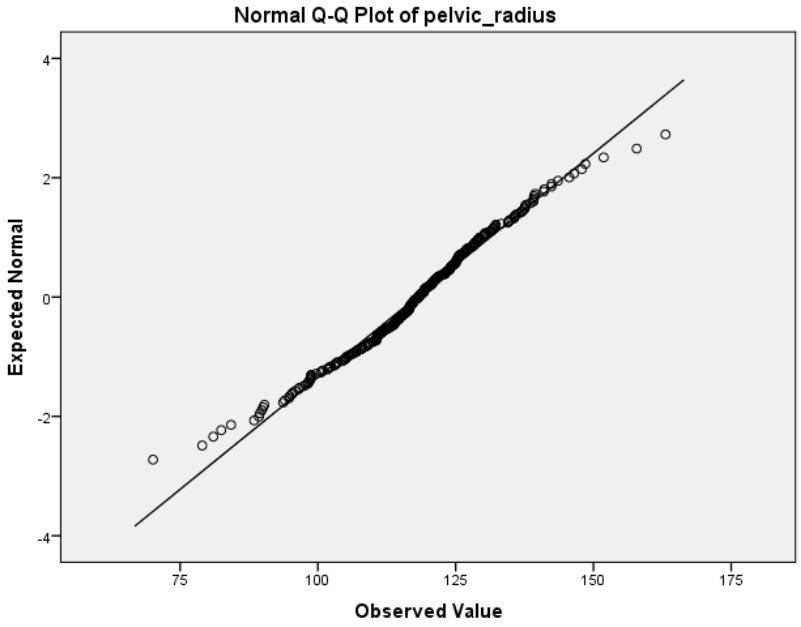
Şekil 5.1.'in başlangıcı



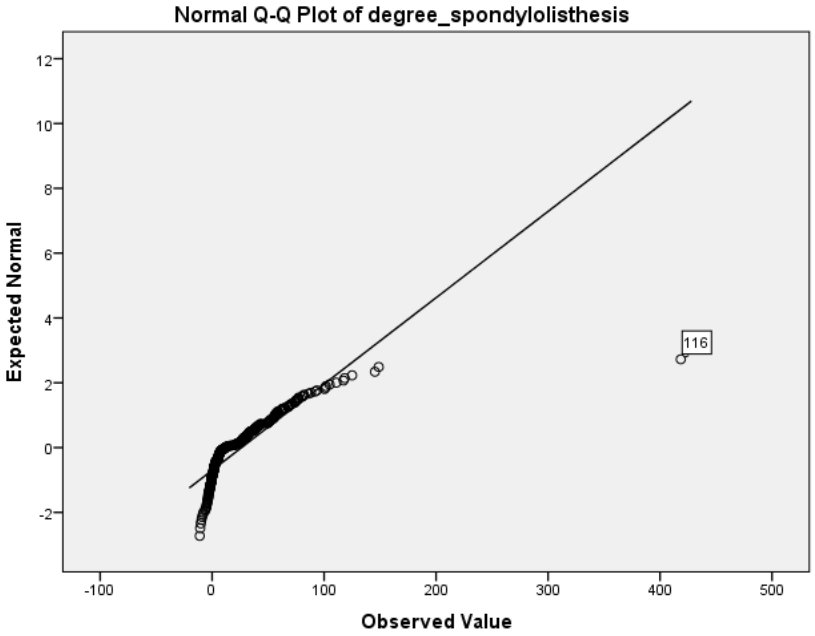


Şekil 5.1.'in devamı





Şekil 5.1.'in devamı



Şekil 5.1. Vertebral Kolon Verisinde Normallik Varsayımı Testi İçin Elde Edilen Normal Q-Q Grafikleri

Bu aykırı gözlem değerlerine sahip bel kayması (Spondilolistezi) olan 116. hastanın verisi hem pelvis insidansı, sakral eğim ve spondilolistezi derecesi değişken ölçümleri silinerek hem de tüm değişken ölçümlerinden yani 116. hastanın verisi tamamen silinerek tekrar normallik testleri yapılmıştır.

Yapılan iki analiz sonuçlarında da normallik test sonuçları ve sonrasında diskriminant analizi varsayımları için yapılan korelasyon, Box's M ve sınıflandırma sonuçları gibi neredeyse tüm analiz sonuçlarının aynı olduğu tespit edilmiştir. Bu sonuçlara dayanarak bundan sonraki diskriminant analizi uygulamaları 116. hastanın verisi tamamen silinerek elde edilen sonuçları içermektedir.

Aykırı gözlem değerlerine sahip, bel kayması (spondilolistezi) olan 116. hastanın verisi silinerek tekrar tanımsal istatistikler hesaplanmış ve normallik grafikleri incelenmiştir. Bulgular incelendiğinde, değişkenlerin normalleştiği

görülmektedir. Aykırı gözlemin silinmesi sonrası tanımsal istatistikler Tablo 5.2' de verilmiştir.

Tablo 5.2. Düzeltilmiş Vertebral Kolon Verisinin Tanımsal İstatistikleri

Değişken	Ortalama	Medyan	Standart Sapma	Çarpıklık Katsayısı	Basıklık Katsayısı
PI	60,272	58,600	16,804	**0,374**	**-0,358**
PT	17,572	16,420	10,011	**0,672**	**0,674**
LuL	51,942	49,780	18,583	**0,597**	**0,151**
SS	41,280	42,370	12,676	**0,229**	**-0,261**
PR	117,953	118,340	13,326	**-0,183**	**0,938**
DS	25,643	11,460	30,234	**1,296**	**1,593**

Tablodan görüldüğü gibi 2'ten büyük çarpıklık ve basıklık değerlerine saptanmamıştır. Ayrıca normallik grafikleri de incelenmiş ve aykırı gözlem sorununun ortadan kalktığı belirlenmiştir.

Çoklu doğrusal bağlantı sorunu olup olmadığının test edilmesi için, değişkenler arası korelasyon katsayılarından oluşan korelasyon matrisi hesaplanmıştır. Korelasyon katsayıları Tablo 5.3. te verilmiştir.

Değişkenler arası korelasyon katsayılarının -0.34 ile 0.80 arasında değiştiği ve pelvis insidansı biyomekanik özelliğinin diğer değişkenlerle 0,70 den büyük yüksek ilişki gösterdiği (Tablo 5.3) belirlenmiştir. Buradan hareketle, veri setinde çoklu ilişki (multicollinearity) bulunduğu ve pelvis insidansı (PI) değişkeninin modelden çıkarılarak analize devam edilmesine karar verilmiştir.

Tablo 5.3. Değişkenler Arası Korelasyon Katsayıları

	pelvic incidence	pelvic tilt	lumbor lordosis	sacral slope	pelvic radius	degree spondylo listhesis
pelvic incidence	1.000					
pelvic tilt	0.659	1.000				
lumbor lordosis	**0.739**	0.432	1.000			
sacral slope	**0.805**	-0.085	0.639	1.000		
pelvic radius	-0.244	0.0305	-0.081	-0.347	1.000	
degree spondylolisthesis	0.6421	0.534	0.672	0.429	-0.000	1.000

Diskriminant analizinin uygulanması aşamasında çoklu regresyona benzer yöntem olan stepwise metodu kullanılmıştır. Burada tahmin edici değişkenler modele gruplar arasındaki diskriminant (ayrıştırma) güçlerine göre art arda girerler. F değeri her bir tahmin edici için hesaplanır ve tahmin edici değişkenler, kriter değişken gibi değerlendirilerek tek değişkenli varyans analizi uygulanır. En büyük F değerine sahip tahmin edici değişken diskriminant fonksiyonuna dahil edilir. İkinci tahmin edici, önceden belirlenmiş kısmi F değerini aşıyorsa modele dahil edilir. Seçilen her bir tahmin edici, önceden eklenen tahmin edici değişkenlerin varlığında gereksiz hale gelip gelmediğini görmek için yeniden incelenir.

Dahil etme işlemi her aşamada tüm istatistikler hesaplanarak tüm tahmin ediciler için devam eder. Bu işlem, giriş için mevcut değişkenler arasındaki en büyük kısmi F önceden belirlenmiş eşik değeri aşmadığında sonlandırılır. Böylece modele dahil edilen ve dahil edilmeyen iki tahmin edici grubu oluşur. Stepwise metodu belirlenen kritere göre uygulanır. Bu kriterler aşağıdaki gibidir.

a) Wilk's Lamda Kriteri
b) Mahalanobis D^2 Kriteri

c) En Küçük F Kriteri

Çalışmamızdaki diskriminant analizi uygulamasında tüm bu kriterlerin sonuçları elde edilmiş ve ilk olarak Wilk's Lamda ve Mahalanobis D^2 kriterleri karşılaştırılmıştır. Wilk's Lamda kriterinin sınıflandırma sonucunun daha yüksek çıkmasına rağmen, tahmin edici değişkenlerin seçiminde analizde değerli görülen pelvis eğimi (pelvic tilt) ve lordoz/bel kemiği açısı (lumbar-lordosis) değişkenlerini diskriminant model tahminine dahil etmediği görülmüştür. İlk aşamada değişkenler için gerçekleştirilen çoklu ilişki varsayımı testi de göz önüne alınarak, Mahalanobis D^2 kriterinin Wilk's Lamda kriterine göre daha gerçekçi sonuçlar verdiği öngörülmüştür.

Ek olarak En Küçük F kriteri sonuçlarının, Mahalanobis D^2 kriteri ile karşılaştırıldığında uygulamaya dahil edilecek tahmin edici değişken seçimi ve sınıflandırma sonuçları açısından bir farklılık göstermedikleri belirlenmiştir. Bu nedenle uygulanan analiz stepwise metodu Mahalanobis D^2 kriterine göre yapılmış ve analiz sonuçları yorumlanmıştır.

Diskriminant analizinde ilk olarak kovaryans matrislerinin eşitliği varsayımının test edilmesi gerekir. Bu varsayımın sınanması için Box's M testi uygulanmıştır. Test sonuçları Tablo 5.4' te verilmiştir.

Testin sıfır ve alternatif hipotezi aşağıdaki şekildedir:

H_0: Bağımlı değişken tarafından oluşturulan gruplar arasında kovaryans matrisleri farklılık göstermemektedir.

H_1: Bağımlı değişken tarafından oluşturulan gruplar arasında kovaryans matrisleri farklılık göstermektedir.

Tablodan görüldüğü gibi 0,05 anlamlılık düzeyinde sıfır hipotezi reddedilmektedir. Yani, grup kovaryans matrisleri eşit değildir.

Tablo 5.4. Box's M Testi Sonuçları

Box's M Değeri	400,310
F (Yaklaşık) Değeri	12,978
Serbestlik Derecesi 1	30
Serbestlik Derecesi 2	126436,215
Anlamlılık	0,000

Daha önce de ifade edildiği gibi diskriminant analizi yönteminin temel varsayımları, değişkenlerin normal dağılması ve grupların ortak kovaryans matrisli olmalarıdır. Bu varsayımları sağlayan diskriminant fonksiyonu lineer diskriminant fonksiyonu (LDF) ve bu sınıflandırma fonksiyonu ile yapılan analiz lineer diskriminant analizi (LDA) olarak adlandırılır. Bu varsayımlardan eşit grup kovaryans matrisleri varsayımının sağlanmaması durumunda yani, gruplar arası kovaryans matrislerinin farklı olmaları durumunda kullanılan fonksiyon karesel diskriminant fonksiyonu (QDF) ve bu sınıflandırma fonksiyonu ile yapılan analiz ise karesel diskriminant analizi (QDA) olarak adlandırılır.

Bu çalışmada yöntemlerin sınıflandırma sonuçlarındaki farklılığını ortaya koyabilmek amacıyla, ilk aşamada eşit kovaryans varsayımının sağlanmadığı ihlal edilerek LDA sınıflandırma sonuçları elde edilmiştir. Daha sonra bu varsayımın sağlanmaması durumunda önerilen QDA sınıflandırma sonuçlarına yer verilmiştir.

Diskriminant analizinde tahmin edilen diskriminant modeli için, uyum iyiliği istatistikleri, model katsayıları, sınıflandırma fonksiyonları ve sınıflandırma tablosuna yer verilmiştir.

Diskriminant fonksiyonunun öneminin değerlendirilmesi için; kanonik korelasyon, özdeğer ve Wilk's Lamda istatistikleri incelenmiştir. Özdeğer ve kanonik korelasyon katsayısı Tablo 5.5' te verilmiştir.

Tablo 5.5. Özdeğer ve Kanonik Korelasyon

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	2,530[a]	91,4	91,4	,847
2	,239[a]	8,6	100,0	,439

a. First 2 canonical discriminant functions were used in the analysis.

Kanonik korelasyon diskriminant skorları ve gruplar arasındaki ilişkiyi ölçer ve açıklanan toplam varyansı gösterir. Bu değerin yorumlanabilmesi için karesinin alınması gereklidir.

Birinci kanonik korelasyon değeri 0,847 olup $(0,847)^2$=0,72 dir. Buna göre birinci diskriminant fonksiyonu bağımlı değişkendeki varyansın %72'sini açıklamaktadır.

Diskriminant analizinde genel istatistiksel anlamlılığı değerlendirme kriterlerinin MANOVA'dakilerle benzer olduğu 2. bölümde ifade edilmişti.

Diskriminant analizinde Wilk's Lamda (Wilk's Λ) aynı zamanda özdeğerlerin bir alt kümesi üzerinde kullanılabildiği için; Roy en büyük karakteristik kökü (Roy's θ), Hotelling izi (Lawley-Hotelling $U^{(s)}$) ve Pillai kriteri (Pillai $V^{(s)}$) gibi diğer üç MANOVA testinden daha yararlıdır. MANOVA anlamlılık test sonuçları Tablo 5.6'da verilmiştir.

Wilk's Lamda istatistiği, grup ortalamalarının birbirinden farklı olup olmadıklarını ve diskriminant skorlarındaki toplam varyansın gruplar arasındaki farklar tarafından açıklanamayan kısmını gösterir. Tablo 5.6'dan görüldüğü gibi diskriminant fonksiyonunda toplam varyansın %23'ü gruplar arasındaki farklar tarafından açıklanamamaktadır. Aynı zamanda $p<0,05$ olup her bir diskriminant fonksiyonu için özdeğer istatistiğinin anlamlı olduğu ve bu fonksiyonların ayırıcı modeller oluşturduğu kabul edilmiştir.

Tablo 5.6. Wilk's Lambda Sonuçları

Number of obs = 309

W = Wilks' lambda L = Lawley-Hotelling trace
P = Pillai's trace R = Roy's largest root

Source	Statistic		df	F(df1,	df2) =	F	Prob>F	
class	W	0.2286	2	10.0	604.0	65.93	0.0000	e
	P	0.9097		10.0	606.0	50.57	0.0000	a
	L	2.7694		10.0	602.0	83.36	0.0000	a
	R	2.5303		5.0	303.0	153.33	0.0000	u
Residual			306					
Total			308					

e = exact, a = approximate, u = upper bound on F

Bağımsız değişkenlerin öneminin değerlendirilmesi için diskriminant fonksiyonu katsayılarına (Tablo 5.7) ve yapı matrisindeki (Tablo 5.8) her bir değişken yüküne bakılması gerekir.

Tablo 5.7. Standartlaştırılmış Kanonik Diskriminant Fonksiyonu Katsayıları

	Function	
	1	2
pelvic tilt	-,256	-,406
lumbar lordosis	,168	,244
sacral slope	,321	,659
pelvic radius	-,330	,821
degree spondylolisthesis	,968	-,303

Mutlak değerce büyük katsayı, bağımsız değişkenin önemliliğine işaret eder. Katsayıların işaretleri ise ilişkinin yönünü göstermektedir. Standartlaştırılmış diskriminant fonksiyonları arasında yer almayan değişkenin etkili olmadığı sonucuna varılır.

Tablo 5.7'ye göre analize dahil ettiğimiz 5 biyomekanik değişkeninde etkili değişkenler olduğu görülmektedir. Pelvis insidansı değişkeni modelden çıkarılmadan gerçekleştirilen diskriminant analizi sonuçlarında da Tablo 5.7'de verilen fonksiyon matrisi elde edilmiş, burada da pelvis insidansı değişkeninin yer almadığı ve bu değişkenin etkili olmadığı desteklenmiştir. Buradaki katsayılar regresyon analizindeki beta katsayılarına karşılık gelmektedir.

Mutlak değerce en büyük katsayının birinci diskriminant fonksiyonunda bel kayması derecesi (degree spondylolisthesis) değişkenine ait olduğu, dolayısıyla burada bağımlı değişkeni gruplara ayırmada en önemli bağımsız değişkenin bel kayması derecesi olduğu görülmektedir. Benzer şekilde ikinci fonksiyon için ise ayırmada en önemli değişken pelvis yarıçapı (pelvic radius) dır. Buna karşılık, ayrıma en az katkıyı sağlayan bağımsız değişken, birinci ve ikinci diskriminant fonksiyonu için lordoz/bel kemiği açısı (lumbar lordosis) olmuştur.

Bağımsız değişkenlerin öneminin değerlendirilmesinde kullanılabilecek başka bir ölçüt yapı matrisidir. Tablo 5.8'de yapı matrisinden elde edilen korelasyonlar verilmiştir. Bu matris, her bir değişkenin diskriminant fonksiyonu ile korelasyonunu gösterir.

Yapı matrisine göre, birinci diskriminant fonksiyonu ile mutlak değerce en yüksek korelasyona sahip bağımsız değişkenler bel kayması derecesi (degree spondylolisthesis) ve lordoz/bel kemiği açısı (lumbar-lordosis), en düşük korelasyona sahip değişken pelvis yarıçapı (pelvic radius) değişkenidir.

Tablo 5.8. Yapı Matrisi

	Function	
	1	2
degree spondylolisthesis	,783*	-,176
lumbar lordosis	,539*	,299
sacral slope	,478	,523*
pelvic radius	-,154	,451*
pelvic tilt	,204	-,392*

*Her bir değişken ve herhangi bir değişken arasındaki mutlak değerce en yüksek korelasyon

LDA için Fisher'in lineer diskriminant fonksiyonu katsayılarına Tablo 5.9'da yer verilmiştir. Her bir grup için sınıflandırma fonksiyonları değişkenlerin aşağıda belirtilen kısaltmaları kullanılarak elde edilmiştir.

- Pelvis/leğen kemiği insidansı (pelvic incidence, PI),
- Pelvis eğimi (pelvic tilt, PT),
- Lordoz/bel kemiği açısı (lumbar lordosis, LuL),
- Sakral/kuyruk sokumu eğim (sacral slope, SS),
- Pelvis yarıçapı (pelvic radius, PR)
- Spondilolistezi /kayma derecesi (degree spondylolisthesis, DS)

Tablo 5.9. Sınıflandırma Fonksiyonu Katsayıları (Fisher'in Lineer Diskriminant Fonksiyonu Katsayıları)

	class		
	Disk Hernia	Spondylolisthesis	Normal
pelvic tilt	,340	,214	,272
lumbar lordosis	-,049	,005	-,021
sacral slope	,783	,945	,885
pelvic radius	1,024	,989	1,106
degree spondylolisthesis	-,267	-,109	-,274
(Constant)	-74,854	-80,852	-87,828

Diskriminant analizinde, elde edilen k-1 diskriminant fonksiyonlarına ihtiyaç duyulmaz. Gözlemleri gruplara atamada tüm k sınıflandırma fonksiyonlarının kullanılmaları gerekir.

Bel fıtığı (disk hernia) grubu için lineer sınıflandırma fonksiyonu;

$$y_1 = -74,854 + 0,340 * \mathrm{PT} - 0,049 * \mathrm{LuL} + 0,783 * \mathrm{SS} + 1,024 * \mathrm{PR} - 0,267 * \mathrm{DS}$$

Bel kayması (spondylolisthesis) grubu için lineer sınıflandırma fonksiyonu;

$$y_2 = -80,852 + 0,214 * \mathrm{PT} + 0,005 * \mathrm{LuL} + 0,945 * \mathrm{SS} + 0,989 * \mathrm{PR} - 0,109 * \mathrm{DS}$$

Normal grubu için lineer sınıflandırma fonksiyonu;

$$y_3 = -87,828 + 0,272 * \mathrm{PT} - 0,021 * \mathrm{LuL} + 0,885 * \mathrm{SS} + 1,106 * \mathrm{PR} - 0,274 * \mathrm{DS}$$

şeklinde yazılır.

Sınıflandırma fonksiyonları yeni gözlemlerin sınıflandırılmasında kullanılan fonksiyonlardır. Katsayıları yorumlanamamaktadır.

Analizin doğru sınıflandırma yüzdesinin, yani analizin başarısının değerlendirilmesi için sınıflandırma tablosu elde edilmiştir. LDA sınıflandırma sonuçları Tablo 5.10'da verilmiştir.

Tablo 5.10. Lineer Diskriminant Analizi Sınıflandırma Tablosu

	Tahmin Grubu / Gerçek Grup	Disk Hernia	Spondylolisthesis	Normal	Toplam
Sayım	Disk Hernia	39	0	21	60
	Spondylolisthesis	4	135	10	149
	Normal	12	3	35	100
%	Disk Hernia	65,0	,0	35,0	100,0
	Spondylolisthesis	2,7	90,6	6,7	100,0
	Normal	12,0	3,0	85,0	100,0

Doğru sınıflandırma oranı %83.8'dır.

LDA ile elde edilen katsayı tahminleri yardımıyla her bir hasta için tahmin edilen grup üyeliklerinin olasılıkları EK-2'te verilmiştir. Tablo 5.10'dan görüldüğü gibi, lineer diskriminant analizinde bel fıtığı (disk hernia) olan hastaları %65, bel kayması (spondylolisthesis) olan hastalar %90,6 ve normal hastaların %85'i doğru sınıflandırılmıştır. Bu durumda bel fıtığı (disk hernia) olan hastaların %35 (n=21)'i, bel kayması (spondylolisthesis) olan hastaların %9,4 (n=14)'ü ve normal hastalarında %15 (n=15)'i hatalı sınıflandırılmıştır. Lineer diskriminant analizi sonucunda bel fıtığı, bel kayması ve normal olan hastaların toplamda doğru sınıflandırma oranı %83,8'dir.

Tablo 5.4'teki Box's M testi sonuçlarına göre grup kovaryans matrisleri eşittir hipotezi reddedilmişti. Bu durumda eşit grup kovaryans matrisleri varsayımının sağlanmamasından dolayı gruplar arası kovaryans matrislerinin farklı olmaları durumunda önerilen QDA uygulanmıştır.

5 biyomekanik değişkenin karesel fonksiyon olarak yazımındaki matematiksel zorluk nedeniyle karesel diskriminant fonksiyonlarının katsayıları kullanılan paket programlar aracılığıyla elde edilememiştir. Bu nedenle karesel sınıflandırma fonksiyonları da oluşturulamamıştır. Bu analiz sonuçları için sadece sınıflandırma sonuçları yorumlanabilmektedir.

Tablo 5.11'de QDA sonuçlarına göre elde edilen sınıflandırma tablosuna yer verilmiştir.

Tablo 5.11. Karesel Diskriminant Analizi (QDA) Sınıflandırma Tablosu

Gerçek Grup \ Tahmin Grubu		Disk Hernia	Spondylolisthesis	Normal	Toplam
Sayım	Disk Hernia	41	1	18	60
	Spondylolisthesis	1	146	2	149
	Normal	14	5	81	100
%	Disk Hernia	68,33	1,67	30,0	100,0
	Spondylolisthesis	0,67	97,99	1,34	100,0
	Normal	14,0	5,0	81,0	100,0

Doğru sınıflandırma oranı %86.7'dır.

QDA ile elde edilen katsayı tahminleri yardımıyla her bir hasta için tahmin edilen grup üyeliklerinin olasılıkları EK-2'te verilmiştir.. Tablo 5.11'dan görüldüğü gibi, karesel diskriminant analizinde bel fıtığı (disk hernia) olan hastaları %68, bel kayması (spondylolisthesis) olan hastalar %98 ve normal hastaların %81'i doğru sınıflandırılmıştır. QDA sonucunda bel fıtığı, bel kayması ve normal olan hastaların toplamda doğru sınıflandırma oranı %86,7'dir.

5.1.3. Lojistik Regresyon Analizi Uygulaması

Bu bölümde lojistik regresyon modeli tahmin edilmiş ve lojistik regresyon analizi ile elde edilen sınıflandırma tablosu sonuçları verilmiştir.

İlk olarak, bağımlı değişkenin multinominal (2'den fazla sınıflı) olmasından dolayı uygulama öncesi bu veri için MNL modelin doğru olup olmadığının göstergesi olan IIA: İlgisiz Alternatiflerin Bağımsızlık (Independent of Irrelevant Alternatives) varsayımının geçerliliği Hausman-McFadden testi ile araştırılmıştır. Bu bağımsızlık varsayımının sağlanma durumu, özellikle bu varsayımı gerektirmeyen MNP model ile karşılaştırma yapabilmek amacıyla test edilmiştir.

Elde edilen test sonucunda;

H_0: Katsayılardaki farklılık sistematik değildir (IIA varsayımı sağlanır)

biçiminde kurulan hipotezi test eden Hausman-McFadden Test istatistik değeri chi2(5)=21,31 (>0,05) bulunmuş ve sıfır hipotezi kabul edilmiştir. Sonuç olarak kullanılan veri seti için IIA varsayımının sağlandığı ve bu veri setine MNL model uygulanabileceğine karar verilmiştir.

Veri setini en iyi açıklayan modelin belirlenmesinde, yaygın olarak kullanılan ölçütlerden birisi Akaike bilgi kriteri (Akaike Information Criteria=AIC), diğeri de Bayesian bilgi kriteri (Bayesian Information Criteria) dir. Karşılaştırmak istenen modellerden; uygun modelin seçiminde, her iki ölçüt bakımından düşük değerli olan model tercih edilir. Diğer bir ifade ile daha az sayıda açıklayıcı değişken içeren ve uyumu iyi olan modelde küçük AIC ve BIC değerleri elde edilir. Eğer herhangi iki modelden de aynı AIC değeri elde edilmişse, bunlardan açıklayıcı değişken sayısı az olanı tercih edilir (Lawles, 1987: Wang ve Putterman, 1998).

Tablo 5.12 incelendiğinde, sadece sabit terim içeren modele göre son modelden elde edilen AIC ve BIC değerlerinin daha düşük olduğu görülmektedir. Buradan yalnızca sabit terim içeren modele göre son modelin uyumunun daha iyi olduğu söylenebilir. Ayrıca olabilirlik oran test (Likelihood Ratio Test) istatistiği de anlamlı bulunmuştur.

Tablo 5.12. Model Uyum Bilgisi

Model	Model Fitting Criteria			Likelihood Ratio Tests		
	AIC	BIC	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	643,674	651,140	639,674			
Final	207,044	259,311	179,044	460,630	12	,000

Sadece sabit terim içeren modelin -2*LL* (Log Likelihood) değeri ile bağımsız değişkenlerin de ilave edildiği modelin –2*LL* değeri arasındaki fark 460,63 bulunmuştur. %5 anlamlılık düzeyinde "bağımsız değişkenler içeren modelin, sadece sabit terim içeren modelden anlamlı farklılık göstermediği" hipotezi reddedilir. Diğer bir deyişle; bağımsız değişkenlerden en az biri bağımlı değişkenle anlamlı derecede ilişkilidir yorumu yapılır.

Tablo 5.13. Pseudo R-Kare

Cox and Snell	,775
Nagelkerke	,887
McFadden	,720

Standart regresyon analizindeki R^2'ye benzer olarak lojistik regresyonda da yapay (pseudo) R^2 değerleri elde edilmektedir. Bunlardan en yaygın olanları Tablo 5.13'te verilmiştir. Burada standart regresyon analizindeki R^2 den daha küçük R^2 değerleri elde edilir ve bu değerler 1'e ulaşmaz. Bu durum, lojistik regresyon analizindeki modelin daha zayıf olduğu ve sonuçlarının da daha zayıf modele göre elde edildiği şeklinde yanlış anlaşılmamalıdır.

Lojistik regresyon parametrelerinin anlamlılığını test eden Tablo 5.14'de bulunan Wald istatistiği, β'nın anlamlılığına ilişkin bir ölçüdür. Bağımsız değişkenlerin β katsayılarının, sıfırdan anlamlı bir şekilde farklılık gösterip göstermediğini tespit etmeye çalışarak her bir değişkenin modele katkısını açıklar.

Lojistik regresyon katsayılarının, standart hatasına oranı;

$$W = Z = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$$

Wald istatistiği olarak adlandırılır. Bu istatistik standart normal dağılımın kritik değerleri ile karşılaştırılarak anlamlılığı belirlenir. Teorik anlatımda da belirtildiği üzere Wald istatistiği

$$W = Z^2 = \left(\frac{\boldsymbol{\beta}}{SE(\boldsymbol{\beta})} \right)^2$$

şeklinde de elde edilebilmektedir. Bu durumda ise Wald test istatistiği 1 serbestlik dereceli χ^2 dağılımı gösterir ve 1 serbestlik dereceli χ^2 dağılımının kritik değerleri ile karşılaştırılarak anlamlılığı belirlenir.

Lojistik regresyon parametrelerinin anlamlılığını test etmede kullanılan diğer ölçüt ise hesaplanan p değerleridir. %5'ten küçük olan p değerlerine ($p<0,05$) sahip değişkenler model için anlamlıdır.

Tablo 5.14'te verilen Wald istatistik değerleri Z^2 değerlerine göre hesaplanmıştır. Wald istatistik değerleri %5 anlamlılık düzeyinde $\chi_1^2 = 3,84$ değerinden büyük olan değişkenler anlamlı model parametrelerine sahiptir yorumu yapılır.

Elde edilen tablo sonucunda hem p değerlerine hem de Wald istatistik değerlerine bakıldığında sırasıyla, bel fıtığı (disk hernia) sınıfı için PI, PT, LuL, SS ve DS değişkenlerinin ve bel kayması(spondylolisthesis) sınıfı için ise PI, PT, LuL, SS ve PR değişkenlerinin bağımlı değişkenin üzerinde anlamlı olmadığı görülmektedir. Diğer bir ifadeyle, bel fıtığı (disk hernia) sınıfı için sadece pelvis yarıçapı (PR) bel kayması(spondylolisthesis) sınıfı için ise sadece kayma derecesi (DS) anlamlı bulunmuştur.

Tablo 5.14. Lojistik Regresyon Parametre Tahminleri

Sınıf[a]		β	Std. Error	Wald	df	p	Exp(β)
DH[b]		20,223	4,264	22,488	1	,000	
Intercept		-3,553	40,656	,008	1	,930	,029
	PI	3,653	40,657	,008	1	,928	38,579
	PT	-,035	,030	1,344	1	,246	,966
	LuL	3,400	40,655	,007	1	,933	29,976
	SS	-,130	,029	20,130	1	,000	,878
	PR	,006	,039	,024	1	,978	1,006
	DS						
SP[c]		-1,563	4,971	,099	1	,753	
Intercept		23,744	73,039	,106	1	,745	2,05E+10
	PI	-23,656	73,052	,105	1	,746	5,33E-11
	PT	-,013	,045	,082	1	,775	,987
	LuL	-23,688	73,044	,105	1	,746	5,16E-11
	SS	-,052	,030	3,037	1	,081	,949
	PR	,315	,054	33,753	1	,000	1,371
	DS						

a. Normal referans grubudur.
b. DH: Disk Hernia
c. SP: Spondylolisthesis

Anlamlılık testi sonuçlarında analizde bulunması gereken bağımsız değişkenlerin anlamlı çıkmamasından dolayı Olabilirlik Oran Testi uygulanmıştır. Değişkenli ve değişkensiz modele göre ayrı ayrı $D = -2LL$ sapma değeri hesaplanmış ve elde edilen

$$G = D(Değişkensiz\ Model) - D(Değişkenli\ Model)$$

G istatistiği χ^2 tablo değeri ile karşılaştırılarak bağımsız değişkenlerin modelde olup olmamasına karar verilmiştir.

İlk olarak lojistik regresyon analizinde model parametreleri arasında anlamlı çıkmayan pelvis insidansı (PI) değişkeni modelden çıkarılmış ve lojistik regresyon analizi yeniden uygulanmıştır.

Tablo 5.15. PI Değişkeni Çıkarılarak Elde Edilen Model Uyum Bilgisi

Model	Model Fitting Criteria			Likelihood Ratio Tests		
	AIC	BIC	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	643,674	651,140	639,674			
Final	203,170	247,970	179,170	460,504	10	,000

Yeni oluşan modelin *-2LL* sapma değeri $D(Değişkensiz\ Model) = 179{,}170$ olarak elde edilmiştir. Hesaplanan *G* istatistik değeri;

$$G = D(Değişkensiz\ Model) - D(Değişkenli\ Model)$$
$$= 179{,}170 - 179{,}044 = 0{,}126$$

$\chi^2_{6-5;0,05} = 3{,}84$ tablo değerinden küçük olduğundan olabilirlik oran testi sonucunda PI değişkeni anlamsız çıkmıştır. Bu testi desteklemek amacıyla ayrıca doğru sınıflandırma oranları karşılaştırılmıştır. Tüm değişkenlerin olduğu modele göre lojistik regresyon analizi doğru sınıflandırma sonucu %87.4 olarak elde edilmiştir. PI değişkeni model dışı bırakıldığında elde edilen sınıflandırma sonucu da %87.4 olarak bulunmuş olup, sınıflandırma sonucuna da bir etkisi olmamasından dolayı PI değişkeninin modelde olmaması gerektiğine karar verilmiştir.

Tablo 5.16'da PI değişkeninin modelden çıkarılarak elde edilen parametre tahminlerine göre, bel fıtığı (disk hernia) ve bel kayması (spondylolisthesis) sınıfları için oluşturulan her iki sınıf modelinde hem Wald istatistik değeri hem de p değerine göre LuL (bel kemiği açısı) değişkeninin ortak anlamsız değişken olduğu görülmektedir.

Bir sonraki aşamada LuL değişkeni de modelden çıkarılmış ve lojistik regresyon analizi yeniden uygulanmıştır.

Tablo 5.16. PI Değişkeni Çıkarılarak Elde Edilen Lojistik Regresyon Parametre Tahminleri

Sınıf[a]		β	Std. Error	Wald	df	p	Exp(β)
DH[b]	Intercept	20,192	4,248	22,597	1	,000	
	PT	,100	,039	6,625	1	,010	1,105
	LuL	-,036	,030	1,388	1	,239	,965
	SS	-,152	,041	13,820	1	,000	,859
	PR	-,130	,029	20,200	1	,000	,878
	DS	,006	,039	,021	1	,885	1,006
SP[c]	Intercept	-1,182	4,801	,061	1	,805	
	PT	,092	,061	2,287	1	,130	1,096
	LuL	-,017	,044	,152	1	,696	,983
	SS	-,057	,065	,778	1	,378	1,059
	PR	-,055	,029	3,666	1	,056	,947
	DS	,313	,054	34,078	1	,000	1,368

a. Normal referans grubudur.
b. DH: Disk Hernia
c. SP: Spondylolisthesis

Tablo 5.17. PI ve LuL Değişkenleri Çıkarılarak Elde Edilen Model Uyum Bilgisi

Model	Model Fitting Criteria			Likelihood Ratio Tests		
	AIC	BIC	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	643,674	651,140	639,674			
Final	200,663	237,996	180,663	459,011	8	,000

Yeni oluşan modelin *-2LL* sapma değeri $D(Değişkensiz\ Model) = 180,663$ olarak elde edilmiştir. Hesaplanan *G* istatistik değeri;

$$G = D(Değişkensiz\ Model) - D(Değişkenli\ Model)$$
$$= 180,663 - 179,170 = 1,493$$

$\chi^2_{5-4;0,05} = 3,84$ tablo değerinden küçük olduğundan LuL değişkeni de anlamsız çıkmıştır. LuL değişkeninin olduğu modele göre, bu değişkenin model dışı bırakılması sonucu elde edilen sınıflandırma sonucu da %87,4 olup, sınıflandırma

sonuçlarında da değişiklik olmamasından dolayı LuL değişkeninin de modelde olmaması gerektiğine karar verilmiştir.

Tablo 5.18'de PI ve LuL değişkenlerinin modelden çıkarılarak elde edilen parametre tahminlerinde Wald istatistik değeri ve p değerine göre, bel fıtığı (disk hernia) sınıfı için DS değişkeninin ve bel kayması (spondylolisthesis) sınıfı için ise PT ve SS değişkenlerinin anlamsız değişkenler olduğu görülmektedir.

Tablo değerine göre anlamsız bulunan bu değişkenler teker teker modelden çıkarılıp, analizler yeniden uygulanarak bu değişkenlerin olabilirlik testi yardımıyla model için anlamlı olup olmadıkları belirlenmiştir.

Tablo 5.18. PI ve LuL Değişkenleri Çıkarılarak Elde Edilen Lojistik Regresyon Parametre Tahminleri

Sınıf[a]		β	Std. Error	Wald	df	p	Exp(β)
DH[b]	Intercept	20,856	4,249	24,091	1	,000	
	PT	,081	,035	5,343	1	,021	1,084
	SS	-,183	,033	30,480	1	,000	,833
	PR	-,136	,029	22,371	1	,000	,873
	DS	,000	,039	,000	1	,997	1,000
SP[c]	Intercept	-,515	4,754	,012	1	,914	
	PT	,083	,060	1,905	1	,167	1,086
	SS	,038	,043	,777	1	,378	1,038
	PR	-,059	,028	4,276	1	,039	,943
	DS	,311	,053	34,339	1	,000	1,368

a. Normal referans grubudur.
b. DH: Disk Hernia
c. SP: Spondylolisthesis

İlk olarak en düşük Wald istatistik değerine sahip olan DS açıklayıcı değişkeni modelden çıkarılmış ve lojistik regresyon analizi yeniden uygulanmıştır.

Yeni oluşan modelin *-2LL* sapma değeri $D(Değişkensiz\ Model) = 408,746$ olarak elde edilmiştir.

Tablo 5.19. PI, LuL ve DS Değişkenleri Çıkarılarak Elde Edilen Model Uyum Bilgisi

Model	Model Fitting Criteria			Likelihood Ratio Tests		
	AIC	BIC	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	643,674	651,140	639,674			
Final	424,746	454,613	408,746	230,928	6	,000

Hesaplanan *G* istatistik değeri;

$$G = D(\textit{Değişkensiz Model}) - D(\textit{Değişkenli Model})$$
$$= 408,746 - 180,663 = 228,083$$

$\chi^2_{4-3;0,05} = 3,84$ tablo değerinden büyük olduğundan DS değişkeni olabilirlik testine göre anlamlı çıkmıştır.Ayrıca bir önceki %87,4 sınıflandırma oranı ile karşılaştırıldığında, DS değişkeninin de model dışı bırakıldığında sınıflandırma oranını %71,8 olarak elde edilmiştir. G istatistik değerine göre anlamlı çıkan DS değişkeni, sınıflandırma sonucunu da zayıflattığı için modelde olması gerektiğine karar verilmiştir.

Tablo 5.20. PI, LuL ve DS Değişkenleri Çıkarılarak Elde Edilen Lojistik Regresyon Parametre Tahminleri

Sınıf[a]		β	Std. Error	Wald	df	p	Exp(β)
DH[b]	Intercept	16,382	2,966	30,497	1	,000	
	PT	,059	,023	6,435	1	,011	1,061
	SS	-,157	,028	30,883	1	,000	,855
	PR	-,103	,020	26,837	1	,000	,902
SP[c]	Intercept	,191	2,058	,009	1	,926	
	PT	,106	,019	31,753	1	,000	1,112
	SS	,093	,018	27,484	1	,000	1,097
	PR	-,048	,015	10,626	1	,001	,953

a. Normal referans grubudur.
b. DH: Disk Hernia
c. SP: Spondylolisthesis

Tablo 5.20'de PI, LuL ve DS değişkenlerinin modelden çıkarılarak elde edilen parametre tahminlerinde, Wald istatistik değeri ve p değerine göre modele dahil edilen 3 değişkenin de model için anlamlı olduğu görülmektedir.

Bu sonuçlardan yararlanarak, DS, PT, SS ve PR olmak üzere 4 açıklayıcı değişkeninde model için anlamlı olduğu sonucuna varılmıştır. Ancak daha az değişkenle optimal sınıflandırma tablosu elde edebilmek amacıyla PT ve SS değişkenleri model dışı bırakılarak da analizler tekrarlanmış ve bu değişkenlerin de model için anlamlı olup olmadıkları test edilmiştir.

PT değişkeni model dışı bırakıldığında elde edilen sonuçlar şöyledir:

Tablo 5.21. PI, LuL ve PT Değişkenleri Çıkarılarak Elde Edilen Model Uyum Bilgisi

Model	Model Fitting Criteria			Likelihood Ratio Tests		
	AIC	BIC	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	643,674	651,140	639,674			
Final	203,351	233,217	187,351	452,323	6	,000

Yeni oluşan modelin *-2LL* sapma değeri $D(Değişkensiz\ Model) = 187,351$ olarak elde edilmiştir. Hesaplanan *G* istatistik değeri;

$$G = D(Değişkensiz\ Model) - D(Değişkenli\ Model)$$
$$= 187,351 - 180,663 = 6,663$$

$\chi^2_{4-3;0,05} = 3,84$ tablo değerinden büyük olduğundan PT değişkeni olabilirlik testine göre anlamlı çıkmıştır.

Tablo 5.22. PI, LuL ve PT Değişkenleri Çıkarılarak Elde Edilen Lojistik Regresyon Parametre Tahminleri

Sınıf[a]		β	Std. Error	Wald	df	p	Exp(β)
DH[b]	Intercept	24,437	4,225	33,448	1	,000	
	SS	-,189	,033	31,963	1	,000	,828
	PR	-,154	,029	27,999	1	,000	,858
	DS	,005	,038	,016	1	,899	1,005
SP[c]	Intercept	2,534	4,593	,304	1	,581	
	SS	,040	,043	,884	1	,347	1,041
	PR	-,073	,029	6,255	1	,012	,930
	DS	,303	,049	38,161	1	,000	1,354

a. Normal referans grubudur.
b. DH: Disk Hernia
c. SP: Spondylolisthesis

Tablo 5.22'de PI, LuL ve PT değişkenlerinin modelden çıkarılmasıyla elde edilen parametre tahminlerinde, Wald istatistik değeri ve p değerine göre, bel fıtığı (disk hernia) sınıfı için DS ve bel kayması (spondylolisthesis) sınıfı için ise SS değişkenlerinin anlamsız değişkenler olduğu görülmektedir.

PT (pelvis eğim açısı) değişkeninin modelden çıkarılması olabilirlik testi sonucuna ek olarak, %87,4 sınıflandırma sonucunu değiştirmemesine rağmen bel

kayması ve bel fıtığı hastalıklarının teşhisindeki önemi açısından modele dahil edilmesi gerektiğine karar verilmiştir.

SS değişkeni model dışı bırakıldığında elde edilen sonuçlar şöyledir:

Tablo 5.23. PI, LuL ve SS Değişkenleri Çıkarılarak Elde Edilen Model Uyum Bilgisi

Model	Model Fitting Criteria			Likelihood Ratio Tests		
	AIC	BIC	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	643,674	651,140	639,674			
Final	255,374	285,241	239,374	400,299	6	,000

Yeni oluşan modelin *-2LL* sapma değeri $D(Değişkensiz\ Model) = 239,374$ olarak elde edilmiştir. Hesaplanan *G* istatistik değeri;

$$G = D(Değişkensiz\ Model) - D(Değişkenli\ Model)$$
$$= 239,374 - 180,663 = 58,711$$

$\chi^2_{4-3;0,05} = 3,84$ tablo değerinden büyük olduğundan SS değişkeni olabilirlik testine göre anlamlı çıkmıştır.

Tablo 5.24'de PI, LuL ve SS değişkenlerinin modelden çıkarılmasıyla elde edilen parametre tahminlerinde, Wald istatistik değeri ve p değerine göre, bel fıtığı (disk hernia) sınıfı için DS ve bel kayması (spondylolisthesis) sınıfı için ise PT değişkenlerinin anlamsız değişkenler olduğu görülmektedir.

Tablo 5.24. PI, LuL ve SS Değişkenleri Çıkarılarak Elde Edilen Lojistik Regresyon Parametre Tahminleri

Sınıf[a]		β	Std. Error	Wald	df	p	Exp(β)
DH[b]	Intercept	6,073	2,382	6,499	1	,011	
	PT	,082	,028	8,597	1	,003	1,085
	PR	-,065	,019	11,623	1	,001	,937
	DS	-,008	,032	,072	1	,789	,992
SP[c]	Intercept	1,373	3,370	,166	1	,684	
	PT	,098	,055	3,179	1	,075	1,103
	PR	-,062	,026	5,837	1	,016	,940
	DS	,310	,050	37,946	1	,000	1,363

a. Normal referans grubudur.
b. DH: Disk Hernia
c. SP: Spondylolisthesis

SS (sakral eğim) değişkeninin modele dahil edilmesi olabilirlik testi sonucuna ek olarak, model dışı bırakıldığında sınıflandırma sonucunu % 82,5'e zayıflatması açısından modele dahil edilmesi gerekiğine karar verilmiştir.

Bütün bu analizlerin sonunda lojistik regresyon modeli için PI ve LuL değişkenlerinin katsayıları anlamsız bulunarak PT, SS, PR ve DS bağımsız değişkenlerin katsayıları, oluşturulan lojistik regresyon modelleri için anlamlı bulunmuştur.

Tahmin edilen sınıflandırma fonksiyonlarını elde etmek için PI ve LuL değişkenlerinin model dışı bırakılarak elde edilmiş olan Tablo 5.18'den yararlanılmıştır.

$\boldsymbol{\beta}'_j = [\beta_{j0}, \beta_{j1}, \beta_{j2}, ..., \beta_{jp}]$, $j = 1, 2, 3$ için sınıflar 1: Disk Hernia, 2: Spondylolisthesis ve 3; Normal olarak kodlanmış, normal hastalar referans alınarak elde edilen bel fıtığı (disk hernia) ve bel kayması (spondylolisthesis) gruplarına ilişkin lojistik fonksiyonlar sırasıyla;

$$\ln\left(\frac{P(y_i=1|x_i)}{P(y_i=3|x_i)}\right) = 20,856 + 0,081*PT - 0,183*SS - 0,136*PR$$

$$\ln\left(\frac{P(y_i=2|x_i)}{P(y_i=3|x_i)}\right) = 0,083*PT + 0,038*SS - 0,059*PR + 0,311*DS$$

şeklinde oluşturulmuştur.

Katsayıların yorumlanması için Tablo 5.18'deki odds oranları (e^{β}) kullanılarak, bağımsız değişkenlerdeki değişikliklerin bağımlı değişken üzerindeki etkileri görülebilir.

Bel fıtığı (disk hernia) grubuna ilişkin modelde sabit terim anlamlı bulunmuştur ancak sabit terimi yorumlamak her zaman mümkün değildir. PT 'deki 1 birim artış bel fıtığı olma riskini normal bireye göre 1,084 kat arttırmakta, SS'deki 1 birim artış bel fıtığı olma riskini normal bireye göre 0,833 kat arttırmakta, PR'deki 1 birim artış bel fıtığı olma riskini normal bireye göre 0,873 kat arttırmaktadır.

Bel kayması (spondylolisthesis) grubuna ait tahmin modeline göre, PT 'deki 1 birim artış bel kayması olma riskini normal bireye göre 1,086 kat arttırmaktadır. SS'deki 1 birim artış bel kayması olma riskini normal bireye göre 1,038 kat arttırmakta, PR'deki 1 birim artış bel kayması olma riskini normal bireye göre 0,943 kat arttırmakta ve DS'deki 1 birim artış bel kayması olma riskini normal bireye göre 1,368 kat arttırmaktadır. İki sınıfında tahmininde ortak olarak en etkili bağımsız değişkenin PT olduğu söylenebilir.

Tablo 5.25. Lojistik Regresyon Analizi Sınıflandırma Tablosu

Gerçek Grup \ Tahmin Grubu		Disk Hernia	Spondylolisthesis		Toplam
Sayım	Disk Hernia	39	1	20	60
	Spondylolisthesis	1	145	3	149
	Normal	12	2	86	100
%	Disk Hernia	65	1,67	33,33	100,0
	Spondylolisthesis	0,67	97,32	2,01	100,0
	Normal	12,0	2,0	86,0	100,0

Doğru sınıflandırma oranı %87,4'dır.

Lojistik regresyon analizinde EÇO katsayı tahminleri yardımıyla her bir hasta için hesaplanan grup olasılıkları EK-2'de verilmiştir. Tablo 5.25'dan görüldüğü gibi, lojistik regresyon analizinde bel fıtığı (disk hernia) olan hastaların %65, bel kayması (spondylolisthesis) olan hastaların %97,3 ve normal hastaların %86'sı doğru sınıflandırılmıştır. Bu durumda bel fıtığı (disk hernia) olan hastaların %33,3 (n=20)'ü, bel kayması (spondylolisthesis) olan hastaların %2,7 (n=4)'si ve normal hastalarında %14 (n=14)'ü hatalı sınıflandırılmıştır. Lojistik regresyon analizi sonucunda bel fıtığı, bel kayması ve normal olan hastaların toplamda doğru sınıflandırma oranı %87,4 olarak elde edilmiştir.

5.1.4. Probit Regresyon Analizi Uygulaması

Probit regresyon analizi ile EÇO yöntemi kullanılarak elde edilen parametre tahminleri Tablo 5.26'da verilmiştir.

Tüm bağımsız değişkenler analize dahil edildiğinde probit model sonuçları lojistik model sonuçları ile benzerlik göstermiştir.

Tablo 5.26. Probit Regresyon Parametre Tahminleri

Sınıf[a]		β	Std. Error	z	p
DH[b]	Intercept	11,496	2,742	4,19	,000
	PI	-1,428	28,088	-,05	,959
	PT	1,486	28,089	,05	,958
	LuL	-,020	,020	-1,01	,314
	SS	1,339	28,088	,05	,962
	PR	-,074	,019	-3,95	,000
	DS	,003	,025	,11	,912
SP[c]	Intercept	-,085	2,882	-,03	,976
	PI	13,606	44,307	,31	,759
	PT	-13,558	44,314	-,31	,760
	LuL	-,012	,026	-,47	,640
	SS	-13,577	44,311	-,31	,759
	PR	-,029	,018	-1,70	,090
	DS	,156	,028	5,66	,000
N=309 Log likelihood= -89,070 Wald chi2(12)= 67,63 Probit>chi2= 0,000					

a. Normal referans grubudur.
b. DH: Disk Hernia
c. SP: Spondylolisthesis

Tablo 5.26'ya göre, bel fıtığı (disk hernia) sınıfı için PI, PT, LuL, SS ve DS değişkenlerinin ve bel kayması(spondylolisthesis) sınıfı için ise PI, PT, LuL, SS ve PR değişkenlerinin bağımlı değişkenin üzerinde anlamlı olmadığı görülmektedir.

Lojistik regresyonda gerçekleştirilen adımlara benzer şekilde, ilk olarak PI değişkeni daha sonra da LuL değişkeni model dışı bırakılarak model için anlamlı olup olmadıkları incelenmiştir. Sınıflandırma sonuçlarına etkileri de göz önüne alınarak, inceleme sonucunda PI ve LuL bağımsız değişkenlerinin modelde olmamasına karar verilmiştir.

PI ve LuL bağımsız değişkenleri modelden çıkarılarak, vertebral kolon verisine probit regresyon analizi yeniden uygulanmıştır. Elde edilen parametre tahminleri Tablo 5.27'de verilmiştir.

Tablo 5.27. PI ve LuL Değişkenleri Çıkarılarak Elde Edilen Probit Regresyon Parametre Tahminleri

Sınıf[a]		β	Std. Error	z	p
DH[b]	Intercept	11,799	2,734	4,32	,000
	PT	,047	,023	2,03	,042
	SS	-,106	,021	-4,93	,000
	PR	-,076	,018	-4,14	,000
	DS	-,001	,024	-,07	,940
SP[c]	Intercept	,372	2,786	,13	,894
	PT	,041	,035	1,20	,231
	SS	,014	,025	,57	,571
	PR	-,033	,016	-1,94	,052
	DS	,151	,026	5,89	,000
N=309 Log likelihood= -90,025 Wald chi2(12)= 68,54 Probit>chi2= 0,000					

a. Normal referans grubudur.
b. DH: Disk Hernia
c. SP: Spondylolisthesis

Tablo 5.27'e göre modelin uyum iyiliği p=0,00<0,05 olduğundan modelin anlamlı olduğu anlaşılmıştır. Parametre tahminlerinde bel fıtığı (disk hernia) sınıfı için DS ve bel kayması (spondylolisthesis) sınıfı için PT, SS ve PR değişkenlerinin anlamsız görünmesine rağmen, bu değişkenler tahmin olasılıklarındaki ve dolayısıyla sınıflandırma gücündeki etkilerinden dolayı analize dahil edilmişlerdir.

Probit modelde normal hastalar referans alınarak, tahmin edilen katsayılarla elde edilen bel fıtığı (disk hernia) grubuna ilişkin probit sınıflandırma fonksiyonu;

$$y_1 = 11,799 + 0,047 * \mathrm{PT} - 0,106 * \mathrm{SS} - 0,076 * \mathrm{PR} - 0,001 * \mathrm{DS}$$

bel kayması (spondylolisthesis) grubuna ilişkin probit sınıflandırma fonksiyonu;

$$y_2 = 0,372 + 0,041*\mathrm{PT} + 0,014*\mathrm{SS} - 0,033*\mathrm{PR} + 0,151*\mathrm{DS}$$

şeklindedir.

Probit model katsayıları doğrudan yorumlanamamaktadır. Tahmin edilen katsayı değerleri omurga hastalığı sınıflarının z skorlarını etkilemektedir.

Pelvis eğimindeki artış, normal hastalara göre bel fıtığı (disk hernia) hastalığı ve bel kayması (spondylolisthesis) hastalığı riskini arttırmaktadır. Sakral (kuyruk sokumu) eğimindeki artış, normal hastalara göre bel fıtığı (disk hernia) hastalığı riskini azaltmakta ancak bel kayması (spondylolisthesis) hastalığı riskini arttırmaktadır. Pelvis yarıçapındaki artış, normal hastalara göre bel fıtığı (disk hernia) hastalığı ve bel kayması (spondylolisthesis) hastalığı riskini azaltmaktadır. Kayması derecesindeki artış ise normal hastalara göre bel fıtığı (disk hernia) hastalığı riskini azaltmakta ancak bel kayması (spondylolisthesis) hastalığı riskini arttırmaktadır.

Probit regresyon analizinde EÇO katsayı tahminleri yardımıyla her bir hasta için hesaplanan grup olasılıkları EK-2'de verilmiştir. Bu tahmin olasılıklarına dayanarak oluşturulan sınıflandırma tablosu Tablo 5.28 ile elde edilmiştir.

Tablo 5.28. Probit Regresyon Analizi Sınıflandırma Tablosu

Gerçek Grup \ Tahmin Grubu		Disk Hernia	Spondylolisthesis		Toplam
Sayım	Disk Hernia	39	1	20	60
	Spondylolisthesis	1	145	3	149
	Normal	12	2	86	100
%	Disk Hernia	65	1,67	33,33	100,0
	Spondylolisthesis	0,67	97,32	2,01	100,0
	Normal	12,0	2,0	86,0	100,0

Doğru sınıflandırma oranı %87,4'dır.

Probit regresyon analizinde bel fıtığı (disk hernia) olan hastaların %65, bel kayması (spondylolisthesis) olan hastaların %97,3 ve normal hastaların %86'sı doğru sınıflandırılmıştır. Probit regresyon analizi sonucunda bel fıtığı, bel kayması ve normal olarak sınıflandırılan hastaların toplamda doğru sınıflandırma oranı %87,4 olarak elde edilmiştir. Lojistik regresyon analizi sınıflandırma sonucu ile aynı sonucu vermektedir.

5.1.5. Analiz Yöntemlerinin Karşılaştırılması

Gerçek veri seti üzerine yapılan uygulamalar sonucunda LDA, QDA, LRA ve PRA yöntemlerini karşılaştırabilmek için aşağıdaki genel tablo oluşturulmuştur.

Tablo 5.29. Gerçek Veri Uygulamasında Yöntemlerin Karşılaştırılması

Yöntem	LDA	QDA	LRA	PRA
Doğru Sınıflandırma Oranı (%)	83,8	86,7	87,4	87,4
Hatalı Sınıflandırma Oranı (%)	16,2	13,3	12,6	12,6

5.2. Simülasyon Veri Seti Üzerine Uygulama

5.2.1 Veri Seti İlişkin Tanımlamalar

Petrografide (yer kabuğu kütlelerini inceleyen bilim dalında) klasik bir ayırt edici analiz örneği QAPF diyagramıdır. Bu diyagram volkanik taşları (özellikle yerin derinliklerinde kayalaşmış plütonik taşları), %100 normalleştirilmiş kuvars (Q: quartz), alkali feldispat (A: alkali feldspar), pilajiyoklaz (P: plagioclase) ve feldispatımsıların (F: feldsparthoids) yüzdelikleri aracılığıyla sınıflandırır. QAPF diyagramını Uluslararası Jeoloji Birimi Birliği (International Union of Geological Sciences) plütonik (derinlik) taşları sınıflandırmak için sıklıkla kullanmaktadır. Örneğin; %50 kuvars, %30 alkali feldispat, %20 pilajiyoklaz ve %0 feldispatımsılar içeren bir plütonik taş granit taşı olarak adlandırılmaktadır.

Uygulamada QAPF yüzdelikleri temel alınarak, x_1 ve x_2 değişkenleri ile tanımlanan 30, 90 ve 180 birimlik sentetik granit taş örnekleri veri setleri oluşturulmuştur. Bu iki değişken oksit olarak ifade edilen iki kimyasal elementin yüzdeliklerini (ağırlık yüzdesi olarak) temsil etmektedir. Yapılan ön araştırmalardan bu taş örneklemlerinin, 3 farklı magmatik olay sırasında farklı zamanlarda oluşan 3 granitten geldiği varsayılmıştır. Bu granit örneklemlerinde ölçülen değerlerin normal dağıldığı varsayılmıştır. Veri setleri için kullanılan simülasyon kodu MATLAB Recipes for Earth Sciences (Trauth, 2015) kitabından elde edilmiştir.

5.2.2. N=30 Birimlik Örneklem İçin Yapılan Uygulamalar

3 farklı granit türü için, her bir granit grubunda 10'ar örneklem olacak şekilde toplam n=30 birimlik veri üretilmiştir. Üretilen bu veri için sırasıyla diskriminant analizi, lojistik ve probit regresyon analizleri uygulanmıştır.

Diskriminant analizi uygulaması için gerekli varsayımlar (normallik, aykırı değer ve çoklu ilişki) test edilmiştir.

Veri normallik varsayımı ile üretilmiş olmasına rağmen normallik testi uygulanmış çarpıklık ve basıklık değerlerine göre verinin normal dağıldığı varsayımı doğrulanmıştır. Değişkenlerin doğrusallıkları Normal Q-Q grafikleri ile incelenmiş aykırı gözlemlere rastlanmamıştır. 30 birimlik granit verisinin tanımsal istatistikleri Tablo 5.30'da verilmiştir.

Tablo 5.30. N=30 Birimli Granit Verisinin Tanımsal İstatistikleri

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness	Kurtosis
class	30	1	3	2,00	,830	,000	-1,554
x1	30	,69	8,43	4,5358	1,96917	,145	-,721
x2	30	-1,16	3,76	1,5953	1,34509	-,556	-,294

Çoklu doğrusal bağlantı sorunu olup olmadığının test edilmesi için, değişkenler arası korelasyon katsayılarından oluşan korelasyon matrisi hesaplanmıştır. Korelasyon katsayıları Tablo 5.31'de verilmiştir.

Tablo 5.31. N=30 Birimli Granit Verisinin Değişkenler Arası Korelasyon Katsayıları

		x1	x2
Correlation	x1	1,000	,262
	x2	,262	1,000

Değişkenler arası korelasyon katsayılarının 0,70 den büyük olmamasından dolayı değişkenlerin yüksek ilişki göstermediği belirlenmiştir.

Diskriminant analizi uygulaması için gerekli tüm varsayımları (normallik, aykırı değer ve çoklu ilişki) sağlayan 30 birimlik granit verisine diskriminant analizi uygulanmıştır.

Diskriminant analizinde ilk olarak kovaryans matrislerinin eşitliği varsayımının test edilmesi gerekir. Bu varsayımın sınanması için Box's M testi uygulanmıştır. Test sonuçları Tablo 5.32' te verilmiştir.

Tablodan görüldüğü gibi 0,05 anlamlılık düzeyinde sıfır hipotezi kabul edilir. Bu durumda gruplar arası kovaryans matrisleri farklılık göstermemektedir. Üretilen bu veri için LDA varsayımları sağlanmıştır.

Tablo 5.32. N=30 Birimli Granit Verisinin Box's M Testi Sonuçları

Box's M Değeri	5,525
F (Yaklaşık) Değeri	2,638
Serbestlik Derecesi 1	2
Serbestlik Derecesi 2	1640,250
Anlamlılık	0,072

Bu çalışmada yöntemlerin sınıflandırma sonuçlarındaki farklılığını ortaya koyabilmek amacıyla, ilk aşamada eşit kovaryans varsayımının sağlandığı LDA

sınıflandırma sonuçları elde edilmiştir. Daha sonra bu varsayımın sağlanmaması durumunda önerilen QDA sınıflandırma sonuçlarına yer verilmiştir.

Diskriminant fonksiyonunun öneminin değerlendirilmesi için; kanonik korelasyon, özdeğer ve Wilk's Lamda istatistikleri Tablo 5.33 ile verilmiştir.

Tablo 5.33. N=30 Birimli Granit Verisinin Özdeğer ve Wilk's Lamda İstatistikleri

Özdeğerler

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	1,755	,933	,933	,798
2	,1269	,067	1,000	,336

Wilks Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	,377	26,303	2	,000

Birinci korelasyon değeri 0,798 olup karesi $(0,798)^2$=0,64'dür. Buna göre birinci diskriminant fonksiyonu bağımlı değişkendeki varyansın %64'ünü açıklamaktadır. Wilk's Lambda tablosundan görüldüğü gibi diskriminant fonksiyonundaki toplam varyansın %38'i gruplar arasındaki farklar tarafından açıklanamamaktadır. Aynı zamanda $p<0,05$ olup her bir diskriminant fonksiyonu için özdeğer istatistiğinin anlamlı olduğu ve bu fonksiyonların ayırıcı model oluşturduğu kabul edilmiştir.

LDA için Fisher'in lineer diskriminant fonksiyonu katsayılarına Tablo 5.34'de yer verilmiştir.

LDA ile elde edilen katsayı tahminleri yardımıyla her bir granit türü için tahmin edilen grup üyeliklerinin olasılıkları EK-3'te verilmiştir.

Tablo 5.34. N=30 Birimli Granit Verisinin LDA Sınıflandırma Fonksiyonu Katsayıları

	class		
	Granite_1	Granite_2	Granite_3
x1	1,245	1,883	1,299
x2	2,073	2,643	-,425
(Constant)	-4,639	-9,212	-2,154

LDA'nın sınıflandırma tablosu Tablo 5.35'te verilmiştir. Tablodan da görüldüğü gibi 30 birimlik granit simülasyon verisinin lineer diskriminant analizine göre doğru sınıflandırma oranı %73,3'tür.

Tablo 5.35. N=30 Birimli Granit Verisinin LDA Sınıflandırma Tablosu

Gerçek Grup / Tahmin Grubu		Granite_1	Granite_2	Granite_3	Toplam
Sayım	Granite_1	7	3	0	10
	Granite_2	3	7	0	10
	Granite_3	2	0	8	10
%	Granite_1	70	30	0	100,0
	Granite_2	30	70	0	100,0
	Granite_3	20	0	80	100,0

Doğru sınıflandırma oranı %73.3'tür.

Bu simülasyon verisine karşılaştırma yapabilmek amacıyla, eşit kovaryans varsayımının sağlanmaması durumunda önerilen QDA da uygulanmıştır. QDA sonuçlarına göre sınıflandırma tablosu Tablo 5.36 'da elde edilmiştir. QDA ile elde edilen katsayı tahminleri yardımıyla her bir granit türü için tahmin edilen grup üyeliklerinin olasılıkları EK-3'te verilmiştir.

Tablo 5.36. N=30 Birimli Granit Verisinin QDA Sınıflandırma Tablosu

Gerçek Grup \ Tahmin Grubu		Granite_1	Granite_2	Granite_3	Toplam
Sayım	Granite_1	6	3	1	10
	Granite_2	3	7	0	10
	Granite_3	2	0	8	10
%	Granite_1	60	30	10	100,0
	Granite_2	30	70	0	100,0
	Granite_3	20	0	80	100,0

Doğru sınıflandırma oranı %70'dir.

Tablodan da görüldüğü gibi 30 birimlik granit simülasyon verisinin karesel diskriminant analizine göre doğru sınıflandırma oranı %70'dir.

İkinci analiz olarak bu veri setine lojistik regresyon analizi (LRA) uygulanmıştır. Granite_3 grubu referans grup olarak alınmış, tahmini lojistik sınıflandırma modelleri elde edilerek LRA sınıflandırma sonuçları verilmiştir.

Yanıt değişkeninin multinominal olmasından dolayı IIA varsayımının geçerliliği Hausman-McFadden testi ile araştırılmıştır. 3 gruplu yanıt değişkeni için Hausman-McFadden test istatistik değeri chi2(3)=-0,08 bulunmuş ve chi2<0 olduğundan bağımsızlık varsayımı reddedilmiştir. Bu sonuç bu veri seti için MNL modelin uygun olmadığını göstermektedir. Ancak probit regresyon analizi ile sonuçları karşılaştırma yapabilmek amacıyla bu varsayım bu veri seti için ihlal edilmiştir.

Lojistik regresyon ve probit regresyon analizlerinin her ikisinden de parametre tahminleri elde edilmiş ve iki analiz için de x_1 açıklayıcı değişkenin istatistiksel olarak anlamsız olduğu görülmüştür. Ancak jeolojik bilgiye dayanarak sınıflandırmalarında x_1 ölçümlerinin önemli olmasından dolayı iki açıklayıcı değişkenin de model için anlamlı olduğu öngörülmüştür.

Tablo 5.37. N=30 Birimli Granit Verisinin Lojistik Regresyon Parametre Tahminleri

Sınıf[a]		β	Std. Error	Wald	df	p	Exp(β)
Granite_1	Intercept	-3,466	2,381	2,119	1	,146	
	x1	,110	,431	,065	1	,798	1,117
	x2	2,432	1,079	5,079	1	,024	11,383
Granite_2	Intercept	-7,766	3,081	6,355	1	,012	
	x1	,578	,497	1,349	1	,245	1,782
	x2	3,258	1,327	6,024	1	,014	25,999

a. Granite_3 referans grubudur.

Tablo 5.38. N=30 Birimli Granit Verisinin LRA Sınıflandırma Tablosu

Gerçek Grup \ Tahmin Grubu		Granite_1	Granite_2	Granite_3	Toplam
Sayım	Granite_1	6	3	1	10
	Granite_2	2	7	1	10
	Granite_3	2	0	8	10
%	Granite_1	60	30	10	100,0
	Granite_2	20	70	10	100,0
	Granite_3	20	0	80	100,0

Doğru sınıflandırma oranı %70'dir.

LRA ile elde edilen katsayı tahminleri yardımıyla her bir granit türü için tahmin edilen grup üyeliklerinin olasılıkları EK-3'te verilmiştir. Tablo 5.38'den görüldüğü gibi 30 birimlik granit simülasyon verisinin lojistik regresyon analizine göre doğru sınıflandırma oranı %70'dir.

PRA uygulaması sonucu elde edilen parametre tahminleri ve sınıflandırma tablosu Tablo 5.39 ve Tablo 5.40'da verilmiştir.

Tablo 5.39. N=30 Birimli Granit Verisinin Probit Regresyon Parametre Tahminleri

Sınıf[a]		β	Std. Error	z	p
Granite_1	Intercept	-2,713	1,824	-1,49	,137
	x1	,083	,337	,25	,806
	x2	1,847	,782	2,36	,018
Granite_2	Intercept	-6,325	2,453	-2,58	,010
	x1	,502	,378	1,33	,185
	x2	2,493	,976	2,55	,011

a. Granite_3 referans grubudur.

Tablo 5.40. N=30 Birimli Granit Verisinin PRA Sınıflandırma Tablosu

Gerçek Grup \ Tahmin Grubu		Granite_1	Granite_2	Granite_3	Toplam
Sayım	Granite_1	6	3	1	10
	Granite_2	2	7	1	10
	Granite_3	2	0	8	10
%	Granite_1	60	30	10	100,0
	Granite_2	20	70	10	100,0
	Granite_3	20	0	80	100,0

Doğru sınıflandırma oranı %70'dir.

PRA ile elde edilen katsayı tahminleri yardımıyla her bir granit türü için tahmin edilen grup üyeliklerinin olasılıkları EK-3'te verilmiştir. Tablo 5.40'dan görüldüğü gibi 30 birimlik granit simülasyon verisinin probit regresyon analizine göre doğru sınıflandırma oranı %70'dir.

5.2.3. N=90 Birimlik Örneklem İçin Yapılan Uygulamalar

3 farklı granit türü için, her bir granit grubunda 30'ar örneklem olacak şekilde toplam n=90 birimlik veri üretilmiştir. Üretilen bu veri için sırasıyla diskriminant analizi, lojistik ve probit regresyon analizleri uygulanmıştır.

Diskriminant analizi uygulaması için gerekli varsayımlar (normallik, aykırı değer ve çoklu ilişki) test edilmiştir. Verinin tüm gerekli varsayımları sağladığı görülmüştür.

90 birimlik granit verisinin tanımsal istatistikleri Tablo 5.41'de verilmiştir.

Tablo 5.41. N=90 Birimli Granit Verisinin Tanımsal İstatistikleri

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness	Kurtosis
class	90	1	3	2,00	,821	,000	-1,517
x1	90	-6,933	9,7308	4,0984	2,1510	,121	-,092
x2	90	-2,4958	4,4513	1,5590	1,589	-,540	-,164

Çoklu doğrusal bağlantı sorunu olup olmadığının test edilmesi için, değişkenler arası korelasyon katsayılarından oluşan korelasyon matrisi hesaplanmıştır. Korelasyon katsayıları Tablo 5.42'de verilmiştir.

Tablo 5.42. N=90 Birimli Granit Verisinin Değişkenler Arası Korelasyon Katsayıları

		x1	x2
Correlation	x1	1,000	,140
	x2	,140	1,000

Değişkenler arası korelasyon katsayılarının 0,70 den büyük olmamasından dolayı değişkenlerin yüksek ilişki göstermediği belirlenmiştir.

Diskriminant analizi uygulaması için gerekli tüm varsayımları (normallik, aykırı değer ve çoklu ilişki) sağlayan 90 birimlik granit verisine diskriminant analizi uygulanmıştır.

Diskriminant analizinde ilk olarak kovaryans matrislerinin eşitliği varsayımının test edilmesi gerekir. Bu varsayımın sınanması için Box's M testi uygulanmıştır. Test sonuçları Tablo 5.43' te verilmiştir.

Tablo 5.43. N=90 Birimli Granit Verisinin Box's M Testi Sonuçları

Box's M Değeri	34,050
F (Yaklaşık) Değeri	5,486
Serbestlik Derecesi 1	6
Serbestlik Derecesi 2	188642,769
Anlamlılık	0,000

Tablodan görüldüğü gibi 0,05 anlamlılık düzeyinde sıfır hipotezi reddedilir. Bu durumda gruplar arası kovaryans matrisleri farklılık göstermektedir. Üretilen bu veri için LDA varsayımları sağlanmamıştır. Ancak yöntemlerin sınıflandırma sonuçlarındaki farklılığını ortaya koyabilmek amacıyla, ilk aşamada eşit kovaryans varsayımının sağlandığı LDA sınıflandırma sonuçları ve daha sonra bu varsayımın sağlanmaması durumunda önerilen QDA sınıflandırma sonuçlarına yer verilmiştir.

Diskriminant fonksiyonunun öneminin değerlendirilmesi için; kanonik korelasyon, özdeğer ve Wilk's Lamda istatistikleri Tablo 5.44 ile verilmiştir.

Tablo 5.44. N=90 Birimli Granit Verisinin Özdeğer ve Wilk's Lamda İstatistikleri

Özdeğerler

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	1,937	99,8	99,8	,812
2	,003	,2	100,0	,054

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	,339	93,444	4	,000

Birinci korelasyon değeri 0,812 olup karesi $(0,812)^2$=0,66'dür. Buna göre birinci diskriminant fonksiyonu bağımlı değişkendeki varyansın %66'sını açıklamaktadır. Wilk's Lambda tablosundan görüldüğü gibi diskriminant fonksiyonundaki toplam varyansın %34'i gruplar arasındaki farklar tarafından

açıklanamamaktadır. Aynı zamanda $p<0,05$ olup her bir diskriminant fonksiyonu için özdeğer istatistiğinin anlamlı olduğu ve bu fonksiyonların ayırıcı model oluşturduğu kabul edilmiştir.

LDA için Fisher'in lineer diskriminant fonksiyonu katsayılarına Tablo 5.45'de yer verilmiştir.

LDA ile elde edilen katsayı tahminleri yardımıyla her bir granit türü için tahmin edilen grup üyeliklerinin olasılıkları EK-4'te verilmiştir.

Tablo 5.45. N=90 Birimli Granit Verisinin LDA Sınıflandırma Fonksiyonu Katsayıları

	class		
	Granite_1	Granite_2	Granite_3
x1	1,132	1,509	,816
x2	2,006	3,545	,045
(Constant)	-3,958	-9,221	-1,079

Tablo 5.46. N=90 Birimli Granit Verisinin LDA Sınıflandırma Tablosu

Gerçek Grup \ Tahmin Grubu		Granite_1	Granite_2	Granite_3	Toplam
Sayım	Granite_1	26	2	2	30
	Granite_2	5	25	0	30
	Granite_3	6	1	27	30
%	Granite_1	86,67	6,67	6,67	100,0
	Granite_2	16,67	83,33	0	100,0
	Granite_3	20,00	3,33	76,67	100,0

Doğru sınıflandırma oranı %82,2'dir.

LDA'nın sınıflandırma tablosu Tablo 5.46'da verilmiştir. Tablodan da görüldüğü gibi 90 birimlik granit simülasyon verisinin lineer diskriminant analizine göre doğru sınıflandırma oranı %82,2'dir.

Bu simülasyon eşit kovaryans varsayımının sağlanmaması durumunda önerilen QDA da uygulanmış ve analiz sonuçlarına göre sınıflandırma tablosu

Tablo 5.47'de elde edilmiştir. QDA ile elde edilen katsayı tahminleri yardımıyla her bir granit türü için tahmin edilen grup üyeliklerinin olasılıkları EK-4'te verilmiştir.

Tablo 5.47. N=90 Birimli Granit Verisinin QDA Sınıflandırma Tablosu

Gerçek Grup \ Tahmin Grubu		Granite_1	Granite_2	Granite_3	Toplam
Sayım	Granite_1	26	2	2	30
	Granite_2	4	26	0	30
	Granite_3	6	1	23	30
%	Granite_1	86,67	6,67	6,67	100,0
	Granite_2	13,33	86,67	0	100,0
	Granite_3	20,00	3,33	76,67	100,0

Doğru sınıflandırma oranı %83,3'tür.

Tablodan da görüldüğü gibi 90 birimlik granit simülasyon verisinin karesel diskriminant analizine göre doğru sınıflandırma oranı %83,3'tür.

İkinci analiz olarak bu veri setine LRA uygulanmıştır. Granite_3 grubu referans grup olarak alınmış, tahmini lojistik sınıflandırma modelleri elde edilerek LRA sınıflandırma sonuçları verilmiştir.

Yanıt değişkeninin multinominal olmasından dolayı IIA varsayımının geçerliliği Hausman-McFadden testi ile araştırılmıştır. 3 gruplu yanıt değişkeni için Hausman-McFadden test istatistik değeri chi2(3)=2,19 (p>chi2=0,53) olduğundan bağımsızlık varsayımı kabul edilmiştir. Bu sonuç bu veri seti için MNL modelin uygun olduğunu göstermektedir

LRA uygulaması sonucu elde edilen parametre tahminleri ve sınıflandırma tablosu Tablo 5.48 ve Tablo 5.49'da verilmiştir.

Tablo 5.48. N=90 Birimli Granit Verisinin Lojistik Regresyon Parametre Tahminleri

Sınıf[a]		β	Std. Error	Wald	df	p	Exp(β)
Granite_1	Intercept	-3,214	1,117	8,279	1	,004	
	x1	,299	,219	1,874	1	,171	1,349
	x2	1,966	,552	12,685	1	,000	7,139
Granite_2	Intercept	-12,713	2,556	24,736	1	,000	
	x1	,842	,302	7,799	1	,005	2,321
	x2	5,099	,991	26,471	1	,000	163,896

b. Granite_3 referans grubudur.

LRA ile elde edilen katsayı tahminleri yardımıyla her bir granit türü için tahmin edilen grup üyeliklerinin olasılıkları EK-4'te verilmiştir. Tablo 5.49'dan görüldüğü gibi 90 birimlik granit simülasyon verisinin lojistik regresyon analizine göre doğru sınıflandırma oranı %83,3'tür.

Tablo 5.49. N=90 Birimli Granit Verisinin LRA Sınıflandırma Tablosu

Gerçek Grup / Tahmin Grubu		Granite_1	Granite_2	Granite_3	Toplam
Sayım	Granite_1	26	2	2	30
	Granite_2	5	25	0	30
	Granite_3	5	1	24	30
%	Granite_1	86,67	6,67	6,67	100,0
	Granite_2	16,67	83,33	0	100,0
	Granite_3	16,67	3,33	80,00	100,0

Doğru sınıflandırma oranı %83,3'dir.

PRA uygulaması sonucu elde edilen parametre tahminleri ve sınıflandırma tablosu Tablo 5.50 ve Tablo 5.51'de verilmiştir.

PRA ile elde edilen katsayı tahminleri yardımıyla her bir granit türü için tahmin edilen grup üyeliklerinin olasılıkları EK-4'te verilmiştir.

Tablo 5.50. N=90 Birimli Granit Verisinin Probit Regresyon Parametre Tahminleri

Sınıf[a]		β	Std. Error	z	p
Granite_1	Intercept	-2,113	,726	-2,91	,004
	x1	,229	,158	1,45	,146
	x2	1,114	,305	3,66	,000
Granite_2	Intercept	-8,746	1,638	-5,34	,000
	x1	,655	,214	3,06	,002
	x2	3,195	,552	5,79	,000

b. Granite_3 referans grubudur.

Tablo 5.51'den görüldüğü gibi 90 birimlik granit simülasyon verisinin probit regresyon analizine göre doğru sınıflandırma oranı %82,2'dir.

Tablo 5.51. N=90 Birimli Granit Verisinin PRA Sınıflandırma Tablosu

	Tahmin Grubu / Gerçek Grup	Granite_1	Granite_2	Granite_3	Toplam
Sayım	Granite_1	25	3	2	30
	Granite_2	5	25	0	30
	Granite_3	5	1	24	30
%	Granite_1	83,33	10,0	6,67	100,0
	Granite_2	16,67	83,33	0	100,0
	Granite_3	16,67	3,33	80,00	100,0

Doğru sınıflandırma oranı %82,2'dir.

5.2.4. N=180 Birimlik Örneklem İçin Yapılan Uygulamalar

3 farklı granit türü için, her bir granit grubunda 60'ar örneklem olacak şekilde toplam n=180 birimlik veri üretilmiştir. Üretilen bu veri için sırasıyla diskriminant analizi, lojistik ve probit regresyon analizleri uygulanmıştır.

Diskriminant analizi uygulaması için gerekli varsayımlar (normallik, aykırı değer ve çoklu ilişki) test edilmiştir. Verinin tüm gerekli varsayımları sağladığı görülmüştür.

180 birimlik granit verisinin tanımsal istatistikleri Tablo 5.52'de verilmiştir.

Tablo 5.52. N=180 Birimli Granit Verisinin Tanımsal İstatistikleri

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness	Kurtosis
class	180	1	3	2,00	,819	,000	-1,508
x1	180	-1,48	10,73	3,93	2,026	,074	,845
x2	180	-2,70	3,91	1,59	1,325	-,754	,443

Çoklu doğrusal bağlantı sorunu olup olmadığının test edilmesi için, değişkenler arası korelasyon katsayılarından oluşan korelasyon matrisi hesaplanmıştır. Korelasyon katsayıları Tablo 5.53'de verilmiştir.

Tablo 5.53. N=180 Birimli Granit Verisinin Değişkenler Arası Korelasyon Katsayıları

		x1	x2
Correlation	x1	1,000	-,043
	x2	-,043	1,000

Değişkenler arası korelasyon katsayılarının 0,70 den büyük olmamasından dolayı değişkenlerin yüksek ilişki göstermediği belirlenmiştir.

Diskriminant analizi uygulaması için gerekli tüm varsayımları (normallik, aykırı değer ve çoklu ilişki) sağlayan 180 birimlik granit verisine diskriminant analizi uygulanmıştır.

Diskriminant analizinde ilk olarak kovaryans matrislerinin eşitliği varsayımının test edilmesi gerekir. Bu varsayımın sınanması için Box's M testi uygulanmıştır. Test sonuçları Tablo 5.54' te verilmiştir.

Tablo 5.54. N=180 Birimli Granit Verisinin Box's M Testi Sonuçları

Box's M Değeri	76,568
F (Yaklaşık) Değeri	12,553
Serbestlik Derecesi 1	6
Serbestlik Derecesi 2	780815,077
Anlamlılık	0,000

Tablodan görüldüğü gibi 0,05 anlamlılık düzeyinde sıfır hipotezi reddedilir. Bu durumda gruplar arası kovaryans matrisleri farklılık göstermektedir. Üretilen bu veri için LDA varsayımları sağlanmamıştır. Ancak bir önceki analizlerde olduğu gibi yöntemlerin sınıflandırma sonuçlarındaki farklılığını ortaya koyabilmek amacıyla, ilk aşamada eşit kovaryans varsayımının sağlandığı LDA sınıflandırma sonuçları ve daha sonra bu varsayımın sağlanmaması durumunda önerilen QDA sınıflandırma sonuçlarına yer verilmiştir.

Diskriminant fonksiyonunun öneminin değerlendirilmesi için; kanonik korelasyon, özdeğer ve Wilk's Lamda istatistikleri Tablo 5.55 ile verilmiştir.

Birinci korelasyon değeri 0,838 olup karesi $(0,838)^2$=0,70'dür. Buna göre birinci diskriminant fonksiyonu bağımlı değişkendeki varyansın %70'ini açıklamaktadır. Wilk's Lambda tablosundan görüldüğü gibi diskriminant fonksiyonundaki toplam varyansın %30'u gruplar arasındaki farklar tarafından açıklanamamaktadır. Aynı zamanda $p<0,05$ olup her bir diskriminant fonksiyonu için özdeğer istatistiğinin anlamlı olduğu ve bu fonksiyonların ayırıcı model oluşturduğu kabul edilmiştir.

Tablo 5.55. N=180 Birimli Granit Verisinin Özdeğer ve Wilk's Lamda İstatistikleri

Özdeğerler				
Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	2,359	99,3	99,3	,838
2	,017	,7	100,0	,128

Wilks' Lambda				
Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	,293	216,767	4	,000

LDA için Fisher'in lineer diskriminant fonksiyonu katsayılarına Tablo 5.56'de yer verilmiştir.

Tablo 5.56. N=180 Birimli Granit Verisinin LDA Sınıflandırma Fonksiyonu Katsayıları

	class		
	Granite_1	Granite_2	Granite_3
x1	1,213	1,750	,918
x2	2,959	4,993	,466
(Constant)	-4,712	-11,729	-1,371

LDA ile elde edilen katsayı tahminleri yardımıyla her bir granit türü için tahmin edilen grup üyeliklerinin olasılıkları EK-5'te verilmiştir.

LDA'nın sınıflandırma tablosu Tablo 5.57'de verilmiştir. Tablodan da görüldüğü gibi 180 birimlik granit simülasyon verisinin lineer diskriminant analizine göre doğru sınıflandırma oranı %83,9'dur.

Bu simülasyona eşit kovaryans varsayımının sağlanmaması durumunda önerilen QDA da uygulanmış ve analiz sonuçlarına göre sınıflandırma tablosu Tablo 5.58'de elde edilmiştir. QDA ile elde edilen katsayı tahminleri yardımıyla her bir granit türü için tahmin edilen grup üyeliklerinin olasılıkları EK-5'te verilmiştir.

Tablo 5.57. N=180 Birimli Granit Verisinin LDA Sınıflandırma Tablosu

Gerçek Grup \ Tahmin Grubu		Granite_1	Granite_2	Granite_3	Toplam
Sayım	Granite_1	53	4	3	60
	Granite_2	6	54	0	60
	Granite_3	16	0	44	60
%	Granite_1	88,33	6,67	5,00	100,0
	Granite_2	10,00	90,00	0	100,0
	Granite_3	26,67	32,22	26,11	100,0

Doğru sınıflandırma oranı %83,9'dur.

Tablo 5.58. N=180 Birimli Granit Verisinin QDA Sınıflandırma Tablosu

Gerçek Grup \ Tahmin Grubu		Granite_1	Granite_2	Granite_3	Toplam
Sayım	Granite_1	49	5	6	60
	Granite_2	5	55	0	60
	Granite_3	13	2	45	60
%	Granite_1	81,67	8,33	10,00	100,0
	Granite_2	8,33	91,67	0	100,0
	Granite_3	21,67	3,33	75,00	100,0

Doğru sınıflandırma oranı %82,8'dir.

Tablodan da görüldüğü gibi 90 birimlik granit simülasyon verisinin karesel diskriminant analizine göre doğru sınıflandırma oranı %82,8'dir.

İkinci analiz olarak bu veri setine LRA uygulanmıştır. Granite_3 grubu referans grup olarak alınmış, tahmini lojistik sınıflandırma modelleri elde edilerek LRA sınıflandırma sonuçları verilmiştir.

Yanıt değişkeninin multinominal olmasından dolayı IIA varsayımının geçerliliği Hausman-McFadden testi ile araştırılmıştır. 3 gruplu yanıt değişkeni için Hausman-McFadden test istatistik değeri chi2(3)=0,33 (p>chi2=0,954) olduğundan

bağımsızlık varsayımı kabul edilmiştir. Bu sonuç bu veri seti için MNL modelin uygun olduğunu göstermektedir

LRA uygulaması sonucu elde edilen parametre tahminleri ve sınıflandırma tablosu Tablo 5.59 ve Tablo 5.60'da verilmiştir.

LRA ile elde edilen katsayı tahminleri yardımıyla her bir granit türü için tahmin edilen grup üyeliklerinin olasılıkları EK-5'te verilmiştir.

Tablo 5.59. N=180 Birimli Granit Verisinin Lojistik Regresyon Parametre Tahminleri

Sınıf[a]		β	Std. Error	Wald	df	p	Exp(β)
Granite_1	Intercept	-3,372	,827	16,630	1	,000	
	x1	,151	,166	,827	1	,363	1,163
	x2	2,531	,508	24,778	1	,000	12,568
Granite_2	Intercept	-17,367	2,809	38,226	1	,000	
	x1	,826	,267	9,567	1	,002	2,285
	x2	7,309	1,069	46,713	1	,000	1493,418

c. Granite_3 referans grubudur.

Tablo 5.60. N=180 Birimli Granit Verisinin LRA Sınıflandırma Tablosu

Gerçek Grup \ Tahmin Grubu		Granite_1	Granite_2	Granite_3	Toplam
Sayım	Granite_1	48	5	7	60
	Granite_2	3	57	0	60
	Granite_3	12	2	46	60
%	Granite_1	80,00	8,33	11,67	100,0
	Granite_2	5,00	95,00	0	100,0
	Granite_3	20,00	3,33	76,67	100,0

Doğru sınıflandırma oranı %83,9'dur.

Tablo 5.60'dan görüldüğü gibi 180 birimlik granit simülasyon verisinin lojistik regresyon analizine göre doğru sınıflandırma oranı %83,9'dur.

PRA uygulaması sonucu elde edilen parametre tahminleri ve sınıflandırma tablosu Tablo 5.61 ve Tablo 5.62'de verilmiştir.

Tablo 5.61. N=180 Birimli Granit Verisinin Probit Regresyon Parametre Tahminleri

Sınıf[a]		β	Std. Error	z	p
Granite_1	Intercept	-2,619	,616	-4,25	,105
	x1	,208	,128	1,62	,000
	x2	1,663	,296	5,61	,000
Granite_2	Intercept	-12,014	1,655	-7,26	,000
	x1	,739	,187	3,96	,000
	x2	4,757	,593	8,02	,000

c. Granite_3 referans grubudur.

PRA ile elde edilen katsayı tahminleri yardımıyla her bir granit türü için tahmin edilen grup üyeliklerinin olasılıkları EK-5'te verilmiştir. Tablo 5.62'den görüldüğü gibi 180 birimlik granit simülasyon verisinin probit regresyon analizine göre doğru sınıflandırma oranı %82,2'dir.

Tablo 5.62. N=180 Birimli Granit Verisinin PRA Sınıflandırma Tablosu

	Tahmin Grubu / Gerçek Grup	Granite_1	Granite_2	Granite_3	Toplam
Sayım	Granite_1	45	5	10	60
	Granite_2	3	57	0	60
	Granite_3	13	1	46	60
%	Granite_1	75,00	8,33	16,67	100,0
	Granite_2	5,00	95,00	0	100,0
	Granite_3	21,67	1,67	76,67	100,0

Doğru sınıflandırma oranı %82,2'dir.

5.2.5. Analiz Yöntemlerin Karşılaştırılması

Simülasyon veri seti üzerine yapılan uygulamalar sonucunda LDA, QDA, LRA ve PRA yöntemlerini karşılaştırabilmek için aşağıdaki genel tablo oluşturulmuştur.

Tablo 5.63. Simülasyon Veri Uygulamasında Yöntemlerin Karşılaştırılması

	Yöntem	LDA	QDA	LRA	PRA
N=30	Doğru Sınıflandırma Oranı (%)	73,3	70	70	70
	Hatalı Sınıflandırma Oranı (%)	26,7	30	30	30
N=90	Doğru Sınıflandırma Oranı (%)	82,2	83,3	83,3	82,2
	Hatalı Sınıflandırma Oranı (%)	17,8	16,7	16,7	17,8
N=180	Doğru Sınıflandırma Oranı (%)	83,9	82,8	83,9	82,2
	Hatalı Sınıflandırma Oranı (%)	16,1	17,2	16,1	17,8

Tablo 5.63'teki yöntemlerin genel sınıflandırma doğrulukları incelendiğinde, örneklem sayısı arttıkça kullanılan 4 farklı analiz yönteminin de sınıflandırma yeteneklerinin arttığı görülmektedir. Klasik diskriminant analizi varsayımlarını sağlayan 30 birimlik örneklemde en yüksek sınıflandırma oranı LDA ile %73,3 elde edilmiştir. Eşit hacimli sınıflara sahip olan ve ortak varyans-kovaryans varsayımını sağlamayan 90 ve 180 birimlik örneklemlerin analiz sonuçlarında ise %82,2 ile %83,9 aralığında birbirine çok yakın sınıflandırma doğrulukları elde edilmiştir.

6. SONUÇLAR VE ÖNERİLER

Verilerde yanıt değişkenin kategorik olması, amaç regresyon analizi olduğunda lineer regresyon modelini kurmayı yetersiz kılmaktadır. İkili (binary) bir yanıt değişkeni tahmin etmek için kukla değişken yardımıyla lineer regresyon uygulanabiliyor olmasına rağmen, kukla değişkeni yaklaşımı ikiden fazla kategoriye sahip yanıt değişkeni karşılamak için kolaylıkla genişletilemez. Bu nedenle kategorik yanıt değerler için sınıflandırma yöntemi kullanılır.

Bir gözlem için kategorik bir yanıt tahmin etmek, bu gözlemi bir kategoriye ya da gruba atamayı içeren "sınıflandırma" olarak adlandırılır. Sınıflandırma için kullanılan yöntemler öncelikle nitel yanıt değişkeninin her bir kategorisinin olasılığını tahmin eder ve yeni bir gözlem değeri için gelecekte kullanılabilir fonksiyonlar verir. Bu anlamda bu yöntemler regresyon yöntemi gibi de davranırlar. Bu çalışmada gözlemlerin gruplara ayrılmasında kullanılan bu sınıflandırma yöntemlerinden; diskriminant analizi ve bu analize alternatif olarak önerilen lojistik regresyon analizi ile probit regresyon analizi ele alınmıştır.

Çalışmanın ikinci bölümünde diskriminant analizi ayrıntılı bir şekilde incelenmiştir. Grup ayrıştırmasının tanımı verildikten sonra ilk olarak iki kategoriye sahip diskriminant fonksiyonları verilmiş ve iki gruplu diskriminant analizi ile çoklu regresyon analizi arasındaki ilişki açıklanmıştır. Sonrasında ikiden fazla kategoriye sahip diskriminant fonksiyonları açıklanarak bu fonksiyonlar için Fisher'in kanonik korelasyonunun nasıl hesaplandığı gösterilmiştir. Ayrıca bu iki tip fonksiyon türünde bağımsız değişkenlerin anlamlılık testleri incelenerek, bu değişkenlerin grup ayrımına katkısının nasıl belirleneceğine yer verilmiştir. Sınıflandırma prosedürü incelenmiş, lineer ve karesel sınıflandırma fonksiyonları ikiden fazla gruba sınıflandırma başlığı altında ele alınmıştır. Diskriminant analizinde hatalı sınıflandırma oranlarının tahmini açıklanarak, hata oranlarındaki yanlılığı azaltmak için önerilen tahmin yöntemlerine yer verilmiştir.

Çalışmanın üçüncü ve dördüncü bölümünde, iki ve ikiden fazla kategoriye sahip yanıt değişkenli sırasıyla lojistik ve probit modeller incelenerek her iki modelde parametrelerin tahmin edilmesi için önerilen EÇO yöntemi açıklanmıştır. Lojistik regresyon parametrelerinin anlamlılık testleri olan olabilirlik oran testi, Wald testi ve skor testi incelenerek bağımsız değişkelerin model için anlamlı olup olmadığının nasıl belirleneceği teorik açıdan gösterilmiştir. Ayrıca probit regresyon ile lojistik regresyon arasındaki ilişki açıklanmıştır.

İncelenen analiz yöntemlerinin pratikteki uygulamalarını göstermek amacı ile iki uygulamaya yer verilmiştir. Bunlardan ilki gerçek ölçümlerden oluşan bir veri seti üzerine diğeri ise simülasyon sonucu elde edilmiş veriler üzerine uygulanmıştır.

Bir hastalığın doğru teşhisi ve tedavisi aşamasındaki karmaşıklıklardan kaçınmak için doğru sınıflandırılması hayati önem taşımaktadır. Bu durum genellikle, çok fazla sinyal işleme yardımı olmaksızın, bir hekim tarafından deneyime dayanarak gerçekleştirilir. Özellikle doğru kararların alınmasında çok yönlü ve mantıklı bir tıbbi tanı / sınıflandırma sisteminin yararlı olabileceği düşünülür. Bu düşünceyle gerçek veri seti uygulamasında her bir kategoride farklı sayıda hasta bulunan, bel fıtığı (disk hernia/60) ve bel kayması (spondylolisthesis/149) şeklinde iki farklı omurga rahatsızlığı olan 209 hasta ile 100 normal hastadan oluşan, 3 kategoriye ayrılmış toplam 309 hastanın bilgisine sahip vertebral kolon verisi analiz edilmiştir. Analiz sonucunda hastaya bel fıtığı, bel kayması ve normal tanısı koyabilen, her bir sınıflandırma yöntemi için en az açıklayıcı değişken içeren optimum sınıflandırma modelleri oluşturulup sınıflandırma doğrulukları karşılaştırılmıştır.

Veri tabanındaki her hasta, leğen kemiğine (pelvis) ve bel kemiğine (lomber) ait omurganın şekli ve hizalamasından kaynaklanan 6 biyomekanik özellik ile karakterize edilmiştir. Verinin açıklayıcı değişkenlerini oluşturan bu özellikler; pelvis/leğen kemiği insidansı (PI), pelvis eğimi (PT), lordoz/bel kemiği

açısı (LuL), sakral/kuyruk sokumu eğimi (SS), pelvis yarıçapı (PR) ve spondilolistezi /kayma derecesi (DS) dir.

Gerçek veri analizinde dört yöntemin sonuçları karşılaştırıldığında pelvis insidansı (PI) ölçümünün hastaları sınıflandırmada anlamlı bir etkiye sahip olmadığı görülmektedir. Diskriminant analizinde elde edilen LDA ve QDA yöntemleri için bel kemiği açısı (LuL) ölçümü en az etkiye sahip ölçüm olarak modele dahil edilerek 5 açıklayıcı değişken içeren optimum sınıflandırma modeli oluşturmalarına rağmen, LRA ve PRA yöntemlerinde LuL değişkeni de model için anlamsız bulunarak bu yöntemler 4 açıklayıcı değişken içeren optimum sınıflandırma modeli oluşturmuşlardır.

Her bir sınıfta farklı örneklem bulunan gerçek veri seti ile gerçekleştirilen analizlerde, doğru sınıflandırma oranlarına göre en iyi model logit ve probit model, sonra karesel diskriminant modeli ve en son lineer diskriminant modeli bulunmuştur. Diskriminant analizinin iki yöntemi olan QDA ile LDA karşılaştırıldığında, veride ortak kovaryansa sahip olma varsayımının bozulmasından dolayı QDA'nın tercih edilmesi gereklidir. Ayrıca daha yüksek doğru sınıflandırma oranı da verdiğine dikkat edilmelidir. LRA ve PRA en az açıklayıcı değişkeni kullanarak en yüksek doğru sınıflandırma oranını verdikleri için bilinen varsayımlar sağlanmağında LDA ve QDA yöntemlerine göre tercih edilebilecekleri söylenebilir.

Sınıflandırma doğruluklarının farklı koşullardaki etkisi bir simülasyon çalışmasıyla da ortaya konmuştur. Simülasyon çalışmasında küçük, orta ve büyük örneklemli, her bir sınıfta eşit örneklem bulunan 3 farklı granit taşı sınıfına sahip jeolojik granit verileri kullanılmıştır

Simülasyon sonuçlarına göre klasik diskriminant analizi varsayımlarını sağlayan küçük örnekleme sahip veri yapısının QDA, LRA ve PRA yöntemlerine göre LDA ile daha doğru sınıflandırma oranı verdiği ortaya konmuştur. Normal dağılan ancak ortak kovaryans varsayımını sağlamayan orta ve büyük örnekleme

sahip veri yapılarının ise, QDA yöntemi ile LDA yöntemine göre daha yüksek doğrulukta sınıflandırma sonuçlarına sahip oldukları tespit edilmiştir. Buna karşılık gerçek veri analizi ile kıyaslandığında, sınıflarında eşit örneklem bulunan yanıt değişkenine sahip bu simülasyon verilerindeki LDA, QDA LRA ve PRA doğru sınıflandırma oranları arasında büyük bir farklılık olmadığı da açığa çıkmıştır.

Elde edilen gerçek ve simülasyon veri sonuçlarına göre klasik varsayımları karşılayan veri yapıları için LDA en iyi sınıflandırma yöntemi olma özelliğini korumuştur.

Klasik varsayımları sağlamayan veri yapıları için farklı bulgular elde edilmiştir. Elde edilen bu bulgulardan yola çıkarak iki farklı sonuç ortaya çıkmıştır.

Yanıt değişkeninin farklı hacimli sınıflara sahip olması durumunda, LRA ve PRA yöntemlerinin daha yüksek doğru sınıflandırma oranları vermesi açısından QDA yöntemine göre iyi birer alternatif sınıflandırma yöntemleri oldukları söylenebilir.

Yanıt değişkeninin eşit hacimli sınıflara sahip olması durumunda ise QDA, LRA ve PRA tahmin olasılıklarının birbirine çok yakın olmasından dolayı çok küçük farklılıklara sahip sınıflandırma sonuçları ortaya çıkmaktadır. Bu nedenle bu yapıdaki verilerde sınıflandırma yöntemi olarak QDA, LRA ve PRA yöntemlerinden herhangi biri tercih edilebilir.

Bu çalışmada diskriminant analizi ve bu analize alternatif yöntemler ele alındığı için normal yapıdaki verilerle çalışılmıştır. Normal dağılmayan veri yapılarında, araştırılan teorik bilgiye göre tercih edilen sınıflandırma yöntemleri farklılık göstermektedir. PRA yöntemindeki yanıt değişken için normal dağılım varsayımı LRA yöntemine göre daha sınırlayıcı olduğundan, veri yapısının normal dağılmadığı durumlarda bu iki yöntem arasında herhangi bir dağılım varsayımı gerektirmeyen LRA yöntemi tercih edilmelidir.

KAYNAKLAR

Afifi, A., Clark,V. A., May, S., 2004. Computer-Aided Multivariate Analysis. 3rd edition, Chapman & Hall/CRC, USA, 476p.

Agresti, A., 1990. Categorical Data Analysis. John Wiley & Sons, New York-USA, 558p..

Albayrak, A. S., 2006. Uygulamalı Çok Değişkenli İstatistik Teknikleri. Asil Yayın Dağıtım Ltd.Şti.

Aldrich, J. H., Nelson, F.D., 1984. Linear Probability, Logit and Probit Models, Sage Publications, London, 94p.

_____, 1986. Linear probability logit and probit models: Sage publications, London, 95 p.

Altınışık, Z., 2007. Multinominal Probit Modeller. H.Ü. Yüksek Lisans Tezi, Ankara, 84s.

Arı, A., Önder, H., 2012. Farklı Veri Yapılarında Kullanılabilecek Regresyon Yöntemleri. Ondokuz Mayıs Üniversitesi, Ziraat Fakültesi, Zootekni Bölümü, Samsun. Anadolu Tarım Bilim. Derg., 2013, 28(3):168-174

Bailey, H., Love, R., J., M., 2008. Bailey & Loves's Short Practice of Surgery. '25th edition.

Başarır, G., 1990. Çok Değişkenli Verilerde Ayrımsama Sorunu ve Lojistik Regresyon Analizi. Hacettepe Üniversitesi, Fen Bilimleri Enstitüsü, Doktora Tezi, Ankara. 145s.

Bek, Y., 2009. R Programında Doz-Yanıt Uygulamaları Çalıştayı Dinleyici Notları. 27-29 Mayıs 2009, Samsun.

Berkson, J., 1944. Application of the Logistic Function to Bio-Assay.Journal of the American Statistical Association, 39(227): 357-365.

Christensen, R., 1997. Log-Linear Models and Logistic Regression. 2nd edition Springer –Verlag New Work, 498p.

Cornfield, J., 1962. Epidemiological Aspects of Coronary Artery Disease. Annals of The New York Academy of Sciences, 97:959.

Cox, D. R., 1970. The Analysis of Binary Data. Methuen, London.

Çamdeviren, H., 2000. Lojistik Regresyon ve Diskriminant Analizi, Doktora Tezi, Ankara Üniversitesi, Ankara.

Elesan, S., 2010. Multinominal Lojistik Regresyon Analizi ve Uygulaması. Yüzüncü Yıl Üniversitesi, Sağlık Bilimleri Enstitüsü, Biyoistatistik Anabilim Dalı, Yüksek Lisans Tezi, Van, 67s.

Ereeş, S., 2013. Some Model Misspecifications in Logistic Regression Model. Dokuz Eylül University, Graduate School of Natural and Applied Sciences, Phd. Thesis, İzmir, 110s.

Finney, D. J., 1971. Probit Analysis: Third Edition, Cambridge, University Press, Cambridge, 318 p.

Fisher, R. A., 1936. The Use Of Multiple Measurements İn Taxonomic Problems. Annals of Eugenics, 7: 179-188.

Frank, A., Asuncion, A., 2010. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Freund, R. J.,Wilson, J.,W., Sa, P., 2006. Regression Analysis. (Statistical Modelling of a Response Variable) 2nd edition, Academic Press.

George, D., Mallery, M., 2010. SPSS for Windows Step by Step : A Simple Guide and Reference, 17.0 update. 10a edition, Pearson, Boston.

Gundabathina, J., L., Kodali, S., R., Reddy, S., K., 2012. Classification of Vertebral Column using Naive Bayes Technique. International Journal of Computer Applications, 58(7), 38-42.

Gujarati, D. N., 2003. Basıc Econometrics Fourth edition. McGraw-Hill Companies, Inc., New York, 1027page.

Halperin, M., Blackwelder, W. C., Verter, J. I., 1971. Estimation Of The Multivariate Logistic Risk Function: A Comparison Of The Discriminant Function And Maximum Likelihood Approaches. Journal of Clinical Epidemiology, 24:125-158

Hastie, T., Tibshirani, R., Friedman, J., 2008. The Elements of Statistical Learning. 2nd edition, Springer, Standford, California 739p.

Hilbe, J. M., 2009. Logistic Regression Models (Texts in Statistical Science). Chapman & Hall/CRC, USA,668p.

Hosmer, D. W. and Lemeshow, S., 2000. Applied Logistic Regression, Second Edition. John Wiley and Sons Inc. New York, 375p.

İçyüz, E., 2010. Doğrusal Olasılık, Logit ve Probit Modellerinin Hayat Sigortacılığına uygulanması. Marmara Üniversitesi, Sosyal Bilimler Enstitüsü, Ekonometri Anabilim Dalı, Yüksek Lisans Tezi, İstanbul, 146s.

James, G., Witten, D., Hastie, T. and Tibshirani, R., 2015. An Introduction to Statistical Learning with Application in R. Springer Science+Business Media, New York, 425p.

Johnson, R. A., Wichern, D. W., 2002. Applied Multivariate Statistical Analysis, Prentice Hall, USA.

Kalaycı, Ş., 2006. SPSS Uygulamalı Çok Değişkenli İstatistik Teknikleri. 5. Baskı, Asil Yayın Dağıtım, Ankara,424s.

Kaşko, Y., 2007. Çoklu Bağlantı Durumunda İkili (Binary) Lojistik Regresyon Modelinde Gerçekleşen 1. Tip Hata ve Testin Gücü. Ankara Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, Ankara, 55s.

Klecka, W. R., 1980. Discriminant Analysis, Sage Pub., Beverly Hills.

Kleinbaum, D. G., Klein, M., 2002. Logistic Regression. 2nd edition, Department of Epidemiology Emory University Atlanta, GA 30333 USA 267-299.

Lachenbruch, P. A., 1975. Discriminant Analysis. Hafner Press, New York.

Lavine, B., Carlson, D.,1987. European Bee or Africanized Bee?: Species Identification through Chemical Analysis. Analytical Chemistry,59(6):468-470

Lawles, J.F., 1987. Negative binomial and mixed poisson regression. The Canadian Journal of Stat., 15(3), 209-225.

Mahalanobis, P. C., 1936. On the Use Generalized Distance in Statistics. Proceedings of the National Institute of Sciences of India, 12, 49-55.

Montgomery, D. C., Peck, E. A., Vining, G.G., 2001. Introduction to Linear Regression Analysis. 3rd edition, John Wiley & Sons, 620p.

Neto, A., R., R., Sousa, R., Barreto, G., A., Cardoso, J., S., 2011. Diagnostic of Pathology on the Vertebral Column with Embedded Reject Option. 5th Iberian Conference on Pattern Recognition and Image Analysis, Spain, 588-595.

Özarıcı, Ö., 1996. Farklı Not Sistemlerinde Öğrencinin Başarılı Olma Olasılığının Probit Regresyon Analiziyle Değerlendirilmesi. Osmangazi Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, Eskişehir, 81s.

Rencher, A. L., 2002. Methods of Multivariate Analysis, John Wiley and Sons, Inc., USA, 708p.

Sakal, M., 2014. Lojistik ve Probit Modelleri için Kullanılan Uyum Kriterlerinin İncelenmesi ve Simülasyonla Karşılaştırılması. Muğla Sıtkı Koçma Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, Muğla, 71s.

Sharma., S. 1996. Applied Multivariate Techniques. John wiley&sons, Inc. Canada USA, 238 page.

Silahlı, N., 2013. Applications of Logistic Regression With Missing Data. Dokuz Eylül University, Graduate School of Natural and Applied Sciences, Msc's Thesis, İzmir,85s.

Tabachnick, B., G., Fidell, L. S., 2015. Using Multivariate Statistics. 6 Edition, Pearson Education, USA, 1018p.

Tatlıdil, H. 2002. Uygulamalı Çok Değişkenli İstatistiksel Analiz. Akademi Matbaası, Ankara.

Trauth, M., H., 2015. MATLAB Recipes for Earth Sciences. 4th Edition, Springer, Germany, 436p.

Türe, M., Kurt, İ., Yavuz, E., Kürüm, T., 2005. Hipertansiyonun tahmini için çoklu tahmin modellerinin karşılaştırılması (Sinir ağları, lojistik regresyon ve esnek ayırma analizleri). Trakya Üniversitesi Tıp Fakültesi Biyoistatistik ve Kardiyoloji Anabilim Dalları, Edirne. Anadolu Kardiyoloji Dergisi 5: 24-28.

Ünal, İ., 2004. Prognozda ve Tanıda Lojistik Regresyon ve Neural Network Yaklaşımı: Epidemiyolojik Veri Setlerinde Karşılaştırmalı Uygulamalar. Çukurova Üniversitesi, Sağlık Bilimleri Enstitüsü, Biyoistatistik Anabilim Dalı, Yüksek Lisans Tezi, Adana, 96s.

Wang ,P., Putterman, M. L., 1998. Mixed logistic regression models. J. of Agriculture, Biological and Environmental Stat., 3(2), 175-200.

Vupa, Ö., 2004. Model Building of Logistic Regression Models. Dokuz Eylül University, Graduate School of Natural and Applied Sciences, Msc's Thesis, İzmir, 94s.

ÖZGEÇMİŞ

1990 yılında Adana'da doğdu. İlk, orta ve lise öğrenimini Adana'da tamamladıktan sonra 2008 yılında Çukurova Üniversitesi Fen Edebiyat Fakültesi İstatistik Bölümü'nde lisans öğrenimine başladı. 2012 yılında bu bölümden mezun olup 2014 yılında Çukurova Üniversitesi Fen Edebiyat Fakültesi İstatistik Bölümü'nde yüksek lisansa başladı. Halen Çukurova Üniversitesi Fen Edebiyat Fakültesi İstatistik Bölümü'nde yüksek lisans öğrencisi olarak eğitimine devam etmektedir.

EKLER

EK-1: Vertebral Column Gerçek Veri Seti Çalışmasında Kullanılan Örneklem

No	PI	PT	LuL	SS	PR	DS	class
1	63,03	22,55	39,61	40,48	98,67	-0,25	Disk Hernia
2	39,06	10,06	25,02	29	114,41	4,56	Disk Hernia
3	68,83	22,22	50,09	46,61	105,99	-3,53	Disk Hernia
4	69,3	24,65	44,31	44,64	101,87	11,21	Disk Hernia
5	49,71	9,65	28,32	40,06	108,17	7,92	Disk Hernia
6	40,25	13,92	25,12	26,33	130,33	2,23	Disk Hernia
7	53,43	15,86	37,17	37,57	120,57	5,99	Disk Hernia
8	45,37	10,76	29,04	34,61	117,27	-10,68	Disk Hernia
9	43,79	13,53	42,69	30,26	125	13,29	Disk Hernia
10	36,69	5,01	41,95	31,68	84,24	0,66	Disk Hernia
11	49,71	13,04	31,33	36,67	108,65	-7,83	Disk Hernia
12	31,23	17,72	15,5	13,52	120,06	0,5	Disk Hernia
13	48,92	19,96	40,26	28,95	119,32	8,03	Disk Hernia
14	53,57	20,46	33,1	33,11	110,97	7,04	Disk Hernia
15	57,3	24,19	47	33,11	116,81	5,77	Disk Hernia
16	44,32	12,54	36,1	31,78	124,12	5,42	Disk Hernia
17	63,83	20,36	54,55	43,47	112,31	-0,62	Disk Hernia
18	31,28	3,14	32,56	28,13	129,01	3,62	Disk Hernia
19	38,7	13,44	31	25,25	123,16	1,43	Disk Hernia
20	41,73	12,25	30,12	29,48	116,59	-1,24	Disk Hernia
21	43,92	14,18	37,83	29,74	134,46	6,45	Disk Hernia
22	54,92	21,06	42,2	33,86	125,21	2,43	Disk Hernia
23	63,07	24,41	54	38,66	106,42	15,78	Disk Hernia
24	45,54	13,07	30,3	32,47	117,98	-4,99	Disk Hernia
25	36,13	22,76	29	13,37	115,58	-3,24	Disk Hernia
26	54,12	26,65	35,33	27,47	121,45	1,57	Disk Hernia
27	26,15	10,76	14	15,39	125,2	-10,09	Disk Hernia
28	43,58	16,51	47	27,07	109,27	8,99	Disk Hernia
29	44,55	21,93	26,79	22,62	111,07	2,65	Disk Hernia
30	66,88	24,89	49,28	41,99	113,48	-2,01	Disk Hernia

31	50,82	15,4	42,53	35,42	112,19	10,87	Disk Hernia
32	46,39	11,08	32,14	35,31	98,77	6,39	Disk Hernia
33	44,94	17,44	27,78	27,49	117,98	5,57	Disk Hernia
34	38,66	12,99	40	25,68	124,91	2,7	Disk Hernia
35	59,6	32	46,56	27,6	119,33	1,47	Disk Hernia
36	31,48	7,83	24,28	23,66	113,83	4,39	Disk Hernia
37	32,09	6,99	36	25,1	132,26	6,41	Disk Hernia
38	35,7	19,44	20,7	16,26	137,54	-0,26	Disk Hernia
39	55,84	28,85	47,69	27	123,31	2,81	Disk Hernia
40	52,42	19,01	35,87	33,41	116,56	1,69	Disk Hernia
41	35,49	11,7	15,59	23,79	106,94	-3,46	Disk Hernia
42	46,44	8,4	29,04	38,05	115,48	2,05	Disk Hernia
43	53,85	19,23	32,78	34,62	121,67	5,33	Disk Hernia
44	66,29	26,33	47,5	39,96	121,22	-0,8	Disk Hernia
45	56,03	16,3	62,28	39,73	114,02	-2,33	Disk Hernia
46	50,91	23,02	47	27,9	117,42	-2,53	Disk Hernia
47	48,33	22,23	36,18	26,1	117,38	6,48	Disk Hernia
48	41,35	16,58	30,71	24,78	113,27	-4,5	Disk Hernia
49	40,56	17,98	34	22,58	121,05	-1,54	Disk Hernia
50	41,77	17,9	20,03	23,87	118,36	2,06	Disk Hernia
51	55,29	20,44	34	34,85	115,88	3,56	Disk Hernia
52	74,43	41,56	27,7	32,88	107,95	5	Disk Hernia
53	50,21	29,76	36,1	20,45	128,29	5,74	Disk Hernia
54	30,15	11,92	34	18,23	112,68	11,46	Disk Hernia
55	41,17	17,32	33,47	23,85	116,38	-9,57	Disk Hernia
56	47,66	13,28	36,68	34,38	98,25	6,27	Disk Hernia
57	43,35	7,47	28,07	35,88	112,78	5,75	Disk Hernia
58	46,86	15,35	38	31,5	116,25	1,66	Disk Hernia
59	43,2	19,66	35	23,54	124,85	-2,92	Disk Hernia
60	48,11	14,93	35,56	33,18	124,06	7,95	Disk Hernia
61	74,38	32,05	78,77	42,32	143,56	56,13	Spondylolisthesis
62	89,68	32,7	83,13	56,98	129,96	92,03	Spondylolisthesis
63	44,53	9,43	52	35,1	134,71	29,11	Spondylolisthesis
64	77,69	21,38	64,43	56,31	114,82	26,93	Spondylolisthesis

65	76,15	21,94	82,96	54,21	123,93	10,43	Spondylolisthesis
66	83,93	41,29	62	42,65	115,01	26,59	Spondylolisthesis
67	78,49	22,18	60	56,31	118,53	27,38	Spondylolisthesis
68	75,65	19,34	64,15	56,31	95,9	69,55	Spondylolisthesis
69	72,08	18,95	51	53,13	114,21	1,01	Spondylolisthesis
70	58,6	-0,26	51,5	58,86	102,04	28,06	Spondylolisthesis
71	72,56	17,39	52	55,18	119,19	32,11	Spondylolisthesis
72	86,9	32,93	47,79	53,97	135,08	101,72	Spondylolisthesis
73	84,97	33,02	60,86	51,95	125,66	74,33	Spondylolisthesis
74	55,51	20,1	44	35,42	122,65	34,55	Spondylolisthesis
75	72,22	23,08	91	49,14	137,74	56,8	Spondylolisthesis
76	70,22	39,82	68,12	30,4	148,53	145,38	Spondylolisthesis
77	86,75	36,04	69,22	50,71	139,41	110,86	Spondylolisthesis
78	58,78	7,67	53,34	51,12	98,5	51,58	Spondylolisthesis
79	67,41	17,44	60,14	49,97	111,12	33,16	Spondylolisthesis
80	47,74	12,09	39	35,66	117,51	21,68	Spondylolisthesis
81	77,11	30,47	69,48	46,64	112,15	70,76	Spondylolisthesis
82	74,01	21,12	57,38	52,88	120,21	74,56	Spondylolisthesis
83	88,62	29,09	47,56	59,53	121,76	51,81	Spondylolisthesis
84	81,1	24,79	77,89	56,31	151,84	65,21	Spondylolisthesis
85	76,33	42,4	57,2	33,93	124,27	50,13	Spondylolisthesis
86	45,44	9,91	45	35,54	163,07	20,32	Spondylolisthesis
87	59,79	17,88	59,21	41,91	119,32	22,12	Spondylolisthesis
88	44,91	10,22	44,63	34,7	130,08	37,36	Spondylolisthesis
89	56,61	16,8	42	39,81	127,29	24,02	Spondylolisthesis
90	71,19	23,9	43,7	47,29	119,86	27,28	Spondylolisthesis
91	81,66	28,75	58,23	52,91	114,77	30,61	Spondylolisthesis
92	70,95	20,16	62,86	50,79	116,18	32,52	Spondylolisthesis
93	85,35	15,84	71,67	69,51	124,42	76,02	Spondylolisthesis
94	58,1	14,84	79,65	43,26	113,59	50,24	Spondylolisthesis
95	94,17	15,38	67,71	78,79	114,89	53,26	Spondylolisthesis
96	57,52	33,65	50,91	23,88	140,98	148,75	Spondylolisthesis
97	96,66	19,46	90,21	77,2	120,67	64,08	Spondylolisthesis
98	74,72	19,76	82,74	54,96	109,36	33,31	Spondylolisthesis

99	77,66	22,43	93,89	55,22	123,06	61,21	Spondylolisthesis
100	58,52	13,92	41,47	44,6	115,51	30,39	Spondylolisthesis
101	84,59	30,36	65,48	54,22	108,01	25,12	Spondylolisthesis
102	79,94	18,77	63,31	61,16	114,79	38,54	Spondylolisthesis
103	70,4	13,47	61,2	56,93	102,34	25,54	Spondylolisthesis
104	49,78	6,47	53	43,32	110,86	25,34	Spondylolisthesis
105	77,41	29,4	63,23	48,01	118,45	93,56	Spondylolisthesis
106	65,01	27,6	50,95	37,41	116,58	7,02	Spondylolisthesis
107	65,01	9,84	57,74	55,18	94,74	49,7	Spondylolisthesis
108	78,43	33,43	76,28	45	138,55	77,16	Spondylolisthesis
109	63,17	6,33	63	56,84	110,64	42,61	Spondylolisthesis
110	68,61	15,08	63,01	53,53	123,43	39,5	Spondylolisthesis
111	63,9	13,71	62,12	50,19	114,13	41,42	Spondylolisthesis
112	85	29,61	83,35	55,39	126,91	71,32	Spondylolisthesis
113	42,02	-6,55	67,9	48,58	111,59	27,34	Spondylolisthesis
114	69,76	19,28	48,5	50,48	96,49	51,17	Spondylolisthesis
115	80,99	36,84	86,96	44,14	141,09	85,87	Spondylolisthesis
116	129,83	8,4	48,38	121,43	107,69	418,54	Spondylolisthesis
117	70,48	12,49	62,42	57,99	114,19	56,9	Spondylolisthesis
118	86,04	38,75	47,87	47,29	122,09	61,99	Spondylolisthesis
119	65,54	24,16	45,78	41,38	136,44	16,38	Spondylolisthesis
120	60,75	15,75	43,2	45	113,05	31,69	Spondylolisthesis
121	54,74	12,1	41	42,65	117,64	40,38	Spondylolisthesis
122	83,88	23,08	87,14	60,8	124,65	80,56	Spondylolisthesis
123	80,07	48,07	52,4	32,01	110,71	67,73	Spondylolisthesis
124	65,67	10,54	56,49	55,12	109,16	53,93	Spondylolisthesis
125	74,72	14,32	32,5	60,4	107,18	37,02	Spondylolisthesis
126	48,06	5,69	57,06	42,37	95,44	32,84	Spondylolisthesis
127	70,68	21,7	59,18	48,97	103,01	27,81	Spondylolisthesis
128	80,43	17	66,54	63,43	116,44	57,78	Spondylolisthesis
129	90,51	28,27	69,81	62,24	100,89	58,82	Spondylolisthesis
130	77,24	16,74	49,78	60,5	110,69	39,79	Spondylolisthesis
131	50,07	9,12	32,17	40,95	99,71	26,77	Spondylolisthesis
132	69,78	13,78	58	56	118,93	17,91	Spondylolisthesis

133	69,63	21,12	52,77	48,5	116,8	54,82	Spondylolisthesis
134	81,75	20,12	70,56	61,63	119,43	55,51	Spondylolisthesis
135	52,2	17,21	78,09	34,99	136,97	54,94	Spondylolisthesis
136	77,12	30,35	77,48	46,77	110,61	82,09	Spondylolisthesis
137	88,02	39,84	81,77	48,18	116,6	56,77	Spondylolisthesis
138	83,4	34,31	78,42	49,09	110,47	49,67	Spondylolisthesis
139	72,05	24,7	79,87	47,35	107,17	56,43	Spondylolisthesis
140	85,1	21,07	91,73	64,03	109,06	38,03	Spondylolisthesis
141	69,56	15,4	74,44	54,16	105,07	29,7	Spondylolisthesis
142	89,5	48,9	72	40,6	134,63	118,35	Spondylolisthesis
143	85,29	18,28	100,74	67,01	110,66	58,88	Spondylolisthesis
144	60,63	20,6	64,54	40,03	117,23	104,86	Spondylolisthesis
145	60,04	14,31	58,04	45,73	105,13	30,41	Spondylolisthesis
146	85,64	42,69	78,75	42,95	105,14	42,89	Spondylolisthesis
147	85,58	30,46	78,23	55,12	114,87	68,38	Spondylolisthesis
148	55,08	-3,76	56	58,84	109,92	31,77	Spondylolisthesis
149	65,76	9,83	50,82	55,92	104,39	39,31	Spondylolisthesis
150	79,25	23,94	40,8	55,3	98,62	36,71	Spondylolisthesis
151	81,11	20,69	60,69	60,42	94,02	40,51	Spondylolisthesis
152	48,03	3,97	58,34	44,06	125,35	35	Spondylolisthesis
153	63,4	14,12	48,14	49,29	111,92	31,78	Spondylolisthesis
154	57,29	15,15	64	42,14	116,74	30,34	Spondylolisthesis
155	41,19	5,79	42,87	35,39	103,35	27,66	Spondylolisthesis
156	66,8	14,55	72,08	52,25	82,46	41,69	Spondylolisthesis
157	79,48	26,73	70,65	52,74	118,59	61,7	Spondylolisthesis
158	44,22	1,51	46,11	42,71	108,63	42,81	Spondylolisthesis
159	57,04	0,35	49,2	56,69	103,05	52,17	Spondylolisthesis
160	64,27	12,51	68,7	51,77	95,25	39,41	Spondylolisthesis
161	92,03	35,39	77,42	56,63	115,72	58,06	Spondylolisthesis
162	67,26	7,19	51,7	60,07	97,8	42,14	Spondylolisthesis
163	118,14	38,45	50,84	79,7	81,02	74,04	Spondylolisthesis
164	115,92	37,52	76,8	78,41	104,7	81,2	Spondylolisthesis
165	53,94	9,31	43,1	44,64	124,4	25,08	Spondylolisthesis
166	83,7	20,27	77,11	63,43	125,48	69,28	Spondylolisthesis

167	56,99	6,87	57,01	50,12	109,98	36,81	Spondylolisthesis
168	72,34	16,42	59,87	55,92	70,08	12,07	Spondylolisthesis
169	95,38	24,82	95,16	70,56	89,31	57,66	Spondylolisthesis
170	44,25	1,1	38	43,15	98,27	23,91	Spondylolisthesis
171	64,81	15,17	58,84	49,64	111,68	21,41	Spondylolisthesis
172	78,4	14,04	79,69	64,36	104,73	12,39	Spondylolisthesis
173	56,67	13,46	43,77	43,21	93,69	21,11	Spondylolisthesis
174	50,83	9,06	56,3	41,76	79	23,04	Spondylolisthesis
175	61,41	25,38	39,1	36,03	103,4	21,84	Spondylolisthesis
176	56,56	8,96	52,58	47,6	98,78	50,7	Spondylolisthesis
177	67,03	13,28	66,15	53,75	100,72	33,99	Spondylolisthesis
178	80,82	19,24	61,64	61,58	89,47	44,17	Spondylolisthesis
179	80,65	26,34	60,9	54,31	120,1	52,47	Spondylolisthesis
180	68,72	49,43	68,06	19,29	125,02	54,69	Spondylolisthesis
181	37,9	4,48	24,71	33,42	157,85	33,61	Spondylolisthesis
182	64,62	15,23	67,63	49,4	90,3	31,33	Spondylolisthesis
183	75,44	31,54	89,6	43,9	106,83	54,97	Spondylolisthesis
184	71	37,52	84,54	33,49	125,16	67,77	Spondylolisthesis
185	81,06	20,8	91,78	60,26	125,43	38,18	Spondylolisthesis
186	91,47	24,51	84,62	66,96	117,31	52,62	Spondylolisthesis
187	81,08	21,26	78,77	59,83	90,07	49,16	Spondylolisthesis
188	60,42	5,27	59,81	55,15	109,03	30,27	Spondylolisthesis
189	85,68	38,65	82,68	47,03	120,84	61,96	Spondylolisthesis
190	82,41	29,28	77,05	53,13	117,04	62,77	Spondylolisthesis
191	43,72	9,81	52	33,91	88,43	40,88	Spondylolisthesis
192	86,47	40,3	61,14	46,17	97,4	55,75	Spondylolisthesis
193	74,47	33,28	66,94	41,19	146,47	124,98	Spondylolisthesis
194	70,25	10,34	76,37	59,91	119,24	32,67	Spondylolisthesis
195	72,64	18,93	68	53,71	116,96	25,38	Spondylolisthesis
196	71,24	5,27	86	65,97	110,7	38,26	Spondylolisthesis
197	63,77	12,76	65,36	51,01	89,82	56	Spondylolisthesis
198	58,83	37,58	125,74	21,25	135,63	117,31	Spondylolisthesis
199	74,85	13,91	62,69	60,95	115,21	33,17	Spondylolisthesis
200	75,3	16,67	61,3	58,63	118,88	31,58	Spondylolisthesis

201	63,36	20,02	67,5	43,34	131	37,56	Spondylolisthesis
202	67,51	33,28	96,28	34,24	145,6	88,3	Spondylolisthesis
203	76,31	41,93	93,28	34,38	132,27	101,22	Spondylolisthesis
204	73,64	9,71	63	63,92	98,73	26,98	Spondylolisthesis
205	56,54	14,38	44,99	42,16	101,72	25,77	Spondylolisthesis
206	80,11	33,94	85,1	46,17	125,59	100,29	Spondylolisthesis
207	95,48	46,55	59	48,93	96,68	77,28	Spondylolisthesis
208	74,09	18,82	76,03	55,27	128,41	73,39	Spondylolisthesis
209	87,68	20,37	93,82	67,31	120,94	76,73	Spondylolisthesis
210	48,26	16,42	36,33	31,84	94,88	28,34	Spondylolisthesis
211	38,51	16,96	35,11	21,54	127,63	7,99	Normal
212	54,92	18,97	51,6	35,95	125,85	2	Normal
213	44,36	8,95	46,9	35,42	129,22	4,99	Normal
214	48,32	17,45	48	30,87	128,98	-0,91	Normal
215	45,7	10,66	42,58	35,04	130,18	-3,39	Normal
216	30,74	13,35	35,9	17,39	142,41	-2,01	Normal
217	50,91	6,68	30,9	44,24	118,15	-1,06	Normal
218	38,13	6,56	50,45	31,57	132,11	6,34	Normal
219	51,62	15,97	35	35,66	129,39	1,01	Normal
220	64,31	26,33	50,96	37,98	106,18	3,12	Normal
221	44,49	21,79	31,47	22,7	113,78	-0,28	Normal
222	54,95	5,87	53	49,09	126,97	-0,63	Normal
223	56,1	13,11	62,64	43	116,23	31,17	Normal
224	69,4	18,9	75,97	50,5	103,58	-0,44	Normal
225	89,83	22,64	90,56	67,2	100,5	3,04	Normal
226	59,73	7,72	55,34	52	125,17	3,24	Normal
227	63,96	16,06	63,12	47,9	142,36	6,3	Normal
228	61,54	19,68	52,89	41,86	118,69	4,82	Normal
229	38,05	8,3	26,24	29,74	123,8	3,89	Normal
230	43,44	10,1	36,03	33,34	137,44	-3,11	Normal
231	65,61	23,14	62,58	42,47	124,13	-4,08	Normal
232	53,91	12,94	39	40,97	118,19	5,07	Normal
233	43,12	13,82	40,35	29,3	128,52	0,97	Normal
234	40,68	9,15	31,02	31,53	139,12	-2,51	Normal

235	37,73	9,39	42	28,35	135,74	13,68	Normal
236	63,93	19,97	40,18	43,96	113,07	-11,06	Normal
237	61,82	13,6	64	48,22	121,78	1,3	Normal
238	62,14	13,96	58	48,18	133,28	4,96	Normal
239	69	13,29	55,57	55,71	126,61	10,83	Normal
240	56,45	19,44	43,58	37	139,19	-1,86	Normal
241	41,65	8,84	36,03	32,81	116,56	-6,05	Normal
242	51,53	13,52	35	38,01	126,72	13,93	Normal
243	39,09	5,54	26,93	33,55	131,58	-0,76	Normal
244	34,65	7,51	43	27,14	123,99	-4,08	Normal
245	63,03	27,34	51,61	35,69	114,51	7,44	Normal
246	47,81	10,69	54	37,12	125,39	-0,4	Normal
247	46,64	15,85	40	30,78	119,38	9,06	Normal
248	49,83	16,74	28	33,09	121,44	1,91	Normal
249	47,32	8,57	35,56	38,75	120,58	1,63	Normal
250	50,75	20,24	37	30,52	122,34	2,29	Normal
251	36,16	-0,81	33,63	36,97	135,94	-2,09	Normal
252	40,75	1,84	50	38,91	139,25	0,67	Normal
253	42,92	-5,85	58	48,76	121,61	-3,36	Normal
254	63,79	21,35	66	42,45	119,55	12,38	Normal
255	72,96	19,58	61,01	53,38	111,23	0,81	Normal
256	67,54	14,66	58	52,88	123,63	25,97	Normal
257	54,75	9,75	48	45	123,04	8,24	Normal
258	50,16	-2,97	42	53,13	131,8	-8,29	Normal
259	40,35	10,19	37,97	30,15	128,01	0,46	Normal
260	63,62	16,93	49,35	46,68	117,09	-0,36	Normal
261	54,14	11,94	43	42,21	122,21	0,15	Normal
262	74,98	14,92	53,73	60,05	105,65	1,59	Normal
263	42,52	14,38	25,32	28,14	128,91	0,76	Normal
264	33,79	3,68	25,5	30,11	128,33	-1,78	Normal
265	54,5	6,82	47	47,68	111,79	-4,41	Normal
266	48,17	9,59	39,71	38,58	135,62	5,36	Normal
267	46,37	10,22	42,7	36,16	121,25	-0,54	Normal
268	52,86	9,41	46,99	43,45	123,09	1,86	Normal

269	57,15	16,49	42,84	40,66	113,81	5,02	Normal
270	37,14	16,48	24	20,66	125,01	7,37	Normal
271	51,31	8,88	57	42,44	126,47	-2,14	Normal
272	42,52	16,54	42	25,97	120,63	7,88	Normal
273	39,36	7,01	37	32,35	117,82	1,9	Normal
274	35,88	1,11	43,46	34,77	126,92	-1,63	Normal
275	43,19	9,98	28,94	33,22	123,47	1,74	Normal
276	67,29	16,72	51	50,57	137,59	4,96	Normal
277	51,33	13,63	33,26	37,69	131,31	1,79	Normal
278	65,76	13,21	44	52,55	129,39	-1,98	Normal
279	40,41	-1,33	30,98	41,74	119,34	-6,17	Normal
280	48,8	18,02	52	30,78	139,15	10,44	Normal
281	50,09	13,43	34,46	36,66	119,13	3,09	Normal
282	64,26	14,5	43,9	49,76	115,39	5,95	Normal
283	53,68	13,45	41,58	40,24	113,91	2,74	Normal
284	49	13,11	51,87	35,88	126,4	0,54	Normal
285	59,17	14,56	43,2	44,6	121,04	2,83	Normal
286	67,8	16,55	43,26	51,25	119,69	4,87	Normal
287	61,73	17,11	46,9	44,62	120,92	3,09	Normal
288	33,04	-0,32	19,07	33,37	120,39	9,35	Normal
289	74,57	15,72	58,62	58,84	105,42	0,6	Normal
290	44,43	14,17	32,24	30,26	131,72	-3,6	Normal
291	36,42	13,88	20,24	22,54	126,08	0,18	Normal
292	51,08	14,21	35,95	36,87	115,8	6,91	Normal
293	34,76	2,63	29,5	32,12	127,14	-0,46	Normal
294	48,9	5,59	55,5	43,32	137,11	19,85	Normal
295	46,24	10,06	37	36,17	128,06	-5,1	Normal
296	46,43	6,62	48,1	39,81	130,35	2,45	Normal
297	39,66	16,21	36,67	23,45	131,92	-4,97	Normal
298	45,58	18,76	33,77	26,82	116,8	3,13	Normal
299	66,51	20,9	31,73	45,61	128,9	1,52	Normal
300	82,91	29,89	58,25	53,01	110,71	6,08	Normal
301	50,68	6,46	35	44,22	116,59	-0,21	Normal
302	89,01	26,08	69,02	62,94	111,48	6,06	Normal

303	54,6	21,49	29,36	33,11	118,34	-1,47	Normal
304	34,38	2,06	32,39	32,32	128,3	-3,37	Normal
305	45,08	12,31	44,58	32,77	147,89	-8,94	Normal
306	47,9	13,62	36	34,29	117,45	-4,25	Normal
307	53,94	20,72	29,22	33,22	114,37	-0,42	Normal
308	61,45	22,69	46,17	38,75	125,67	-2,71	Normal
309	45,25	8,69	41,58	36,56	118,55	0,21	Normal
310	33,84	5,07	36,64	28,77	123,95	-0,2	Normal

EK-2: Vertebral Column Gerçek Veri Uygulamasında LDA, QDA, LRA ve PRA Yöntemlerinde Tahmin Edilen Grup Üyeliklerinin Olasılıkları

No	ldapr1	ldapr2	ldapr3	qdap1	qdap2	qdap3	logitp1	logitp2	logitp3	probitp1	probitp2	probitp3
1	0,74	0,02	0,24	0,87	0,02	0,10	0,86	0,01	0,13	0,84	0,00	0,15
2	0,64	0,01	0,35	0,82	0,00	0,18	0,69	0,01	0,30	0,67	0,00	0,33
3	0,39	0,02	0,59	0,29	0,04	0,67	0,43	0,01	0,56	0,42	0,00	0,58
4	0,54	0,16	0,31	0,60	0,16	0,24	0,44	0,38	0,19	0,41	0,40	0,19
5	0,44	0,08	0,48	0,67	0,09	0,24	0,38	0,07	0,56	0,35	0,08	0,57
6	0,45	0,00	0,55	0,42	0,00	0,58	0,37	0,00	0,63	0,38	0,00	0,62
7	0,33	0,01	0,66	0,33	0,01	0,66	0,24	0,04	0,73	0,24	0,04	0,73
8	0,41	0,00	0,59	0,22	0,01	0,77	0,37	0,00	0,63	0,37	0,00	0,63
9	0,35	0,02	0,63	0,31	0,02	0,67	0,31	0,13	0,56	0,30	0,14	0,55
10	0,79	0,10	0,11	0,14	0,86	0,00	0,98	0,00	0,02	0,99	0,00	0,01
11	0,56	0,00	0,44	0,59	0,02	0,39	0,61	0,00	0,39	0,59	0,00	0,41
12	0,92	0,00	0,08	0,86	0,00	0,14	0,97	0,00	0,03	0,98	0,00	0,02
13	0,61	0,01	0,38	0,73	0,00	0,27	0,70	0,03	0,27	0,68	0,02	0,29
14	0,72	0,01	0,27	0,84	0,00	0,16	0,77	0,03	0,20	0,75	0,03	0,22
15	0,58	0,01	0,41	0,71	0,00	0,28	0,69	0,03	0,29	0,67	0,03	0,31
16	0,35	0,01	0,64	0,37	0,00	0,63	0,30	0,01	0,68	0,31	0,01	0,68
17	0,29	0,02	0,68	0,20	0,02	0,78	0,33	0,01	0,66	0,32	0,01	0,67
18	0,23	0,00	0,76	0,16	0,00	0,84	0,17	0,00	0,82	0,18	0,00	0,82
19	0,58	0,00	0,42	0,65	0,00	0,35	0,64	0,00	0,36	0,63	0,00	0,37
20	0,58	0,00	0,41	0,71	0,00	0,29	0,65	0,00	0,35	0,63	0,00	0,37
21	0,23	0,00	0,76	0,18	0,00	0,81	0,15	0,01	0,84	0,16	0,01	0,83
22	0,37	0,00	0,63	0,47	0,00	0,53	0,34	0,01	0,65	0,35	0,01	0,65
23	0,52	0,18	0,30	0,43	0,31	0,26	0,38	0,52	0,10	0,39	0,51	0,10
24	0,49	0,00	0,51	0,48	0,00	0,52	0,48	0,00	0,51	0,48	0,00	0,52
25	0,94	0,00	0,06	0,93	0,00	0,07	0,99	0,00	0,01	1,00	0,00	0,00
26	0,73	0,00	0,27	0,88	0,00	0,12	0,81	0,00	0,19	0,80	0,00	0,20
27	0,79	0,00	0,21	0,57	0,00	0,43	0,87	0,00	0,13	0,87	0,00	0,13
28	0,73	0,02	0,25	0,85	0,04	0,11	0,90	0,01	0,08	0,90	0,01	0,09
29	0,91	0,00	0,09	0,94	0,00	0,06	0,97	0,00	0,03	0,97	0,00	0,03
30	0,41	0,01	0,58	0,42	0,01	0,58	0,44	0,01	0,55	0,43	0,00	0,56
31	0,48	0,06	0,46	0,68	0,02	0,30	0,52	0,12	0,36	0,49	0,14	0,37

32	0,71	0,07	0,23	0,94	0,04	0,02	0,85	0,02	0,13	0,83	0,02	0,15
33	0,71	0,00	0,29	0,80	0,00	0,20	0,76	0,01	0,23	0,74	0,01	0,25
34	0,46	0,00	0,54	0,50	0,00	0,50	0,56	0,00	0,44	0,55	0,00	0,45
35	0,77	0,00	0,23	0,92	0,00	0,07	0,89	0,00	0,10	0,89	0,00	0,11
36	0,74	0,00	0,26	0,91	0,00	0,09	0,84	0,00	0,16	0,83	0,00	0,17
37	0,27	0,00	0,72	0,18	0,00	0,82	0,24	0,01	0,75	0,25	0,00	0,74
38	0,67	0,00	0,33	0,73	0,00	0,27	0,68	0,00	0,32	0,68	0,00	0,32
39	0,67	0,00	0,33	0,86	0,00	0,14	0,81	0,01	0,18	0,80	0,00	0,20
40	0,56	0,00	0,43	0,69	0,00	0,31	0,60	0,01	0,39	0,59	0,00	0,41
41	0,89	0,00	0,11	0,94	0,00	0,06	0,95	0,00	0,05	0,95	0,00	0,05
42	0,34	0,02	0,65	0,39	0,01	0,60	0,24	0,01	0,75	0,25	0,01	0,75
43	0,46	0,01	0,53	0,53	0,00	0,47	0,38	0,03	0,60	0,38	0,02	0,60
44	0,35	0,00	0,65	0,45	0,01	0,55	0,31	0,01	0,68	0,31	0,01	0,68
45	0,24	0,02	0,74	0,08	0,08	0,84	0,36	0,00	0,64	0,35	0,00	0,64
46	0,67	0,00	0,33	0,77	0,00	0,23	0,84	0,00	0,16	0,83	0,00	0,17
47	0,76	0,00	0,23	0,85	0,00	0,15	0,86	0,01	0,13	0,86	0,01	0,14
48	0,79	0,00	0,21	0,88	0,00	0,12	0,91	0,00	0,09	0,90	0,00	0,10
49	0,72	0,00	0,28	0,80	0,00	0,19	0,85	0,00	0,15	0,84	0,00	0,16
50	0,81	0,00	0,19	0,83	0,00	0,17	0,86	0,00	0,14	0,85	0,00	0,15
51	0,58	0,01	0,41	0,70	0,00	0,30	0,59	0,02	0,40	0,57	0,01	0,42
52	0,94	0,00	0,06	0,88	0,08	0,04	0,96	0,01	0,03	0,97	0,01	0,02
53	0,80	0,00	0,20	0,93	0,00	0,07	0,88	0,01	0,11	0,89	0,00	0,11
54	0,85	0,01	0,14	0,91	0,02	0,07	0,95	0,01	0,04	0,96	0,00	0,04
55	0,75	0,00	0,24	0,78	0,00	0,21	0,89	0,00	0,11	0,88	0,00	0,12
56	0,74	0,06	0,20	0,95	0,03	0,02	0,90	0,01	0,09	0,88	0,01	0,10
57	0,43	0,03	0,54	0,68	0,02	0,30	0,39	0,02	0,59	0,38	0,02	0,61
58	0,54	0,01	0,45	0,69	0,00	0,31	0,63	0,00	0,37	0,61	0,00	0,39
59	0,65	0,00	0,35	0,75	0,00	0,25	0,76	0,00	0,24	0,75	0,00	0,25
60	0,36	0,01	0,63	0,38	0,00	0,62	0,28	0,04	0,68	0,28	0,04	0,67
61	0,01	0,89	0,10	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
62	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
63	0,07	0,24	0,69	0,00	1,00	0,00	0,00	0,94	0,06	0,00	0,89	0,11
64	0,01	0,85	0,14	0,00	1,00	0,00	0,00	0,99	0,01	0,00	0,99	0,01
65	0,02	0,16	0,82	0,00	0,32	0,68	0,01	0,33	0,66	0,00	0,33	0,67
66	0,47	0,23	0,30	0,00	1,00	0,00	0,01	0,99	0,00	0,00	0,99	0,00
67	0,02	0,77	0,21	0,00	1,00	0,00	0,00	0,99	0,01	0,00	0,99	0,01

68	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
69	0,12	0,05	0,83	0,02	0,03	0,95	0,05	0,03	0,92	0,04	0,04	0,92
70	0,00	0,99	0,01	0,00	1,00	0,00	0,00	0,99	0,01	0,00	0,98	0,02
71	0,01	0,87	0,12	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
72	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
73	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
74	0,22	0,48	0,31	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
75	0,00	0,99	0,01	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
76	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
77	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
78	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
79	0,01	0,94	0,05	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
80	0,30	0,23	0,47	0,08	0,76	0,16	0,08	0,78	0,15	0,07	0,74	0,19
81	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
82	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
83	0,00	0,99	0,01	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
84	0,00	0,98	0,02	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
85	0,27	0,63	0,10	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
86	0,01	0,00	0,99	0,00	0,99	0,01	0,00	0,16	0,83	0,00	0,16	0,84
87	0,12	0,37	0,50	0,01	0,84	0,16	0,02	0,90	0,08	0,01	0,87	0,12
88	0,07	0,60	0,33	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
89	0,16	0,16	0,68	0,00	0,92	0,08	0,01	0,91	0,09	0,00	0,86	0,14
90	0,14	0,44	0,42	0,00	1,00	0,00	0,00	0,99	0,01	0,00	0,99	0,01
91	0,04	0,80	0,16	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
92	0,01	0,89	0,09	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
93	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
94	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
95	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
96	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
97	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
98	0,00	0,98	0,02	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
99	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
100	0,06	0,76	0,18	0,00	1,00	0,00	0,00	0,99	0,01	0,00	0,99	0,01
101	0,05	0,80	0,15	0,00	0,99	0,01	0,00	0,99	0,00	0,00	1,00	0,00
102	0,00	0,98	0,02	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
103	0,00	0,96	0,03	0,00	1,00	0,00	0,00	0,99	0,01	0,00	0,99	0,01

104	0,03	0,83	0,14	0,00	1,00	0,00	0,01	0,94	0,05	0,00	0,92	0,07
105	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
106	0,51	0,01	0,48	0,63	0,01	0,36	0,55	0,08	0,37	0,53	0,09	0,38
107	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
108	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
109	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
110	0,00	0,95	0,05	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
111	0,00	0,98	0,01	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
112	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
113	0,00	0,97	0,03	0,00	1,00	0,00	0,00	0,94	0,06	0,00	0,90	0,10
114	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
115	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
116	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
117	0,01	0,98	0,01	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
118	0,12	0,02	0,87	0,04	0,51	0,45	0,02	0,52	0,46	0,01	0,48	0,51
119	0,05	0,82	0,12	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
120	0,02	0,93	0,06	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
121	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
122	0,06	0,94	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
123	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
124	0,00	0,98	0,01	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
125	0,01	0,99	0,01	0,00	1,00	0,00	0,00	0,99	0,00	0,00	1,00	0,00
126	0,04	0,91	0,06	0,00	1,00	0,00	0,00	0,99	0,00	0,00	1,00	0,00
127	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
128	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
129	0,00	0,99	0,01	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
130	0,11	0,82	0,07	0,00	1,00	0,00	0,02	0,96	0,01	0,03	0,96	0,02
131	0,02	0,51	0,47	0,00	0,64	0,36	0,00	0,79	0,21	0,00	0,71	0,29
132	0,00	0,99	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
133	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
134	0,00	0,96	0,04	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
135	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
136	0,00	0,99	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
137	0,00	0,99	0,01	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
138	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
139	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00

140	0,00	0,98	0,02	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
141	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
142	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
143	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
144	0,02	0,94	0,04	0,00	1,00	0,00	0,00	0,99	0,00	0,00	1,00	0,00
145	0,05	0,93	0,02	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
146	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
147	0,00	0,99	0,01	0,00	1,00	0,00	0,00	0,99	0,01	0,00	0,99	0,01
148	0,00	0,99	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
149	0,01	0,98	0,01	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
150	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
151	0,01	0,88	0,12	0,00	1,00	0,00	0,00	0,99	0,01	0,00	0,99	0,01
152	0,02	0,91	0,07	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
153	0,04	0,81	0,15	0,00	1,00	0,00	0,00	0,99	0,01	0,00	0,99	0,01
154	0,10	0,82	0,08	0,00	1,00	0,00	0,05	0,93	0,02	0,06	0,92	0,02
155	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
156	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
157	0,00	0,99	0,01	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
158	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
159	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
160	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
161	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
162	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
163	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
164	0,06	0,43	0,51	0,00	0,98	0,02	0,00	0,92	0,08	0,00	0,86	0,14
165	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
166	0,00	0,98	0,01	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
167	0,02	0,97	0,01	0,00	1,00	0,00	0,45	0,51	0,04	0,43	0,54	0,03
168	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
169	0,03	0,92	0,05	0,00	1,00	0,00	0,04	0,91	0,05	0,03	0,89	0,07
170	0,04	0,72	0,24	0,00	0,92	0,08	0,01	0,93	0,06	0,00	0,90	0,10
171	0,00	0,87	0,13	0,00	0,80	0,20	0,01	0,68	0,32	0,00	0,62	0,38
172	0,14	0,80	0,07	0,01	0,99	0,00	0,12	0,85	0,03	0,14	0,82	0,03
173	0,03	0,97	0,01	0,00	1,00	0,00	0,25	0,74	0,01	0,30	0,70	0,00
174	0,69	0,17	0,14	0,06	0,91	0,02	0,20	0,78	0,02	0,26	0,73	0,01
175	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00

176	0,00	0,99	0,01	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
177	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
178	0,00	0,99	0,01	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
179	0,71	0,23	0,05	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
180	0,03	0,03	0,95	0,00	1,00	0,00	0,00	0,91	0,09	0,00	0,82	0,18
181	0,00	0,99	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
182	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
183	0,00	0,99	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
184	0,00	0,97	0,03	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
185	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
186	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
187	0,00	0,98	0,02	0,00	1,00	0,00	0,00	0,99	0,01	0,00	0,99	0,01
188	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
189	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
190	0,01	0,99	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
191	0,01	0,99	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
192	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
193	0,00	0,97	0,03	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
194	0,02	0,79	0,20	0,00	0,98	0,02	0,00	0,98	0,02	0,00	0,97	0,03
195	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
196	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
197	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
198	0,00	0,97	0,03	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
199	0,00	0,92	0,08	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
200	0,02	0,73	0,25	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
201	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
202	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
203	0,00	0,99	0,01	0,00	1,00	0,00	0,00	0,99	0,01	0,00	0,99	0,01
204	0,11	0,80	0,09	0,00	1,00	0,00	0,02	0,97	0,01	0,02	0,96	0,01
205	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
206	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
207	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
208	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
209	0,42	0,54	0,05	0,00	1,00	0,00	0,14	0,86	0,00	0,21	0,79	0,00
210	0,62	0,00	0,38	0,65	0,00	0,35	0,71	0,01	0,28	0,70	0,01	0,29
211	0,23	0,01	0,76	0,17	0,00	0,82	0,21	0,01	0,78	0,22	0,01	0,77

212	0,12	0,01	0,87	0,04	0,00	0,95	0,08	0,01	0,91	0,07	0,01	0,92
213	0,28	0,00	0,72	0,23	0,00	0,77	0,29	0,00	0,71	0,30	0,00	0,70
214	0,14	0,00	0,86	0,04	0,00	0,96	0,08	0,00	0,91	0,09	0,00	0,91
215	0,34	0,00	0,66	0,26	0,00	0,74	0,35	0,00	0,65	0,38	0,00	0,62
216	0,15	0,02	0,83	0,05	0,03	0,93	0,06	0,00	0,94	0,05	0,00	0,95
217	0,11	0,01	0,88	0,03	0,01	0,96	0,09	0,01	0,90	0,09	0,01	0,91
218	0,23	0,00	0,77	0,17	0,00	0,83	0,12	0,01	0,87	0,13	0,00	0,87
219	0,67	0,02	0,32	0,78	0,01	0,21	0,82	0,02	0,16	0,80	0,02	0,19
220	0,87	0,00	0,13	0,92	0,00	0,08	0,95	0,00	0,05	0,96	0,00	0,04
221	0,03	0,02	0,96	0,00	0,01	0,99	0,01	0,00	0,99	0,00	0,00	1,00
222	0,02	0,86	0,12	0,00	1,00	0,00	0,00	0,99	0,01	0,00	0,99	0,01
223	0,14	0,19	0,67	0,01	0,63	0,36	0,28	0,03	0,70	0,26	0,03	0,71
224	0,01	0,68	0,30	0,00	0,69	0,31	0,03	0,25	0,73	0,01	0,28	0,71
225	0,02	0,05	0,93	0,00	0,02	0,98	0,01	0,01	0,98	0,00	0,01	0,98
226	0,01	0,01	0,98	0,00	0,07	0,93	0,00	0,02	0,98	0,00	0,02	0,98
227	0,23	0,03	0,74	0,16	0,01	0,83	0,20	0,05	0,76	0,19	0,05	0,75
228	0,40	0,00	0,60	0,40	0,00	0,60	0,32	0,01	0,67	0,33	0,00	0,67
229	0,11	0,00	0,89	0,02	0,00	0,97	0,04	0,00	0,96	0,04	0,00	0,96
230	0,14	0,00	0,85	0,04	0,03	0,94	0,13	0,00	0,87	0,13	0,00	0,87
231	0,24	0,03	0,73	0,20	0,01	0,79	0,16	0,03	0,81	0,15	0,03	0,81
232	0,31	0,00	0,69	0,28	0,00	0,72	0,30	0,00	0,70	0,31	0,00	0,69
233	0,12	0,00	0,88	0,03	0,00	0,97	0,04	0,00	0,96	0,04	0,00	0,96
234	0,18	0,01	0,81	0,06	0,02	0,91	0,11	0,07	0,82	0,10	0,09	0,81
235	0,34	0,00	0,66	0,12	0,04	0,84	0,28	0,00	0,72	0,29	0,00	0,71
236	0,05	0,03	0,91	0,00	0,02	0,97	0,03	0,01	0,96	0,02	0,01	0,97
237	0,03	0,01	0,96	0,00	0,02	0,98	0,01	0,02	0,97	0,00	0,02	0,98
238	0,02	0,11	0,86	0,00	0,11	0,89	0,00	0,20	0,80	0,00	0,21	0,79
239	0,10	0,00	0,90	0,08	0,01	0,92	0,04	0,00	0,96	0,03	0,00	0,97
240	0,39	0,00	0,60	0,39	0,01	0,60	0,43	0,00	0,57	0,43	0,00	0,57
241	0,21	0,03	0,76	0,11	0,06	0,83	0,07	0,23	0,69	0,05	0,26	0,69
242	0,16	0,00	0,84	0,04	0,00	0,96	0,06	0,00	0,94	0,06	0,00	0,94
243	0,32	0,00	0,67	0,22	0,01	0,77	0,41	0,00	0,59	0,42	0,00	0,58
244	0,58	0,02	0,40	0,69	0,01	0,30	0,68	0,07	0,26	0,65	0,07	0,27
245	0,13	0,01	0,87	0,03	0,01	0,96	0,11	0,00	0,89	0,11	0,00	0,89
246	0,50	0,01	0,49	0,61	0,01	0,39	0,55	0,04	0,41	0,53	0,04	0,43
247	0,49	0,00	0,51	0,52	0,00	0,48	0,41	0,01	0,58	0,41	0,00	0,58

248	0,21	0,01	0,78	0,14	0,01	0,85	0,13	0,01	0,87	0,13	0,00	0,87
249	0,53	0,00	0,46	0,67	0,00	0,33	0,57	0,01	0,43	0,56	0,00	0,44
250	0,05	0,00	0,95	0,00	0,00	1,00	0,01	0,00	0,99	0,01	0,00	0,99
251	0,02	0,00	0,97	0,00	0,01	0,99	0,01	0,00	0,99	0,00	0,00	1,00
252	0,02	0,05	0,94	0,00	0,48	0,52	0,01	0,00	0,99	0,00	0,00	1,00
253	0,16	0,11	0,73	0,04	0,14	0,82	0,12	0,37	0,51	0,10	0,38	0,52
254	0,11	0,09	0,80	0,01	0,04	0,95	0,08	0,04	0,89	0,06	0,05	0,89
255	0,02	0,64	0,34	0,00	0,99	0,01	0,00	0,97	0,03	0,00	0,95	0,05
256	0,08	0,06	0,86	0,02	0,02	0,96	0,03	0,06	0,91	0,02	0,08	0,90
257	0,01	0,00	0,99	0,00	0,09	0,91	0,00	0,00	1,00	0,00	0,00	1,00
258	0,27	0,00	0,73	0,19	0,00	0,81	0,23	0,00	0,77	0,24	0,00	0,76
259	0,16	0,02	0,82	0,05	0,01	0,94	0,10	0,01	0,89	0,09	0,01	0,90
260	0,14	0,01	0,85	0,05	0,00	0,95	0,07	0,01	0,92	0,07	0,00	0,93
261	0,07	0,26	0,68	0,01	0,23	0,76	0,03	0,06	0,91	0,02	0,08	0,90
262	0,44	0,00	0,56	0,40	0,00	0,60	0,34	0,00	0,66	0,35	0,00	0,64
263	0,24	0,00	0,76	0,11	0,00	0,89	0,14	0,00	0,86	0,15	0,00	0,85
264	0,12	0,04	0,85	0,03	0,08	0,89	0,08	0,00	0,92	0,07	0,00	0,93
265	0,07	0,01	0,92	0,01	0,00	0,98	0,02	0,01	0,97	0,01	0,01	0,98
266	0,23	0,01	0,76	0,16	0,00	0,84	0,20	0,00	0,80	0,20	0,00	0,80
267	0,09	0,02	0,89	0,02	0,01	0,97	0,04	0,01	0,95	0,04	0,01	0,96
268	0,35	0,03	0,62	0,40	0,01	0,59	0,32	0,04	0,64	0,31	0,05	0,65
269	0,74	0,00	0,26	0,74	0,00	0,26	0,80	0,01	0,19	0,79	0,00	0,21
270	0,06	0,01	0,93	0,00	0,01	0,98	0,03	0,00	0,97	0,02	0,00	0,97
271	0,59	0,01	0,40	0,68	0,00	0,32	0,73	0,02	0,26	0,71	0,01	0,28
272	0,36	0,01	0,63	0,50	0,01	0,50	0,37	0,00	0,62	0,37	0,00	0,63
273	0,10	0,01	0,89	0,02	0,01	0,97	0,07	0,00	0,93	0,06	0,00	0,94
274	0,33	0,00	0,67	0,28	0,00	0,72	0,23	0,00	0,76	0,24	0,00	0,76
275	0,02	0,01	0,97	0,00	0,05	0,95	0,00	0,02	0,97	0,00	0,02	0,98
276	0,16	0,00	0,84	0,07	0,00	0,93	0,06	0,01	0,94	0,05	0,00	0,94
277	0,03	0,01	0,96	0,00	0,02	0,97	0,01	0,00	0,99	0,00	0,00	1,00
278	0,11	0,01	0,89	0,02	0,05	0,93	0,04	0,00	0,96	0,03	0,00	0,97
279	0,14	0,00	0,85	0,08	0,03	0,89	0,09	0,05	0,86	0,09	0,06	0,85
280	0,35	0,01	0,64	0,37	0,00	0,63	0,27	0,01	0,71	0,28	0,01	0,71
281	0,13	0,08	0,79	0,04	0,04	0,91	0,06	0,08	0,86	0,04	0,11	0,86
282	0,31	0,03	0,66	0,36	0,01	0,63	0,29	0,02	0,70	0,28	0,01	0,70
283	0,16	0,01	0,83	0,06	0,01	0,94	0,14	0,00	0,86	0,14	0,00	0,85

284	0,15	0,02	0,83	0,06	0,01	0,94	0,07	0,02	0,91	0,06	0,02	0,92
285	0,10	0,04	0,86	0,02	0,04	0,94	0,03	0,06	0,91	0,02	0,08	0,91
286	0,16	0,02	0,82	0,07	0,01	0,92	0,08	0,02	0,89	0,08	0,03	0,89
287	0,29	0,03	0,68	0,36	0,08	0,55	0,16	0,03	0,81	0,15	0,03	0,82
288	0,07	0,23	0,70	0,01	0,16	0,83	0,05	0,04	0,91	0,03	0,06	0,91
289	0,28	0,00	0,72	0,20	0,00	0,80	0,19	0,00	0,81	0,21	0,00	0,79
290	0,66	0,00	0,34	0,63	0,00	0,37	0,67	0,00	0,33	0,66	0,00	0,34
291	0,41	0,03	0,56	0,53	0,01	0,47	0,37	0,04	0,59	0,36	0,05	0,60
292	0,19	0,00	0,80	0,09	0,00	0,91	0,11	0,00	0,89	0,11	0,00	0,89
293	0,02	0,11	0,87	0,00	0,61	0,39	0,00	0,42	0,57	0,00	0,39	0,61
294	0,16	0,00	0,84	0,04	0,00	0,95	0,09	0,00	0,91	0,09	0,00	0,91
295	0,06	0,01	0,93	0,01	0,00	0,99	0,03	0,00	0,97	0,02	0,00	0,98
296	0,44	0,00	0,56	0,44	0,00	0,56	0,49	0,00	0,51	0,50	0,00	0,50
297	0,72	0,00	0,27	0,84	0,00	0,16	0,83	0,00	0,17	0,81	0,00	0,18
298	0,15	0,00	0,85	0,05	0,06	0,88	0,03	0,01	0,95	0,03	0,02	0,96
299	0,23	0,09	0,68	0,10	0,08	0,81	0,13	0,29	0,57	0,10	0,32	0,57
300	0,15	0,02	0,82	0,07	0,02	0,91	0,07	0,00	0,92	0,06	0,00	0,93
301	0,05	0,26	0,70	0,00	0,12	0,88	0,02	0,33	0,65	0,00	0,35	0,65
302	0,62	0,00	0,38	0,70	0,00	0,30	0,61	0,00	0,39	0,60	0,00	0,40
303	0,16	0,00	0,84	0,05	0,00	0,95	0,09	0,00	0,91	0,09	0,00	0,91
304	0,05	0,00	0,95	0,00	0,01	0,99	0,01	0,00	0,99	0,01	0,00	0,99
305	0,42	0,00	0,58	0,42	0,00	0,58	0,43	0,00	0,57	0,43	0,00	0,57
306	0,68	0,00	0,32	0,77	0,00	0,23	0,71	0,00	0,29	0,69	0,00	0,31
307	0,25	0,00	0,75	0,25	0,00	0,75	0,19	0,00	0,81	0,20	0,00	0,80
308	0,25	0,01	0,74	0,23	0,01	0,77	0,22	0,00	0,77	0,23	0,00	0,77
309	0,30	0,00	0,70	0,27	0,00	0,72	0,30	0,00	0,70	0,31	0,00	0,69

EK-3: N=30 Birimlik Simülasyon Verisi İçin LDA, QDA, LRA ve PRA Yöntemlerinde Tahmin Edilen Grup Üyeliklerinin Olasılıkları

No	ldapr1	ldapr2	ldapr3	qdapr1	qdapr2	qdapr3	logitp1	logitp2	logitp3	probitp1	probitp2	probitp3
1	0,48	0,13	0,39	0,23	0,16	0,61	0,41	0,10	0,50	0,40	0,08	0,53
2	0,34	0,65	0,00	0,20	0,79	0,01	0,33	0,67	0,00	0,33	0,67	0,00
3	0,88	0,04	0,08	0,97	0,01	0,03	0,75	0,07	0,18	0,73	0,04	0,22
4	0,59	0,27	0,14	0,58	0,24	0,18	0,57	0,24	0,19	0,55	0,24	0,21
5	0,69	0,24	0,07	0,69	0,17	0,14	0,63	0,26	0,11	0,62	0,26	0,12
6	0,76	0,06	0,18	0,84	0,03	0,14	0,61	0,07	0,31	0,59	0,05	0,36
7	0,72	0,12	0,16	0,70	0,07	0,23	0,62	0,13	0,26	0,60	0,10	0,30
8	0,68	0,29	0,03	0,66	0,23	0,11	0,62	0,33	0,05	0,62	0,33	0,05
9	0,16	0,83	0,01	0,44	0,56	0,00	0,23	0,75	0,01	0,22	0,77	0,01
10	0,27	0,72	0,02	0,46	0,54	0,00	0,32	0,66	0,02	0,31	0,68	0,01
11	0,16	0,84	0,00	0,02	0,98	0,00	0,16	0,84	0,00	0,16	0,84	0,00
12	0,74	0,22	0,04	0,72	0,15	0,13	0,67	0,27	0,07	0,67	0,26	0,07
13	0,26	0,73	0,02	0,46	0,54	0,00	0,32	0,66	0,02	0,31	0,68	0,01
14	0,10	0,89	0,00	0,43	0,57	0,00	0,16	0,83	0,01	0,15	0,85	0,00
15	0,32	0,36	0,32	0,17	0,72	0,11	0,37	0,25	0,37	0,35	0,27	0,39
16	0,09	0,91	0,00	0,00	1,00	0,00	0,10	0,90	0,00	0,09	0,91	0,00
17	0,18	0,82	0,00	0,10	0,89	0,00	0,19	0,81	0,00	0,19	0,81	0,00
18	0,50	0,49	0,02	0,51	0,45	0,04	0,48	0,50	0,02	0,48	0,50	0,01
19	0,20	0,80	0,00	0,00	0,99	0,00	0,17	0,83	0,00	0,18	0,82	0,00
20	0,65	0,26	0,09	0,66	0,19	0,15	0,61	0,26	0,13	0,59	0,26	0,15
21	0,01	0,00	0,99	0,00	0,00	1,00	0,01	0,00	0,99	0,00	0,00	1,00
22	0,16	0,01	0,83	0,01	0,01	0,98	0,10	0,01	0,90	0,10	0,00	0,89
23	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00
24	0,01	0,00	0,99	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00
25	0,14	0,00	0,86	0,01	0,01	0,98	0,07	0,00	0,93	0,07	0,00	0,93
26	0,77	0,18	0,05	0,76	0,10	0,14	0,69	0,22	0,09	0,69	0,21	0,10
27	0,01	0,00	0,98	0,00	0,00	1,00	0,01	0,00	0,99	0,01	0,00	0,99
28	0,28	0,06	0,66	0,03	0,09	0,88	0,22	0,04	0,74	0,23	0,02	0,74
29	0,06	0,02	0,92	0,00	0,03	0,97	0,05	0,01	0,94	0,06	0,00	0,94
30	0,53	0,33	0,14	0,53	0,33	0,13	0,53	0,29	0,18	0,51	0,29	0,20

EK-4: N=90 Birimlik Simülasyon Verisi İçin LDA, QDA, LRA ve PRA Yöntemlerinde Tahmin Edilen Grup Üyeliklerinin Olasılıkları

No	ldapr1	ldapr2	ldapr3	qdapr1	qdapr2	qdapr3	logitp1	logitp2	logitp3	probitp1	probitp2	probitp3
1	0,61	0,34	0,05	0,70	0,20	0,10	0,64	0,27	0,08	0,59	0,27	0,14
2	0,64	0,19	0,17	0,76	0,07	0,17	0,67	0,05	0,27	0,62	0,06	0,32
3	0,45	0,02	0,53	0,42	0,00	0,58	0,38	0,00	0,62	0,35	0,00	0,65
4	0,66	0,15	0,19	0,81	0,04	0,15	0,67	0,04	0,29	0,60	0,04	0,36
5	0,34	0,02	0,65	0,09	0,00	0,91	0,26	0,00	0,74	0,32	0,00	0,68
6	0,69	0,23	0,08	0,62	0,18	0,20	0,71	0,18	0,12	0,62	0,15	0,22
7	0,69	0,18	0,14	0,81	0,06	0,13	0,71	0,08	0,21	0,62	0,07	0,31
8	0,65	0,12	0,22	0,80	0,03	0,16	0,64	0,03	0,33	0,58	0,03	0,40
9	0,25	0,75	0,00	0,09	0,88	0,03	0,15	0,84	0,00	0,17	0,82	0,00
10	0,62	0,19	0,18	0,61	0,12	0,27	0,66	0,05	0,30	0,62	0,06	0,32
11	0,66	0,09	0,26	0,76	0,02	0,21	0,63	0,02	0,35	0,53	0,01	0,45
12	0,51	0,45	0,04	0,59	0,33	0,08	0,61	0,32	0,08	0,53	0,38	0,09
13	0,64	0,28	0,08	0,79	0,11	0,09	0,70	0,16	0,13	0,62	0,18	0,20
14	0,68	0,20	0,12	0,82	0,07	0,11	0,72	0,10	0,18	0,63	0,10	0,28
15	0,65	0,14	0,21	0,80	0,04	0,16	0,65	0,03	0,32	0,59	0,03	0,38
16	0,68	0,16	0,16	0,83	0,04	0,13	0,70	0,06	0,24	0,61	0,05	0,34
17	0,68	0,15	0,18	0,83	0,04	0,13	0,69	0,05	0,26	0,60	0,04	0,36
18	0,55	0,41	0,04	0,67	0,26	0,08	0,60	0,34	0,07	0,54	0,35	0,10
19	0,50	0,47	0,03	0,53	0,40	0,08	0,49	0,47	0,04	0,47	0,46	0,06
20	0,50	0,48	0,03	0,52	0,41	0,08	0,48	0,48	0,04	0,47	0,47	0,06
21	0,65	0,13	0,22	0,80	0,04	0,16	0,65	0,03	0,32	0,58	0,03	0,39
22	0,68	0,10	0,22	0,78	0,03	0,19	0,66	0,03	0,31	0,56	0,02	0,42
23	0,63	0,10	0,27	0,75	0,03	0,22	0,60	0,02	0,38	0,55	0,02	0,43
24	0,65	0,17	0,18	0,77	0,06	0,17	0,67	0,05	0,28	0,61	0,05	0,34
25	0,67	0,21	0,12	0,83	0,07	0,10	0,71	0,10	0,19	0,63	0,10	0,28
26	0,48	0,50	0,02	0,40	0,52	0,08	0,42	0,55	0,03	0,43	0,52	0,05
27	0,66	0,15	0,20	0,81	0,04	0,15	0,66	0,04	0,30	0,59	0,04	0,37
28	0,68	0,19	0,12	0,81	0,07	0,12	0,72	0,09	0,19	0,62	0,08	0,29
29	0,67	0,17	0,16	0,83	0,05	0,12	0,70	0,06	0,24	0,61	0,06	0,33

30	0,68	0,24	0,08	0,71	0,14	0,15	0,71	0,17	0,12	0,62	0,15	0,22
31	0,60	0,35	0,05	0,62	0,27	0,11	0,61	0,32	0,07	0,57	0,31	0,12
32	0,20	0,80	0,00	0,01	0,96	0,03	0,06	0,94	0,00	0,09	0,91	0,00
33	0,09	0,91	0,00	0,00	0,99	0,01	0,01	0,99	0,00	0,01	0,99	0,00
34	0,01	0,99	0,00	0,00	0,99	0,01	0,00	1,00	0,00	0,00	1,00	0,00
35	0,15	0,85	0,00	0,02	0,97	0,02	0,06	0,94	0,00	0,07	0,93	0,00
36	0,18	0,82	0,00	0,00	0,97	0,03	0,05	0,95	0,00	0,07	0,93	0,00
37	0,45	0,54	0,01	0,07	0,82	0,11	0,27	0,72	0,01	0,35	0,62	0,02
38	0,66	0,24	0,10	0,82	0,08	0,10	0,71	0,12	0,17	0,63	0,13	0,25
39	0,46	0,53	0,01	0,17	0,73	0,10	0,32	0,66	0,02	0,38	0,59	0,03
40	0,23	0,77	0,01	0,06	0,91	0,03	0,18	0,81	0,01	0,18	0,82	0,00
41	0,16	0,84	0,00	0,00	0,97	0,03	0,03	0,97	0,00	0,05	0,95	0,00
42	0,28	0,72	0,01	0,10	0,86	0,04	0,16	0,83	0,00	0,19	0,80	0,00
43	0,22	0,77	0,00	0,01	0,96	0,03	0,07	0,93	0,00	0,11	0,89	0,00
44	0,20	0,80	0,00	0,03	0,94	0,03	0,08	0,92	0,00	0,11	0,89	0,00
45	0,26	0,74	0,00	0,01	0,95	0,04	0,08	0,92	0,00	0,14	0,86	0,00
46	0,54	0,43	0,03	0,30	0,56	0,14	0,47	0,50	0,03	0,49	0,44	0,07
47	0,32	0,68	0,00	0,01	0,93	0,06	0,11	0,89	0,00	0,18	0,81	0,00
48	0,09	0,91	0,00	0,00	0,99	0,01	0,01	0,99	0,00	0,02	0,98	0,00
49	0,06	0,94	0,00	0,00	0,98	0,02	0,00	1,00	0,00	0,00	1,00	0,00
50	0,28	0,72	0,00	0,03	0,92	0,04	0,12	0,88	0,00	0,17	0,83	0,00
51	0,42	0,55	0,02	0,39	0,55	0,07	0,49	0,46	0,05	0,43	0,52	0,05
52	0,33	0,66	0,01	0,16	0,79	0,05	0,22	0,78	0,01	0,25	0,74	0,01
53	0,07	0,93	0,00	0,00	0,99	0,01	0,00	1,00	0,00	0,01	0,99	0,00
54	0,20	0,79	0,00	0,05	0,93	0,02	0,11	0,89	0,00	0,13	0,87	0,00
55	0,17	0,83	0,00	0,00	0,97	0,03	0,03	0,97	0,00	0,05	0,95	0,00
56	0,13	0,86	0,00	0,00	0,98	0,02	0,03	0,97	0,00	0,04	0,96	0,00
57	0,04	0,96	0,00	0,00	0,99	0,01	0,00	1,00	0,00	0,00	1,00	0,00
58	0,62	0,31	0,08	0,78	0,13	0,09	0,69	0,18	0,13	0,61	0,20	0,18
59	0,22	0,77	0,00	0,02	0,95	0,03	0,08	0,92	0,00	0,12	0,88	0,00
60	0,62	0,33	0,05	0,66	0,23	0,12	0,64	0,28	0,08	0,59	0,27	0,14
61	0,59	0,28	0,13	0,40	0,31	0,29	0,68	0,08	0,24	0,63	0,13	0,24
62	0,28	0,01	0,71	0,02	0,00	0,98	0,21	0,00	0,79	0,29	0,00	0,71
63	0,34	0,02	0,64	0,03	0,01	0,97	0,26	0,00	0,74	0,34	0,00	0,66

64	0,02	0,00	0,98	0,00	0,00	1,00	0,02	0,00	0,98	0,03	0,00	0,97
65	0,63	0,08	0,29	0,76	0,02	0,22	0,59	0,02	0,39	0,52	0,01	0,47
66	0,14	0,00	0,86	0,00	0,00	1,00	0,10	0,00	0,90	0,16	0,00	0,84
67	0,08	0,00	0,92	0,00	0,00	1,00	0,06	0,00	0,94	0,13	0,00	0,87
68	0,69	0,18	0,13	0,79	0,07	0,14	0,72	0,09	0,19	0,62	0,07	0,31
69	0,19	0,01	0,81	0,00	0,00	1,00	0,13	0,00	0,87	0,22	0,00	0,78
70	0,02	0,00	0,98	0,00	0,00	1,00	0,01	0,00	0,99	0,03	0,00	0,97
71	0,10	0,00	0,90	0,00	0,00	1,00	0,07	0,00	0,93	0,12	0,00	0,88
72	0,02	0,00	0,98	0,00	0,00	1,00	0,02	0,00	0,98	0,03	0,00	0,97
73	0,01	0,00	0,99	0,00	0,00	1,00	0,01	0,00	0,99	0,01	0,00	0,99
74	0,21	0,79	0,00	0,00	0,96	0,04	0,04	0,95	0,00	0,08	0,92	0,00
75	0,58	0,38	0,04	0,55	0,33	0,12	0,57	0,37	0,06	0,55	0,35	0,10
76	0,35	0,02	0,64	0,15	0,00	0,85	0,27	0,00	0,73	0,32	0,00	0,68
77	0,01	0,00	0,99	0,00	0,00	1,00	0,00	0,00	1,00	0,01	0,00	0,99
78	0,06	0,00	0,94	0,00	0,00	1,00	0,04	0,00	0,96	0,10	0,00	0,90
79	0,17	0,00	0,82	0,00	0,00	1,00	0,12	0,00	0,88	0,20	0,00	0,80
80	0,56	0,05	0,39	0,65	0,01	0,33	0,49	0,01	0,50	0,45	0,00	0,54
81	0,01	0,00	0,99	0,00	0,00	1,00	0,01	0,00	0,99	0,02	0,00	0,98
82	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00
83	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00
84	0,32	0,01	0,66	0,14	0,00	0,86	0,25	0,00	0,75	0,29	0,00	0,71
85	0,29	0,01	0,70	0,11	0,00	0,89	0,22	0,00	0,78	0,26	0,00	0,74
86	0,30	0,01	0,69	0,14	0,00	0,86	0,23	0,00	0,77	0,26	0,00	0,74
87	0,09	0,00	0,91	0,00	0,00	1,00	0,07	0,00	0,93	0,11	0,00	0,89
88	0,23	0,01	0,77	0,04	0,00	0,96	0,17	0,00	0,83	0,23	0,00	0,77
89	0,03	0,00	0,97	0,00	0,00	1,00	0,02	0,00	0,98	0,04	0,00	0,96
90	0,65	0,16	0,19	0,80	0,05	0,15	0,67	0,04	0,29	0,60	0,04	0,35

EK-5: N=180 Birimlik Simülasyon Verisi İçin LDA, QDA, LRA ve PRA Yöntemlerinde Tahmin Edilen Grup Üyeliklerinin Olasılıkları

No	ldapr1	ldapr2	ldapr3	qdapr1	qdapr2	qdapr3	logitp1	logitp2	logitp3	probitp1	probitp2	probitp3
1	0,68	0,07	0,25	0,79	0,00	0,21	0,61	0,00	0,39	0,59	0,00	0,41
2	0,60	0,36	0,05	0,80	0,11	0,08	0,77	0,10	0,13	0,70	0,18	0,12
3	0,80	0,05	0,15	0,60	0,02	0,38	0,82	0,01	0,17	0,69	0,00	0,30
4	0,56	0,42	0,02	0,69	0,25	0,06	0,65	0,30	0,04	0,59	0,37	0,04
5	0,56	0,43	0,02	0,55	0,37	0,08	0,57	0,40	0,03	0,55	0,43	0,03
6	0,77	0,06	0,17	0,80	0,01	0,19	0,78	0,01	0,21	0,67	0,00	0,32
7	0,55	0,02	0,43	0,65	0,00	0,35	0,46	0,00	0,54	0,45	0,00	0,55
8	0,71	0,09	0,20	0,84	0,01	0,15	0,68	0,01	0,32	0,64	0,00	0,36
9	0,53	0,40	0,06	0,36	0,27	0,37	0,70	0,05	0,25	0,70	0,12	0,18
10	0,12	0,88	0,00	0,04	0,95	0,01	0,04	0,96	0,00	0,05	0,95	0,00
11	0,67	0,03	0,30	0,75	0,00	0,25	0,64	0,00	0,36	0,55	0,00	0,45
12	0,30	0,69	0,01	0,31	0,66	0,04	0,41	0,56	0,03	0,35	0,64	0,01
13	0,72	0,18	0,10	0,88	0,02	0,10	0,78	0,03	0,20	0,73	0,04	0,23
14	0,71	0,25	0,04	0,81	0,10	0,09	0,80	0,12	0,08	0,74	0,15	0,11
15	0,71	0,11	0,18	0,84	0,01	0,15	0,68	0,01	0,31	0,66	0,01	0,34
16	0,59	0,03	0,38	0,69	0,00	0,31	0,50	0,00	0,50	0,48	0,00	0,52
17	0,59	0,03	0,38	0,69	0,00	0,31	0,50	0,00	0,50	0,48	0,00	0,52
18	0,57	0,40	0,03	0,78	0,15	0,06	0,74	0,18	0,08	0,66	0,26	0,08
19	0,67	0,26	0,07	0,86	0,05	0,09	0,78	0,05	0,17	0,73	0,09	0,18
20	0,67	0,25	0,07	0,86	0,05	0,10	0,78	0,05	0,17	0,73	0,08	0,19
21	0,53	0,46	0,02	0,58	0,36	0,06	0,56	0,41	0,03	0,53	0,45	0,02
22	0,78	0,08	0,13	0,81	0,01	0,18	0,81	0,01	0,18	0,71	0,01	0,28
23	0,70	0,23	0,07	0,87	0,05	0,08	0,80	0,06	0,14	0,74	0,09	0,17
24	0,35	0,65	0,01	0,34	0,62	0,03	0,31	0,68	0,01	0,32	0,68	0,00
25	0,71	0,09	0,20	0,84	0,01	0,16	0,67	0,01	0,33	0,64	0,00	0,36
26	0,60	0,37	0,03	0,79	0,14	0,06	0,75	0,18	0,07	0,67	0,26	0,07
27	0,62	0,35	0,03	0,79	0,15	0,07	0,75	0,19	0,07	0,67	0,25	0,07
28	0,75	0,09	0,16	0,86	0,01	0,13	0,75	0,01	0,24	0,68	0,01	0,31
29	0,73	0,19	0,08	0,88	0,03	0,09	0,81	0,04	0,15	0,75	0,06	0,19

30	0,60	0,02	0,38	0,72	0,00	0,28	0,53	0,00	0,47	0,48	0,00	0,52
31	0,68	0,08	0,23	0,77	0,01	0,23	0,61	0,00	0,39	0,60	0,00	0,39
32	0,77	0,07	0,16	0,82	0,01	0,17	0,78	0,01	0,20	0,69	0,01	0,31
33	0,79	0,13	0,08	0,79	0,04	0,17	0,84	0,04	0,12	0,76	0,04	0,20
34	0,53	0,46	0,01	0,09	0,79	0,11	0,33	0,67	0,01	0,40	0,60	0,01
35	0,54	0,01	0,45	0,40	0,00	0,60	0,56	0,00	0,44	0,43	0,00	0,57
36	0,63	0,33	0,05	0,83	0,09	0,07	0,78	0,10	0,12	0,71	0,17	0,12
37	0,74	0,15	0,11	0,88	0,02	0,10	0,79	0,02	0,19	0,73	0,03	0,24
38	0,42	0,01	0,57	0,45	0,00	0,55	0,34	0,00	0,66	0,35	0,00	0,65
39	0,69	0,21	0,10	0,85	0,03	0,12	0,76	0,03	0,22	0,72	0,05	0,23
40	0,38	0,01	0,62	0,40	0,00	0,60	0,34	0,00	0,66	0,31	0,00	0,69
41	0,72	0,24	0,05	0,82	0,09	0,10	0,81	0,10	0,08	0,74	0,14	0,12
42	0,68	0,05	0,27	0,81	0,00	0,19	0,62	0,00	0,38	0,57	0,00	0,43
43	0,74	0,17	0,09	0,88	0,03	0,09	0,80	0,04	0,16	0,74	0,05	0,21
44	0,73	0,10	0,17	0,86	0,01	0,13	0,71	0,01	0,28	0,67	0,01	0,32
45	0,78	0,10	0,12	0,84	0,02	0,14	0,81	0,02	0,17	0,73	0,02	0,26
46	0,72	0,08	0,20	0,85	0,01	0,15	0,69	0,01	0,30	0,64	0,00	0,36
47	0,75	0,18	0,07	0,86	0,04	0,10	0,83	0,05	0,12	0,75	0,07	0,18
48	0,65	0,31	0,04	0,81	0,11	0,07	0,77	0,15	0,08	0,70	0,21	0,09
49	0,41	0,59	0,01	0,37	0,58	0,04	0,35	0,63	0,01	0,36	0,63	0,01
50	0,70	0,22	0,09	0,87	0,04	0,09	0,78	0,04	0,18	0,73	0,06	0,21
51	0,37	0,01	0,63	0,35	0,00	0,65	0,29	0,00	0,71	0,30	0,00	0,70
52	0,70	0,07	0,24	0,82	0,00	0,17	0,65	0,00	0,35	0,60	0,00	0,39
53	0,77	0,19	0,05	0,62	0,16	0,22	0,82	0,11	0,06	0,76	0,13	0,11
54	0,59	0,02	0,38	0,71	0,00	0,29	0,54	0,00	0,46	0,48	0,00	0,52
55	0,69	0,27	0,04	0,79	0,12	0,09	0,79	0,14	0,07	0,72	0,18	0,10
56	0,62	0,33	0,05	0,83	0,09	0,08	0,78	0,10	0,12	0,71	0,16	0,12
57	0,72	0,25	0,03	0,62	0,22	0,16	0,77	0,18	0,05	0,72	0,21	0,08
58	0,51	0,02	0,47	0,47	0,00	0,53	0,38	0,00	0,62	0,42	0,00	0,58
59	0,75	0,10	0,15	0,87	0,01	0,12	0,76	0,01	0,23	0,69	0,01	0,30
60	0,68	0,09	0,23	0,75	0,01	0,25	0,60	0,00	0,39	0,61	0,00	0,39
61	0,05	0,95	0,00	0,00	1,00	0,00	0,04	0,96	0,00	0,03	0,97	0,00
62	0,04	0,96	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
63	0,11	0,89	0,00	0,03	0,96	0,01	0,06	0,94	0,00	0,06	0,94	0,00

64	0,40	0,60	0,00	0,12	0,83	0,05	0,21	0,79	0,00	0,27	0,73	0,00
65	0,11	0,89	0,00	0,00	0,99	0,01	0,01	0,99	0,00	0,01	0,99	0,00
66	0,05	0,95	0,00	0,00	0,99	0,01	0,00	1,00	0,00	0,00	1,00	0,00
67	0,18	0,82	0,00	0,10	0,88	0,01	0,15	0,84	0,01	0,15	0,85	0,00
68	0,04	0,96	0,00	0,00	0,99	0,01	0,00	1,00	0,00	0,00	1,00	0,00
69	0,03	0,97	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
70	0,55	0,44	0,01	0,03	0,83	0,13	0,29	0,70	0,00	0,38	0,61	0,01
71	0,22	0,78	0,00	0,05	0,93	0,02	0,07	0,93	0,00	0,10	0,90	0,00
72	0,31	0,68	0,00	0,07	0,89	0,03	0,12	0,87	0,00	0,18	0,82	0,00
73	0,51	0,48	0,01	0,11	0,80	0,10	0,31	0,68	0,01	0,38	0,62	0,01
74	0,09	0,91	0,00	0,01	0,98	0,01	0,01	0,99	0,00	0,02	0,98	0,00
75	0,10	0,90	0,00	0,01	0,98	0,01	0,02	0,98	0,00	0,02	0,98	0,00
76	0,05	0,95	0,00	0,00	0,99	0,01	0,00	1,00	0,00	0,00	1,00	0,00
77	0,08	0,92	0,00	0,00	0,98	0,02	0,00	1,00	0,00	0,00	1,00	0,00
78	0,03	0,97	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
79	0,13	0,87	0,00	0,03	0,96	0,01	0,03	0,97	0,00	0,04	0,96	0,00
80	0,09	0,91	0,00	0,00	0,99	0,01	0,01	0,99	0,00	0,01	0,99	0,00
81	0,11	0,89	0,00	0,01	0,98	0,01	0,02	0,98	0,00	0,02	0,98	0,00
82	0,41	0,58	0,01	0,46	0,50	0,04	0,42	0,57	0,02	0,40	0,59	0,01
83	0,21	0,79	0,00	0,00	0,96	0,04	0,02	0,98	0,00	0,03	0,97	0,00
84	0,13	0,87	0,00	0,01	0,98	0,01	0,02	0,98	0,00	0,02	0,98	0,00
85	0,23	0,77	0,00	0,02	0,96	0,02	0,05	0,95	0,00	0,08	0,92	0,00
86	0,18	0,82	0,00	0,00	0,98	0,02	0,02	0,98	0,00	0,04	0,96	0,00
87	0,26	0,74	0,00	0,01	0,96	0,03	0,05	0,95	0,00	0,09	0,91	0,00
88	0,45	0,54	0,01	0,46	0,49	0,05	0,44	0,55	0,02	0,43	0,56	0,01
89	0,57	0,42	0,01	0,05	0,81	0,14	0,34	0,66	0,01	0,42	0,58	0,01
90	0,07	0,93	0,00	0,01	0,98	0,00	0,01	0,99	0,00	0,01	0,99	0,00
91	0,15	0,84	0,00	0,07	0,92	0,01	0,06	0,94	0,00	0,08	0,92	0,00
92	0,02	0,98	0,00	0,00	0,99	0,01	0,00	1,00	0,00	0,00	1,00	0,00
93	0,32	0,67	0,00	0,29	0,68	0,03	0,26	0,73	0,01	0,28	0,72	0,00
94	0,38	0,61	0,00	0,18	0,78	0,05	0,23	0,77	0,00	0,28	0,72	0,00
95	0,12	0,88	0,00	0,05	0,94	0,01	0,06	0,94	0,00	0,06	0,94	0,00
96	0,43	0,56	0,01	0,52	0,44	0,04	0,48	0,50	0,02	0,44	0,55	0,01
97	0,29	0,71	0,00	0,06	0,91	0,03	0,10	0,89	0,00	0,15	0,85	0,00

98	0,06	0,94	0,00	0,01	0,99	0,00	0,01	0,99	0,00	0,01	0,99	0,00
99	0,03	0,97	0,00	0,00	0,99	0,01	0,00	1,00	0,00	0,00	1,00	0,00
100	0,27	0,72	0,00	0,07	0,90	0,03	0,10	0,89	0,00	0,15	0,85	0,00
101	0,43	0,56	0,01	0,49	0,47	0,05	0,45	0,53	0,02	0,43	0,56	0,01
102	0,05	0,95	0,00	0,00	0,99	0,01	0,00	1,00	0,00	0,00	1,00	0,00
103	0,09	0,91	0,00	0,00	0,98	0,02	0,00	1,00	0,00	0,00	1,00	0,00
104	0,02	0,98	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
105	0,17	0,83	0,00	0,06	0,92	0,02	0,22	0,77	0,01	0,18	0,82	0,00
106	0,17	0,83	0,00	0,07	0,92	0,01	0,06	0,93	0,00	0,08	0,92	0,00
107	0,39	0,60	0,00	0,06	0,88	0,05	0,17	0,83	0,00	0,24	0,76	0,00
108	0,21	0,79	0,00	0,01	0,97	0,02	0,04	0,96	0,00	0,06	0,94	0,00
109	0,21	0,79	0,00	0,06	0,92	0,02	0,07	0,93	0,00	0,10	0,90	0,00
110	0,04	0,96	0,00	0,00	0,99	0,00	0,00	1,00	0,00	0,00	1,00	0,00
111	0,26	0,74	0,00	0,01	0,96	0,03	0,04	0,96	0,00	0,08	0,92	0,00
112	0,81	0,12	0,07	0,61	0,08	0,31	0,86	0,05	0,09	0,77	0,05	0,18
113	0,66	0,32	0,02	0,56	0,31	0,12	0,69	0,28	0,04	0,65	0,30	0,05
114	0,22	0,78	0,00	0,14	0,84	0,02	0,13	0,86	0,00	0,15	0,85	0,00
115	0,17	0,83	0,00	0,07	0,92	0,01	0,07	0,93	0,00	0,09	0,91	0,00
116	0,13	0,87	0,00	0,04	0,95	0,01	0,04	0,96	0,00	0,05	0,95	0,00
117	0,15	0,85	0,00	0,02	0,97	0,01	0,03	0,97	0,00	0,05	0,95	0,00
118	0,65	0,12	0,22	0,55	0,01	0,44	0,56	0,00	0,44	0,61	0,00	0,39
119	0,30	0,69	0,00	0,14	0,83	0,03	0,16	0,84	0,00	0,20	0,80	0,00
120	0,01	0,99	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00
121	0,12	0,00	0,88	0,03	0,00	0,97	0,10	0,00	0,90	0,11	0,00	0,89
122	0,02	0,00	0,98	0,00	0,00	1,00	0,01	0,00	0,99	0,02	0,00	0,98
123	0,42	0,01	0,56	0,33	0,00	0,67	0,31	0,00	0,69	0,35	0,00	0,65
124	0,56	0,02	0,42	0,66	0,00	0,34	0,47	0,00	0,53	0,46	0,00	0,54
125	0,68	0,05	0,26	0,81	0,00	0,19	0,62	0,00	0,38	0,58	0,00	0,42
126	0,01	0,00	0,99	0,00	0,00	1,00	0,01	0,00	0,99	0,01	0,00	0,99
127	0,40	0,01	0,60	0,37	0,00	0,63	0,31	0,00	0,69	0,33	0,00	0,67
128	0,36	0,02	0,62	0,02	0,00	0,98	0,19	0,00	0,81	0,31	0,00	0,69
129	0,66	0,05	0,29	0,78	0,00	0,21	0,59	0,00	0,41	0,55	0,00	0,45
130	0,47	0,02	0,51	0,50	0,00	0,50	0,37	0,00	0,63	0,39	0,00	0,61
131	0,63	0,02	0,35	0,71	0,00	0,28	0,59	0,00	0,41	0,51	0,00	0,49

132	0,11	0,00	0,89	0,01	0,00	0,99	0,07	0,00	0,93	0,11	0,00	0,89
133	0,12	0,00	0,88	0,00	0,00	1,00	0,07	0,00	0,93	0,11	0,00	0,89
134	0,73	0,11	0,16	0,87	0,01	0,12	0,73	0,01	0,26	0,68	0,01	0,31
135	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00
136	0,03	0,00	0,97	0,00	0,00	1,00	0,02	0,00	0,98	0,03	0,00	0,97
137	0,01	0,00	0,99	0,00	0,00	1,00	0,00	0,00	1,00	0,01	0,00	0,99
138	0,02	0,00	0,98	0,00	0,00	1,00	0,02	0,00	0,98	0,01	0,00	0,99
139	0,63	0,22	0,15	0,36	0,06	0,58	0,60	0,01	0,39	0,66	0,02	0,32
140	0,68	0,06	0,26	0,80	0,00	0,20	0,61	0,00	0,39	0,58	0,00	0,42
141	0,01	0,00	0,99	0,00	0,00	1,00	0,01	0,00	0,99	0,01	0,00	0,99
142	0,02	0,00	0,98	0,00	0,00	1,00	0,02	0,00	0,98	0,02	0,00	0,98
143	0,17	0,00	0,83	0,05	0,00	0,95	0,13	0,00	0,87	0,15	0,00	0,85
144	0,06	0,00	0,94	0,00	0,00	1,00	0,04	0,00	0,96	0,06	0,00	0,94
145	0,36	0,01	0,64	0,18	0,00	0,82	0,24	0,00	0,76	0,30	0,00	0,70
146	0,20	0,00	0,80	0,11	0,00	0,89	0,20	0,00	0,80	0,17	0,00	0,83
147	0,01	0,00	0,99	0,00	0,00	1,00	0,01	0,00	0,99	0,01	0,00	0,99
148	0,05	0,00	0,95	0,00	0,00	1,00	0,05	0,00	0,95	0,05	0,00	0,95
149	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00
150	0,72	0,18	0,09	0,88	0,03	0,09	0,79	0,03	0,18	0,74	0,05	0,22
151	0,03	0,00	0,97	0,00	0,00	1,00	0,02	0,00	0,98	0,03	0,00	0,97
152	0,01	0,00	0,99	0,00	0,00	1,00	0,00	0,00	1,00	0,01	0,00	0,99
153	0,13	0,00	0,87	0,00	0,00	1,00	0,07	0,00	0,93	0,13	0,00	0,87
154	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00
155	0,16	0,00	0,84	0,02	0,00	0,98	0,10	0,00	0,90	0,15	0,00	0,85
156	0,02	0,00	0,98	0,00	0,00	1,00	0,02	0,00	0,98	0,02	0,00	0,98
157	0,67	0,32	0,01	0,01	0,74	0,25	0,41	0,58	0,01	0,51	0,48	0,01
158	0,69	0,23	0,08	0,87	0,04	0,09	0,79	0,04	0,17	0,74	0,07	0,19
159	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00
160	0,39	0,01	0,60	0,36	0,00	0,64	0,30	0,00	0,70	0,33	0,00	0,67
161	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00
162	0,10	0,00	0,90	0,00	0,00	1,00	0,06	0,00	0,94	0,10	0,00	0,90
163	0,26	0,00	0,73	0,07	0,00	0,93	0,17	0,00	0,83	0,23	0,00	0,77
164	0,68	0,12	0,21	0,71	0,01	0,28	0,61	0,01	0,39	0,63	0,01	0,37
165	0,12	0,00	0,88	0,00	0,00	1,00	0,06	0,00	0,94	0,12	0,00	0,88

166	0,65	0,32	0,03	0,71	0,20	0,09	0,73	0,22	0,05	0,67	0,26	0,06
167	0,03	0,00	0,97	0,00	0,00	1,00	0,02	0,00	0,98	0,03	0,00	0,97
168	0,25	0,00	0,74	0,18	0,00	0,82	0,21	0,00	0,79	0,22	0,00	0,78
169	0,65	0,09	0,26	0,63	0,01	0,37	0,54	0,00	0,45	0,58	0,00	0,42
170	0,09	0,00	0,91	0,02	0,00	0,98	0,09	0,00	0,91	0,09	0,00	0,91
171	0,68	0,05	0,27	0,80	0,00	0,19	0,62	0,00	0,37	0,57	0,00	0,43
172	0,14	0,00	0,85	0,05	0,00	0,95	0,15	0,00	0,85	0,13	0,00	0,87
173	0,03	0,00	0,97	0,00	0,00	1,00	0,01	0,00	0,99	0,03	0,00	0,97
174	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00
175	0,56	0,43	0,01	0,29	0,60	0,10	0,46	0,52	0,01	0,49	0,50	0,01
176	0,03	0,00	0,97	0,00	0,00	1,00	0,02	0,00	0,98	0,03	0,00	0,97
177	0,06	0,00	0,94	0,01	0,00	0,99	0,07	0,00	0,93	0,06	0,00	0,94
178	0,55	0,03	0,42	0,54	0,00	0,46	0,43	0,00	0,57	0,46	0,00	0,54
179	0,07	0,00	0,93	0,01	0,00	0,99	0,08	0,00	0,92	0,07	0,00	0,93
180	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	0,00	1,00