Farklı Ortalama Vektörü ve Farklı Kovaryans Matrisi Koşullarında Dört Değişkenli Lojistik Regresyon Modeli ve Diskriminant Analizine Ait Doğru Sınıflandırma Olasılıklarının Simülasyon Tekniği Yardımıyla Karşılaştırılması

Handan ÇAMDEVİREN* E.Arzu KANIK* Fikret GÜRBÜZ**

ÖZET

Bu çalışmada, farklı değişken yapısına sahip iki veri setinde, sürekli değişkenlere ait ortalama, varyans ve kovaryanslardaki değişmenin her bir populasyona doğru sınıflandırma olasılıkları üzerine etkisini araştırmak amacıyla bir simülasyon çalışması yapılmıştır. Bu çalışma sonucunda, her bir populasyona doğru sınıflandırma olasılıkları bakımından doğrusal diskriminant analizi ve lojistik regresyon analizinin benzer sonuçlar verdiği görülmüştür.

Anahtar Kelimeler: *Diskriminant analizi, lojistik regresyon modeli, sınıflandırma olasılıkları, simulasyon*

1. GİRİŞ

Son yıllarda yaygın bir şekilde kullanılan ve amaçlarından birisi sınıflandırma, diğeri ise bağımlı ve bağımsız değişkenler arasındaki ilişkilerin araştırılması olan lojistik regresyon analizi; diskriminant analizi ve regresyon modellerinde gerekli olan bir takım ön şartların yerine gelmediği durumlarda tercih edilen bir tekniktir. Başta sağlık ve davranış bilimleri olmak üzere hemen hemen bütün bilim dallarında bağımlı değişkenin var-yok, evet-hayır gibi iki seviye (binary veya dichotomous) veya ikiden çok seviye (polychotomous) ile ifade edildiği sıralayıcı veya sınıflayıcı ölçekteki verilerin bağımsız değişkenlerle olan ilişkilerini belirlemede ve sınıflamada kullanılan lojistik regresyon modeli, ilk defa 1970 yılında Cox tarafından bağımsız değişkenlerin yapılarına hiç bir sınırlama getirilmeden bu değişkenlerin, sınıflandırılmış yapıdaki bağımlı değişken ile ilişkilerini incelemede kullanılmıştır (Agresti, 1990; Qin ve Zhang, 1997; Fears ve Brown, 1986; Hosmer et. al., 1989; Alho, 1990; Everitt, 1992). Karmaşık olmayan bir modele sahip olması ve kurulan modelde yer alan bağımlı ve bağımsız değişkenlere ait ön şartların olmaması gün geçtikçe bu modelin kullanım sıklığını artırmaktadır.

* Mersin Üniversitesi Tıp Fakültesi Biyoistatistik Anabilim Dalı Yenişehir Kampüsü 33160 Mersin Email: hcamdeviren@ixir.com, Tel: 0 324 3412311 (Haberleşme adresi)
** Ankara Üniversitesi Ziraat Fakültesi Biyometri ve Genetik Anabilim Dalı

Uygulamada sık kullanılan çok değişkenli istatistik metodlarından birisi olan diskriminant analizi yukarıda da belirtildiği gibi bireylerin veya nesnelerin çeşitli özelliklerine ait ölçümlerinden yararlanarak, mevcut grupları ayırmada kullanılacak uygun fonksiyonları belirlemede ve bireylerin ait oldukları gruplara sınıflandırılmasında kullanılmaktadır. Botanikte çeşitli bitkilerin türlere göre, antropolojide bireylerin ırklara göre, arkeolojide fosillerin tarih çağlarına göre sınıflandırılması bu teknik yardımıyla kolaylıkla yapılabilir. Diskriminant fonksiyonları gerektirdikleri ön şartların farklılıklarına göre esas itibariyle doğrusal (linear), karesel (quadratic) ve yüksek dereceden terimli (polynomial) diskriminant fonksiyonları olmak üzere üç kısımda incelenir. Bunlardan doğrusal diskriminant analizi ilk olarak kesikli değişkenler içeren bir veri seti üzerinde uygulanmış fakat daha sonra, sınıflandırmada daha doğru sonuçlar elde edebilmek için bu tip veri setlerine uygulanabilecek farklı diskriminant metotları geliştirilmiştir (Schmitz et al., 1983). Doğrusal diskriminant fonksiyonundan (LDF) etkin bir şekilde yararlanabilmek için bir takım ön şartların yerine gelmiş olması gerekir. Bu ön şartlar: bağımlı değişkenin her bir seviyesinde yer alan bağımsız değişkenlerin çok değişkenli normal dağılım göstermesi ve gruplara ait kovaryans matrislerinin homojen olması şeklinde özetlenebilir. Yaygın bir şekilde kullanılan bir diğer diskriminant metodu karesel diskriminant metodudur. Gruplara ait kovaryans matrislerinin homojenliği ön şartı yerine gelmediği durumlarda, karesel diskriminant fonksiyonunu (QDF) doğrusal diskriminant fonksiyonuna tercih edilmektedir. Her iki ön şartın sağlanmadığı durumlarda ise sınıflandırmada, lojistik regresyon analizinin daha iyi sonuçlar verdiği bilinmektedir. Gerek diskriminant analizinde gerekse de lojistik regresyon analizinde sınıflandırılacak grup sayısı önceden belirlidir.

2. MATERYAL VE YÖNTEM

İki populayon mevcut iken bireylerin gruplara sınıflandırılmasında kullanılan diskiriminant analizinde son olasılıkların (posterior probability) hesaplanmasında (1) nolu eşitlik kullanılmıştır. x_0 bir bireye ait ölçüm değerleri veya bağımsız değişken değerleri olmak üzere bu bireyin birinci populasyona ait olma olasılık değeri;

$$P(\pi_1/x_0)=\frac{p_1 f_1(x_0)}{p_1 f_1(x_0)+p_2 f_2(x_0)} \qquad (1)$$

olur ve ikinci populasyona ait olma olasılık değeri (2) nolu eşitlikte tanımlanmıştır.

$$P(\pi_2/x_0)=1-P(\pi_1/x_0) \qquad (2)$$

Bu iki son olasılık değerinden büyük olan hangisi ise birey o populasyona aittir denilir ve bu sonuç x_0 bireyinin o populasyona ait olma olasılığının daha yüksek olduğunun bir göstergesidir. (1) ve (2) nolu eşitliklerde yer alan;

n_1 : 1. populasyonun örneklem genişliği
n_2 : 2. populasyonun örneklem genişliği ve
$N = n_1 + n_2$ olmak üzere

p_1 : 1. populasyona ait ön olasılık (prior probability) değeri olup; $p_1 = n_1 / N$ ve
p_2 : 2. populasyona ait ön olasılık (prior probability) değeri olup; $p_2 = n_2 / N$ eşitliği yardımıyla hesaplanır. Ayrıca her bir populasyona ait diskiriminant fonksiyon değerleri ise i =1, 2 olmak üzere

$$f_i(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_i\right)' \Sigma^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_i\right)\right]$$ eşitliği yardımıyla

hesaplanır.

Lojistik regresyon modelinde ise bir bireyin ait olduğu grubu belirlemede kullanılan son olasılık değeri;

$$\pi(\mathbf{x}) = \frac{\exp(g(\mathbf{x}))}{[1+\exp(g(\mathbf{x}))]}$$ eşitliği yardımıyla hesaplanır. Bu eşitlikte yer alan;

$g(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ olup logit fonksiyon olarak adlandırılır. Burada yer alan;

p ; bireylerden ölçülen değişken sayısını, x ; sözkonusu değişkenlere ait değerleri göstermektedir.

Bu çalışmada simulasyon tekniği yardımıyla yukarıda tanımlanan eşitliklerden hesaplanan son olasılık değerleri kullanılarak gerek diskiriminant analizi gerekse lojistik regresyon modeli ile bireylerin gruplara sınıflandırılması yapılmış ve bu sınıflandırma sonuçları kullanılarak her iki metodun doğru sınıflandırma olasılıkları hesaplanmıştır. Bu olasılığın hesaplanmasında ise gerçek gruplarına atanan birey sayısının toplam birey sayısına oranı dikkate alınmıştır.

Yapılan simulasyon çalışmasında, dört değişkenli normal dağılım gösteren iki populasyonun her birisinden, her defasında 100' er gözlem bulunan 10000 örnek alınarak, söz konusu iki populasyonun, 1) değişkenlere ait ortalama vektörlerinin değişimi, 2) değişkenlere ait varyans ve kovaryansların değişimi ve 3) dört değişkenin hepsinin sürekli ve ikisinin kesikli ikisinin ise sürekli olduğu durumlarda, Fisher' in doğrusal diskriminant fonksiyonları ve lojistik regresyon analizi yardımıyla mevcut gözlemlerin son olasılıklarından yararlanarak kendi gruplarına (populasyon) doğru sınıflandırma olasılıkları üzerine etkisi araştırılmıştır.

Söz konusu farklı koşullar daha ayrıntılı bir şekilde aşağıdaki gibi düşünülmüştür. İlk aşamada dört değişkenli (x_1, x_2, x_3, x_4) standart normal dağılımdan ortalama vektörü sıfır, kovaryans matrisi birim matris yani varyansları bir, kovaryansları sıfır olan ve her birisinde 100' er gözlem bulunan iki populasyondan 10000 defa örnek alınmıştır. Ayrıca dört değişken başlangıçta sürekli değişken olarak hesaplamalara dahil edilmiştir. Simulasyon denemesinde değişken sayısı 4 ile sınırlandırılmıştır. Ele alınan koşullardan birinde dört değişkenin hepsi sürekli yapıda düşünülmüş diğerinde ise ikisi sürekli ikisi kategorik yapıda (karışık değişkenli) ele alınmıştır. Ayrıca her bir populasyondan alınan örnek genişliğinin yeterince büyük

olmasını sağlamak amacıyla örnek genişliği 100 olarak seçilmiştir. Uygun ve güvenilir bir örnek genişliğinin tespitinde N/p oranı kullanılmıştır. Bu oran 5 veya daha büyük olduğu durumlarda çok değişkenli tekniklerden güvenilir sonuçlar elde edilebilmektedir. Ayrıca sınıflandırma tekniklerinde, doğru sınıflandırma olasılığının gerçek değerine ulaşabilmek için bu oranın büyük olması istenir (Schmitz et al. 1983, Stevens 1986). Bu çalışmada söz konusu olasılıkların güvenilir sonuçlar vermesi için N/p oranı 20 olarak alınmıştır (Çamdeviren, 2000).

Deneme yapılan farklı koşulların her birisinde, dört sürekli değişken içeren birinci populasyondan alınan örneklerde, söz konusu sürekli değişkenlere ait ortalamalar 0 ve varyanslar ise 1 olarak seçilmiştir. Ayrıca sırasıyla değişkenlerin bağımsız, düşük korelasyona sahip ve yüksek korelasyona sahip olduğu durumlarda, doğru sınıflandırma olasılıklarında ortaya çıkacak değişmeleri araştırmak amacıyla, birinci populasyondaki sürekli değişkenler arasındaki kovaryanslar birbirine eşit olmak üzere farklı kombinasyonlarda, 0.0, 0.25 ve 0.75 olarak belirlenmiştir. Aynı koşulda, dört değişkenin de sürekli yapıda olduğu ikinci populasyonda söz konusu değişkenlere ait ortalama değerler standart sapma cinsinden ve bir birine eşit olmak üzere farklı kombinasyonlarda sırasıyla 0.0, 1.0, 1.5 ve 2.0 olarak alınmıştır. Ayrıca yine ikinci populasyonda farklı kombinasyonlarda varyans-kovaryans matrisi sırasıyla birim matris, kovaryanslar sıfır varyanslar, birinci populasyondaki varyansın 3 katı, kovaryanslar sıfır varyanslar birinci populasyondaki varyansın 9 katı, kovaryanslar 0.25 varyanslar 1, kovaryanslar 0.25 varyanslar birinci populasyondaki varyansın 3 katı, kovaryanslar 0.25 varyanslar birinci populasyondaki varyansın 9 katı, kovaryanslar 0.75 varyanslar 1, kovaryanslar 0.75 varyanslar birinci populasyondaki varyansın 3 katı ve kovaryanslar 0.75 varyanslar birinci populasyondaki varyansın 9 katı olacak şekilde belirlenmiştir.

İkinci aşamada, dört değişkenden ilk ikisi ($x, ve\ x_2,$) sürekli, üçüncü değişken ($x_3,$) dört kategorili ve dördüncü değişken ise (x_4) iki kategorili kesikli bir değişken olarak düşünülmüştür. Deneme yapılan farklı koşulların her birisinde, iki sürekli ve iki kategorik değişken içeren birinci populasyondan alınan örneklerde, söz konusu sürekli değişkenlere ait ortalamalar 0 ve varyanslar ise 1 olarak seçilmiştir. Ayrıca birinci populasyondaki sürekli değişkenler arasındaki kovaryans farklı kombinasyonlarda sırasıyla 0.0, 0.25 ve 0.75 olarak belirlenmiştir. Aynı koşulda, iki sürekli ve iki kategorik değişkene sahip ikinci populasyonda, söz konusu sürekli değişkenlere ait ortalama değerler bir birine eşit olmak üzere standart sapma cinsinden farklı kombinasyonlarda sırasıyla 0.0, 1.0, 1.5 ve 2.0 olarak alınmıştır. Ayrıca yine ikinci populasyonda farklı kombinasyonlarda sürekli değişkenlere ait varyans-kovaryans matrisi sırasıyla birim matris, kovaryans 0, varyanslar birinci populasyondaki varyansın 3 katı, kovaryans 0 varyanslar birinci populasyondaki varyansın 9 katı, kovaryans 0.25 varyanslar 1, kovaryans 0.25 varyanslar birinci populasyondaki varyansın 3 katı, kovaryans 0.25 varyanslar birinci populasyondaki varyansın 9 katı, kovaryans 0.75 varyanslar 1, kovaryans 0.75 varyanslar birinci populasyondaki varyansın 3 katı ve kovaryans 0.75 varyanslar birinci populasyondaki varyansın 9 katı olacak şekilde belirlenmiştir. Denemede düşünülen her bir kombinasyon için ayrı ayrı doğrusal diskriminant analizi ve lojistik regresyon analizi uygulanmış, analiz sonucunda her bir gözleme ait tahmin edilen son olasılık değerleri yardımıyla sınıflandırma tabloları

oluşturulmuştur. 10000 deneme sonucunda elde edilen sınıflandırma tablolarında yer alan frekanslar kullanılarak her bir populasyona ortalama doğru sınıflandırma olasılıkları hesaplanmıştır.

Sonuçta farklı koşullara uygulanan diskriminant analizi ve lojistik regresyon analizi yardımıyla elde edilen doğru sınıflandırma olasılıkları karşılaştırmalı olarak incelenmiştir. Hesaplamalarda FORTRAN-90 dilinde yazılan programlardan yararlanılmıştır. Bu programlarda IMSL alt programları kullanılmıştır.

3. SONUÇLAR ve TARTIŞMA

Simulasyon çalışmasında elde edilen sonuçlar, Tablo 1., Tablo 2., Tablo 3. ve Tablo 4. 'te topluca verilmiştir.

Tablo 1.' de, her birinde 100' er gözlem bulunan dört değişkenli normal dağılıma sahip iki populasyonun farklı ortalama, varyans ve kovaryans kombinasyonlarında Fisher'in doğrusal diskriminant fonksiyonları yardımıyla sınıflandırmada, 10000 deneme sonucunda bu fonksiyonların her bir populasyona doğru sınıflandırma olasılıkları (%) yer almaktadır.

Ortalamaları 0, varyans-kovaryans matrisleri birim matris olan iki populasyondan alınan ve her birinde 100' er gözlem bulunan tesadüf örneklerinde 4 sürekli değişken kullanılarak, söz konusu gözlemlerin, doğrusal diskriminant fonksiyonları yardımıyla kendi gruplarına sınıflandırılma olasılıkları ampirik olarak (1. populasyona doğru sınıflandırma olasılığı %50.1 ve 2. populasyona doğru sınıflandırma olasılığı %49.9) bulunmuştur. Bunların, teorik olarak beklenen sonuçlara (%50) çok yakın çıktığı görülmektedir (Tablo 1). Bir başka ifadeyle ortalama vektörleri ve kovaryans matrisleri aynı olan iki populasyondan her hangi birine ait olduğu bilinen bir bireyin, tekrar bu iki populasyondan her hangi birine sınıflandırılması istenirse, bu iki populasyondan her hangi birisine ait olma veya doğru sınıflandırma olasılığının %50 olması beklenir. Geriye kalan %50' lik yanlış sınıflandırma olasılığı ise aslında yine aynı özelliklere sahip diğer populasyona sınıflandırılma olasılığını verir. Bu olasılık teorik olarak beklenen sonuçtur ve Tablo 1' de de bu durum gözlenmektedir. Bu olasılık değerleri, her bir gözleme ait son olasılık değerlerinden hesaplanmıştır. Bu durumda 4 sürekli değişken yardımıyla populasyonların sınıflandırılması amaçlandığında, her bir populasyondan 100 gözlem içeren örneklerle çalışmanın uygun ve güvenilir sonuçlar vereceği söylenebilir. Tablo 1.' in aynı sütününda yer alan diğer doğru sınıflandırma olasılıkları, iki populasyondaki değişkenlere ait varyans ve kovaryans değerleri sabit kalmak şartıyla, yukarıdan aşağıya doğru 2. populasyondaki değişken ortalamalarının birbirine eşit ve sırasıyla 1 standart sapma, 1.5 standart sapma ve 2 standart sapma olduğu durumlarda hesaplanmıştır. Bulunan sonuçlar incelendiğinde, genel olarak populasyon ortalama vektörleri arasındaki fark arttıkça, her bir populasyona doğru sınıflandırma olasılığının da arttığı görülür. Bu fark 2 standart sapma olduğunda, söz konusu olasılıklar birbirine eşit ve %97.6 olarak hesaplanmıştır. Bu koşullarda elde edilen sonuçlar teorik olarak beklenen sonuçlara çok yakın olup pratik uygulamalarda bilinen durumlardır. Ancak Tablo 1.' in birinci sütununda yer alan sonuçlar, diğer koşullarla karşılaştırmalı olarak incelemek amacıyla ele alınmıştır.

Tablo 1.' in ikinci sütununda yer alan olasılıklar, 1. populasyondaki ortalama vektörünün elemanlarının 0 ve varyans-kovaryans matrisinin birim matris olduğu, buna karşılık 2. populasyonun ortalama vektörünün elemanlarının standart sapma cinsinden sırasıyla 0, 1, 1.5 ve 2 ve yine 2. populasyonda değişken varyanslarının eşit olup 1. populasyondaki varyansların 3 katı, değişkenler arasındaki kovaryansların ise 0 olduğu durumlarda hesaplanmıştır. Bu koşulda birinci koşula nazaran, ikinci populasyondan alınan örneklerdeki 4 sürekli değişkenin varyansı, birinci populasyondaki değişkenlerin varyansının 3 katıdır. Varyanslar arasındaki bu oran, populasyon varyanslarının heterojenlik sınırındaki değeridir. Bu durumda, iki populasyonun varyansları heterojen sayılabilecek düzeyde farklı ise 1. populasyona doğru sınıflandırma olasılıkları, birinci koşula nazaran biraz daha büyük buna karşılık 2. populasyona doğru sınıflandırma olasılıkları birinci koşuldaki sonuçlara çok yakın çıkmıştır. Ancak, 1. populasyona doğru sınıflandırma olasılığında gözlenen artış, sınıflandırmanın iyi olduğu anlamında yorumlanmamalıdır. Dolayısıyla iki populasyona doğrusınıflandırma olasılıkları birbirinden farklılaşmıştır. Bu farklılık varyansların heterojenliğinin bir sonucudur. Bu sonuç varyansı küçük olan populasyona sınıflandırılan birey sayısında bir artış olacağını göstermektedir.

Tablo 1.' in üçüncü sütununda yer alan olasılıklar ise 2. populasyondaki değişkenlere ait varyansların, 1. sütunda 1. populasyondaki değişken varyanslarının 9 katı olarak belirlendiği durumda bulunan olasılıklardır. İki populasyondaki değişkenlere ait varyanslar arasındaki fark büyüdükçe, 1. populasyona doğru sınıflandırma olasılığı daha da artmış buna karşılık 2. populasyona doğru sınıflandırma olasılığı hemen hemen aynı kalmıştır. Bunun sonucu olarak iki populasyon için hesaplanan doğru sınıflandırma olasılıkları arasındaki fark daha da büyümüştür. Tablo 1.' in 4., 5. ve 6. sütunlarında, 1. ve 2. populasyonlarda, değişkenler arasındaki kovaryans birbirine eşit ve 0.25 olarak belirlenmiştir. Bunun dışında ortalama ve varyanslardaki değişim 1., 2. ve 3. koşuldaki sırayı takip etmektedir. 4. sütun incelendiğinde, her bir populasyona doğru sınıflandırma olasılıklarının birbirine eşit olduğu görülür. Bu durumda her iki populasyondaki değişkenler arasında varyans ve kovaryansların eşit olması sonucunda, her iki populasyona doğru sınıflandırma olasılıklarının aynı olacağı söylenebilir. Buna karşılık 4. sütunda yer alan olasılıklar, 1. sütunda yer alan teorik olarak beklenen olasılıklarla karşılaştırıldığında, populasyonlara ait ortalama vektörleri arasındaki fark arttıkça, teorik olarak beklenen olasılıklardan daha küçük olasılıklar elde edilmiştir. Bunun sonucu olarak, değişkenler arasında kovaryans veya korelasyon mevcut iken yani değişkenler birbirinden bağımsız olmadığı zaman doğru sınıflandırma olasılığı beklenenin altında çıkabileceği söylenebilir.

Kovaryanslar sabit kalmak şartıyla ikinci populasyonda değişkenlere ait varyanslar 1. populasyondaki varyansların 3 katı olacak şekilde belirlendiği zaman her iki populasyona doğru sınıflandırma olasılıkları arasındaki fark, kovaryansların sıfır olduğu 2. sütundaki olasılıklar arasındaki farktan biraz daha büyük çıkmıştır. Ayrıca populasyon varyansları değiştikçe söz konusu iki olasılık arasında fark oluşmuştur. 6. sütun incelendiğinde varyanslar arasındaki fark arttıkça iki olasılık değeri arasındaki farkında arttığı söylenebilir. Tablo 1.' in 7., 8. ve 9. sütunları incelendiğinde her iki populasyonda da değişkenler arasındaki kovaryans değerinin birbirine eşit ve 0.75

olarak belirlendiği görülür. Bu koşullarda elde edilen doğru sınıflandırma olasılıklarına ilişkin sonuçların ise yukarıda açıklanan diğer koşulların sonuçlarına benzer olduğu söylenebilir.

Tablo 1. Her birinde 100' er gözlem bulunan dört değişkenli normal dağılıma sahip iki populasyonun farklı ortalama, varyans ve kovaryans kombinasyonlarında doğrusal diskriminant fonksiyonu yardımıyla sınıflandırmada 10000 deneme sonucunda bu fonksiyonun her bir gruba doğru sınıflandırma olasılıkları (%)*

1. populasyondaki 4 sürekli değişkene ait ortalama ve varyanslar			$\mu_{i(1)} = (0, 0, 0, 0)$ ve $\sigma^2_{i(1)} = (1, 1, 1, 1)$								
1. ve 2. Populasyondaki kovaryanslar (σ_{ij})			0.0			0.25			0.75		
2. populastondaki varyanslar ($\sigma^2_{i(2)}$)			1	3	9	1	3	9	1	3	9
2. populasyon ortalamaları	$\mu_{i(2)} = (0, 0, 0, 0)$	Pop-1	50.1	53.7	59.3	50.0	53.7	59.3	50.1	53.7	59.4
		Pop-2	**49.9**	**47.9**	**47.0**	**50.0**	**47.9**	**46.9**	**50.0**	**48.0**	**47.0**
	$\mu_{i(2)} = (1, 1, 1, 1)$	Pop-1	83.7	95.6	99.8	77.0	90.2	98.8	70.4	82.8	94.9
		Pop-2	**83.7**	**83.6**	**83.5**	**77.0**	**76.9**	**76.8**	**70.4**	**70.3**	**70.0**
	$\mu_{i(2)} = (1.5, 1.5, 1.5, 1.5)$	Pop-1	93.1	99.5	100	86.8	97.3	100	79.3	92.3	99.3
		Pop-2	**93.0**	**93.0**	**92.9**	**86.8**	**86.8**	**86.6**	**79.4**	**79.2**	**79.1**
	$\mu_{i(2)} = (2, 2, 2, 2)$	Pop-1	97.6	100	100	93.2	99.5	100	86.3	97.1	99.9
		Pop-2	**97.6**	**97.6**	**97.5**	**93.2**	**93.2**	**93.1**	**86.3**	**86.1**	**86.0**

*Açık renkte yazılmış olasılık 1. populasyona, koyu yazılmış olasılık ise 2. populasyona doğru sınıflandırma olasılığını göstermektedir.

Tablo 2.' de, her birinde 100' er gözlem bulunan dört değişkenli normal dağılıma sahip iki populasyonun farklı ortalama, varyans ve kovaryans kombinasyonlarında lojistik regresyon modeli yardımıyla sınıflandırmada, 10000 deneme sonucunda her bir populasyona doğru sınıflandırma olasılıkları (%) yer almaktadır. Tablo 2. genel olarak incelendiğinde, lojistik regresyon modeli yardımıyla her bir populasyona doğru sınıflandırma olasılığının, doğrusal diskriminant fonksiyonları yardımıyla elde edilen olasılıklara çok yakın çıktığı görülür. Ancak özellikle her iki populasyonun varyans ve kovaryanslarının eşit olduğu koşullarda, doğrusal diskriminant fonksiyonları yardımıyla elde edilen sınıflandırma olasılıkları, lojistik regresyon modelinden elde edilen olasılıklara göre teorik sonuçlara biraz daha yakın çıkmıştır. Bunun dışında, lojistik regresyon modelinden elde edilen doğru sınıflandırma olasılıkları üzerine ortalama, varyans ve kovaryanslardaki değişmelerin etkisi, aynı koşullarda uygulanan doğrusal diskriminant analizindeki gibidir. Bulunan bu sonuçlara göre, hepsi sürekli değişkenlerden oluşan iki çok değişkenli normal dağılım gösteren populasyondaki bireylerin sınıflandırılmasında, doğrusal diskriminant analizinden elde edilecek doğru sınıflandırma olasılıklarının, lojistik regresyon modeli yardımıyla elde edilen doğru sınıflandırma olasılıklarına göre teorik olarak beklenen olasılıklara biraz daha yakın çıkacağı söylenebilir.

Tablo 2. Her birinde 100' er gözlem bulunan dört değişkenli normal dağılıma sahip iki populasyonun farklı ortalama, varyans ve kovaryans kombinasyonlarında lojistik regresyon modeli yardımıyla sınıflandırmada 10000 deneme sonucunda bu fonksiyonun her bir gruba doğru sınıflandırma olasılıkları (%)

1. populasyondaki 4 sürekli değişkene ait ortalama ve varyanslar			$\mu_{i(1)} = (0, 0, 0, 0)$ ve $\sigma^2_{i(1)} = (1, 1, 1, 1)$								
1. ve 2. Populasyondaki kovaryanslar (σ_{ij})			0.0			0.25			0.75		
2. populastondaki varyanslar ($\sigma^2_{i(2)}$)			1	3	9	1	3	9	1	3	9
2. populasyon ortalamaları	$\mu_{i(2)} = (0, 0, 0, 0)$	Pop-1	52.1	54.2	59.6	54.4	57.4	63.1	56.1	57.6	62.0
		Pop-2	**51.1**	**48.0**	**53.7**	**54.4**	**53.4**	**53.1**	**56.6**	**54.1**	**54.6**
	$\mu_{i(2)} = (1, 1, 1, 1)$	Pop-1	83.2	92.6	97.0	78.2	87.7	93.6	71.1	81.4	89.5
		Pop-2	**83.1**	**88.9**	**91.8**	**78.7**	**82.6**	**83.2**	**69.9**	**74.7**	**75.2**
	$\mu_{i(2)} = (1.5, 1.5, 1.5, 1.5)$	Pop-1	94.2	98.2	98.8	87.2	95.6	97.8	80.2	89.0	94.5
		Pop-2	**94.8**	**98.1**	**98.2**	**86.8**	**92.8**	**93.7**	**78.8**	**84.6**	**86.8**
	$\mu_{i(2)} = (2, 2, 2, 2)$	Pop-1	97.4	99.5	99.8	94.2	97.6	99.0	88.1	92.7	98.3
		Pop-2	**97.4**	**98.5**	**99.5**	**95.1**	**97.1**	**98.7**	**88.0**	**90.2**	**95.0**

Tablo 3.' te, her birinde 100' er gözlem bulunan ikisi sürekli ikisi kategorik yapıdaki 4 değişkenden oluşan iki populasyonun farklı ortalama, varyans ve kovaryans kombinasyonlarında Fisher'in doğrusal diskriminant fonksiyonları yardımıyla sınıflandırmada, 10000 deneme sonucunda bu fonksiyonların her bir populasyona doğru sınıflandırma olasılıkları (%) yer almaktadır. Ortalamaları 0, varyans-kovaryans matrisleri birim matris olan iki populasyondan alınan ve her birinde 100' er gözlem bulunan tesadüf örneklerinde, söz konusu gözlemlerin, doğrusal diskriminant fonksiyonları yardımıyla kendi gruplarına sınıflandırılma olasılıkları (1. populasyona doğru sınıflandırma olasılığı %93.2 ve 2. populasyona doğru sınıflandırma olasılığı %100) incelendiği zaman, bulunan sonuçların teorik olarak beklenen sonuçlardan oldukça farklı çıktığı görülür (Tablo 3). Populasyonların ortalamaları arasındaki fark büyüdükçe söz konusu olasılıklarda da biraz artış gözlenmiştir. Ayrıca Tablo 1.' de populasyonlara ait varyans ve kovaryans değerlerinin aynı olduğu 1., 4. ve 7. sütunlardaki olasılıklar incelendiğinde her iki populasyona doğru sınıflandırma olasılığı aynı bulunurken, değişkenlerden ikisinin kategorik yapıda olduğu bu koşulda (Tablo 3.) aynı numaraya sahip sütunlarda 1. ve 2. populasyona ait doğru sınıflandırma olasılıkları birbirinden farklı çıkmıştır. Populasyonların varyansları birbirinden farklılaştıkça, her bir populasyona doğru sınıflandırma olasılığı biraz daha birbirine yaklaşmış ve varyansı küçük olan populasyona doğru sınıflandırma olasılığı, varyansı büyük olan populasyona doğru sınıflandırma olasılığından biraz daha küçük çıkmıştır. Ancak, Tablo 1.' den de görüleceği üzere populasyonlardaki 4 değişkenin de sürekli olduğu durumda, varyansı küçük olan populasyona doğru sınıflandırma olasılığının daha büyük olmaktaydı.

Tablo 3. Her birinde 100' er gözlem bulunan ikisi sürekli ikisi kategorik olmak üzere dört değişken içeren iki populasyonun farklı ortalama, varyans ve kovaryans kombinasyonlarında doğrusal diskriminant fonksiyonu yardımıyla sınıflandırmada 10000 deneme sonucunda bu fonksiyonun her bir gruba doğru sınıflandırma olasılıkları (%)

1. populasyondaki 2 sürekli değişkene ait ortalama ve varyanslar			$\mu_{i(1)} = (0, 0)$ ve $\sigma^2_{i(1)} = (1, 1)$								
1. ve 2. Populasyondaki kovaryanslar (σ_{ij})			0.0			0.25			0.75		
2. populastondaki varyanslar ($\sigma^2_{i(2)}$)			1	3	9	1	3	9	1	3	9
2. populasyon ortalamaları	$\mu_{i(2)} = (0, 0)$	Pop-1	93.2	93.3	93.3	92.5	93.2	93.3	92.6	92.8	91.6
		Pop-2	**100**	**100**	**100**	**100**	**100**	**100**	**99.7**	**99.9**	**100**
	$\mu_{i(2)} = (1, 1)$	Pop-1	91.4	93.0	93.4	92.0	92.7	93.3	90.7	91.0	91.6
		Pop-2	**99.9**	**99.4**	**98.8**	**100**	**99.8**	**99.3**	**100**	**99.9**	**99.7**
	$\mu_{i(2)} = (1.5, 1.5)$	Pop-1	92.6	95.1	96.1	91.5	93.6	94.6	90.4	93.0	95.0
		Pop-2	**99.8**	**99.0**	**98.3**	**99.9**	**99.2**	**98.5**	**100**	**99.6**	**99.3**
	$\mu_{i(2)} = (2, 2)$	Pop-1	95.0	98.0	99.4	92.7	99.2	98.1	91.7	99.8	98.9
		Pop-2	**99.8**	**99.1**	**98.7**	**99.8**	**99.0**	**98.5**	**99.9**	**89.8**	**99.3**

Ayrıca populasyonlardaki sürekli değişkenler arasında bir korelasyon mevcut olduğunda ise her bir populasyona doğru sınıflandırma olasılıkları arasındaki fark biraz daha büyümüştür. 4 değişkeninde sürekli olduğu durumda ortalama, varyans ve kovaryanslardaki değişmelere bağlı olarak her bir gruba doğru sınıflandırma olasılıkları yaklaşık olarak %50 ile %100 arasında hesaplanırken, değişkenlerden 2' si sürekli 2' si kategorik olduğu zaman söz konusu olasılıklar yaklaşık olarak %90 ile %100 arasında bulunmuştur.

Tablo 4.' te, her birinde 100' er gözlem bulunan ikisi sürekli ikisi kategorik yapıdaki 4 değişkenden oluşan iki populasyonun farklı ortalama, varyans ve kovaryans kombinasyonlarında lojistik regresyon modeli yardımıyla sınıflandırmada, 10000 deneme sonucunda her bir populasyona doğru sınıflandırma olasılıkları (%) yer almaktadır. Tablo 4. genel olarak incelendiğinde, lojistik regresyon modeli yardımıyla her bir populasyona doğru sınıflandırma olasılıklarının, Tablo 3.' teki doğrusal diskriminant fonksiyonları yardımıyla elde edilen olasılıklara çok yakın olduğu görülür. Yani lojistik regresyon modelinden elde edilen doğru sınıflandırma olasılıkları üzerine ortalama, varyans ve kovaryanslardaki değişmelerin etkisi, aynı koşullarda uygulanan doğrusal diskriminant analizindeki gibidir. Bunlara ilaveten populasyon varyansları farklılaştıkça, lojistik regresyon modeli yardımıyla elde edilen her bir populasyona doğru sınıflandırma olasılıkları, aynı koşullardaki doğrusal diskriminant fonksiyonlarından elde edilen olasılıklara göre biraz daha birbirine yakın çıkmıştır. 4 değişkeninde sürekli olduğu durumda ortalama, varyans ve kovaryanslardaki değişmelere bağlı olarak her bir populasyona doğru sınıflandırma olasılıkları yaklaşık

olarak %50 ile %100 arasında hesaplanırken, değişkenlerden 2' si sürekli 2' si kategorik olduğu zaman bu olasılıklar yaklaşık olarak %90 ile %100 arasında bulunmuştur.

Tablo 4. Her birinde 100' er gözlem bulunan ikisi sürekli ikisi kategorik olmak üzere dört değişken içeren iki populasyonun farklı ortalama, varyans ve kovaryans kombinasyonlarında lojistik regresyon modeli yardımıyla sınıflandırmada 10000 deneme sonucunda bu fonksiyonun her bir gruba doğru sınıflandırma olasılıkları (%)

1. populasyondaki 2 sürekli değişkene ait ortalama ve varyanslar			$\mu_{i(1)} = (0, 0)$ ve $\sigma^2_{i(1)} = (1, 1)$								
1. ve 2. Populasyondaki kovaryanslar (σ_{ij})			0.0			0.25			0.75		
2. populastondaki varyanslar ($\sigma^2_{i(2)}$)			1	3	9	1	3	9	1	3	9
2. populasyon ortalamaları	$\mu_{i(2)} = (0, 0)$	Pop-1	91.4	92.9	92.2	93.8	94.8	93.1	94.6	93.9	93.7
		Pop-2	**100**	**100**	**100**	**99.8**	**100**	**100**	**99.3**	**100**	**100**
	$\mu_{i(2)} = (1, 1)$	Pop-1	93.7	94.4	95.0	94.4	93.2	94.3	93.1	93.5	96.0
		Pop-2	**99.6**	**98.5**	**99.3**	**99.6**	**99.0**	**99.8**	**99.7**	**99.1**	**98.3**
	$\mu_{i(2)} = (1.5, 1.5)$	Pop-1	97.9	96.8	97.3	94.7	95.0	96.0	94.7	96.6	98.2
		Pop-2	**98.8**	**98.3**	**98.0**	**99.3**	**98.6**	**98.3**	**99.5**	**98.8**	**98.8**
	$\mu_{i(2)} = (2, 2)$	Pop-1	97.9	98.3	98.3	96.8	98.1	97.8	96.2	98.7	99.0
		Pop-2	**99.2**	**99.4**	**98.5**	**98.7**	**98.7**	**98.3**	**99.3**	**99.2**	**99.5**

Bulunan bu sonuçlara göre iki populasyona ait veri setinde yer alan 4 değişkenin sürekli yapıda veya 2' sinin sürekli 2' sinin kategorik yapıda olması durumunda Fisher' in doğrusal diskriminant fonksiyonları ve lojistik regresyon modeli yardımıyla, her bir populasyona doğru sınıflandırma olasılıklarının birbirine benzer çıkacağı söylenebilir. Ancak değişkenlerden 2' sinin kategorik olması sonucunda, her iki teknikte de doğru sınıflandırma olasılıkları, teorik olarak beklenen sonuçlardan oldukça farklı çıkmaktadır. Bu durumda, sadece doğru sınıflandırma olasılıklarına bakarak doğrusal diskriminant fonksiyonları veya lojistik regresyon modelinin kullanımı yönünde tercih yapmanın yanıltıcı sonuçlar vereceği söylenebilir. Özellikle sürekli ve kategorik yapıda değişken içeren karışık veri setlerinde sınıflandırma amacıyla uygun tekniğin seçiminde tahmin edilen katsayıların yanlılık ve tutarlılık gibi durumları da göz önüne alınmalıdır. Kesikli ve sürekli değişkenlerden oluşan karışık yapıdaki veri setleri için Fisher' in doğrusal diskriminant fonksiyonu, karesel diskriminant fonksiyonu, yüksek dereceden terimler içeren doğrusal diskriminant fonksiyonu, lojistik diskriminant fonksiyonu ve konum modeli (location model) kullanıldığında, çok değişkenli normallik ve kovaryans matrislerinin homojenliği ön şartları yerine geldiği zaman doğrusal diskriminant fonksiyonun en iyi sonuçlar verdiği, modelde interaksiyon terimlerinin yer aldığı ve kovaryans matrislerinin homojen olmadığı durumlarda ise lojistik diskriminant fonksiyonunun hata oranının diğer metotlardan daha düşük olduğu görülmüştür. Bazı çalışmalarda, çok sayıda değişken içeren veri setlerinde, var-yok gibi kategorize edilerek ifade edilen az sayıda değişken mevcut olduğu zaman, çok değişkenli normallik varsayımının bozulmadığı görülmüştür. Bu gibi durumlarda, ilk birkaç diskriminant

fonksiyonun değerlerine ilişkin normallik testi yapılmış ve sonuçta dağılımlarının normale çok yakın çıktığı gözlenmiştir (Knoke, 1982; Schmitz et al., 1983; Johnson ve Wichern, 1982). Karışık yapıda değişken içeren veri setlerinde, orijini bilinmeyen bir bireyin ait olduğu populasyonu belirlemede veya populasyonları birbirinden en iyi bir şekilde ayırmada, lojistik ayırıcı katsayıların, diğer metotlara nazaran daha güçlü sonuçlar verdiği söylenebilir. Bu katsayıların, Newton-Raphson algoritması yardımıyla en çok olabilirlik tahmin edicileri elde edilerek veya son olasılıkların lojistik şekli bulunarak sınıflandırma işlemi yapılmaktadır. Bunun dışında, modeldeki değişkenler çok değişkenli normal veya çok değişkenli bağımsız binomiyal dağılım gösterdiği durumlarda, kovaryans matrisleri eşit iken veya model ikinci ve daha yüksek dereceden terimleri içerdiği durumlarda da bu katsayılar, ayırma amacıyla kullanılabilir. (Anderson, 1972). Yapılan bir diğer simülasyon çalışmasında, doğrusal diskriminant fonksiyonu, karışık değişkenli konum modeli (mixed location model) ve lojistik regresyon modeli karışık yapıda değişken içeren veri setlerine sınuflandırma yapmak amacıyla uygulanmış ve bu tekniklerin hata oranları karşılaştırılmıştır. Tek değişkenli odds oranlarının değerinin 2' yi geçtiği durumlarda, modeldeki kesikli değişkenlere ait katsayıların tahmininde diskriminant fonksiyonu yanlı sonuçlar vermiştir. Buna karşılık odds oranlarının değeri, 2' nin altında olduğu durumlarda ise bu üç teknik yardımıyla kesikli değişkenlere ait katsayı tahminlerinin yanlılık durumları arasında bir fark bulunmamıştır. Ayrıca sürekli değişkenlere ait katsayı tahminlerinde de bu üç metodun nispi yanlılıkları arasında önemli bir farklılığın olmadığı, bunlara ilaveten karışık değişkenli konum modeli ile katsayıların tahmin edilen örnekleme varyansları diğer iki tahmin metodundan daha küçük veya en fazla eşit olduğu gösterilmiştir (Hosmer et al., 1983).

KAYNAKLAR

AGRESTİ, A. (1990), *Categorical data analysis,* John Wiley & Sons, 558 p., New York-USA.

ALHO, J..M. (1990), *"Logistic regression in capture-recapture models"*, Biometrics, 46, 623-635.

ANDERSON, J.A. (1972), *"Separate sample logistic discrimination"*, Biometrika, 59(1), 19-35.

ÇAMDEVIREN, H. (2000), *Lojistik regresyon ve Diskriminant analizi*, A.Ü. Fen Bilm. Enst. Doktora Tezi (Yayınlanmamış), 185 s.

EVERITT, B.S. (1992), *The analysis of contingency tables*, Chapman & Hall, secon Edition, 164 p., London-UK.

FEARS, T.R. and BROWN, C.C. (1986), "*Logistic regression methods for redrospective case-control studies using complex sampling procedures*", Biometrics, 42, 955-960.

HOSMER, T., HOSMER, D., and FISHER, L. (1983), *"A comparison of the maximum likelihood and discriminant function estimators of the coefficients of the logistic regression model for mixed continuous and discrede variables"*, Commun. Statist.-Simula. Computa., 12(1), 23-43.

HOSMER, D.W., JAVANOVIC, B. and LEMESHOW, S. (1989), "*Best subsets logistic regression*", Biometrics, 45, 1265-1270.

JOHNSON, R.A. and WICHERN, D.W. (1982), *Applied multivariate statistical analysis,* Prentice-Hall, INC., Englewood Cliffs, 594 p., New Jersey-USA.

KNOKE, J.D. (1982), "*Discriminant analysis with discrede and continuous variables*", Biometrics, 38, 191-200.

QIN, J. and ZHANG, B. (1997), "*A goodness-of-fit test for logistic regression models based on case-control data*", Biometrika, 84(3), 609-618.

SCHMITZ, P.I.M., HABBEMA, J.D.F., HERMANS, J. and RAATGEVER, J.W. (1983), "*Comparative performance of four discriminant analysis methods for mixtures of continuous and discrede variables*", Commun. Statist.-Simula. Computa., 12(6), 727-751.

STEVENS, J. (1986), *Applied multivariate statistics for the social sciences*, Hillsdale, 509 p., New Jersey-USA.

Probabilities of Correctly Classifying Belong to Logistic Regression Model and Discriminant Analysis with four Variables Compares with Simulation Technique in Condition that Different Mean Vector and Different Covariance Matrix

ABSTRACT

In this study, to show that the effect of various combination mean vectors, variance and covariance structure of continuous variables on the true classification proportions each population was made a simulation study. Simulation results show that both linear discriminant analysis and logistic regression analysis same.

Key Words: *Diskriminant analysis, logistic regression model, classification probabilities, simulation.*