Volume: 09 Number: 02 Category: 05 Page: 54-64 ISSN No: 1303-6319 Cilt: 09 Sayı: 02 Kategori: 05

LOJİSTİK REGRESYON VE DOĞRUSAL DİSKRİMİNANT ANALİZLERİNDE KULLANILAN BAZI İNDEKSLERİN KARŞILAŞTIRILMASI

Atilla GÖKTAŞ* **Barış KESKİN**** **Selen ÇAKMAKYAPAN*****

ÖZET

Lojistik regresyon ve doğrusal diskriminant analizi, bireylerin ya da gözlemlerin sınıflandırılmasında yaygın olarak kullanılan iki yöntemdir. Bu analizlerin aynı amaçla kullanılabilmesi, hangisinin daha iyi sonuçlar elde ettiği sorusunu akla getirmektedir. Bu konu üzerine çalışmalar yapılmış ve bu iki analiz karşılaştırılmıştır. Diskriminant analizi için gerekli varsayımların, lojistik regresyon için gerekli olmaması bu iki analizin farklı koşullarda tercih edilebilirliğini de değiştirmiştir. Bu çalışmada ise, değişkenlerin normal dağılım varsayımını sağlamadığı durumda bu iki analizin ve analizleri değerlendirmede kullanılan indekslerin karşılaştırılması amacıyla bir benzetim çalışması gerçekleştirilmiştir. Verilerin çok değişkenli normal dağılım varsayımını sağlamayıp, farklı dağılımlar gösterdiği durumda lojistik regresyon analizinin diskriminant analizine göre genel olarak daha iyi sonuç verdiği görülmüştür. Örneklem büyüklüğü arttıkça iki analizden elde edilen sonuçlar arasındaki farklılık azalmıştır. Örneklem büyüklüğü ne olursa olsun, tüm indeks ölçütlerine göre lojistik regresyon analizinin doğrusal diskriminant analizine göre sınıflandırmada daha başarılı olduğu görülmüştür. Cohen'in Kappa katsayısı yeni bir indeks olarak kullanılmıştır. Ayrıca, hangi modelin iyi olduğu bilindiğinde elde edilen indeks değerleri sayesinde, indekslerin iki yöntemin tahmindeki doğruluklarını karşılaştırmadaki başarısı değerlendirmeye alınmıştır.

Anahtar Kelimeler: Diskriminant analizi, Cohen'in Kappası, Lojistik regresyon.

1. GİRİŞ

Lojistik regresyon (LR) ve doğrusal diskriminant analizi (DA), bireylerin ya da gözlemlerin sınıflandırılmasında yaygın olarak kullanılan iki yöntemdir. Aynı amaçla kullanılmalarına rağmen, aralarındaki en önemli farklılık varsayımlarından ve parametre tahminlerinde kullandıkları yöntemlerden kaynaklanmaktadır. DA'da açıklayıcı değişkenlerin normal dağılımdan gelmiş olması ve grupların aynı kovaryans matrisine sahip olması gerekirken, LR'de bu varsayımlara gerek yoktur.

Literatürde bu iki analizin karşılaştırılmasına yönelik çalışmalar yapılmış ve bu çalışmalarda varsayımlar sağladığında DA'nın daha iyi sonuçlar verebileceği, ancak diğer tüm durumlarda LR'nin daha başarılı olacağı ifade edilmiştir (Press vd., 1978).

Bu analizlerin başarısını değerlendirmede ya da iki analizin karşılaştırılması için önerilen ölçütler vardır. Bu ölçütlerden bazıları B, Q ve C indeksleridir. Bu çalışmada, LR ve DA ile ilgili elde edilen bilgiler ışığında, bu indekslerin ve ayrıca Cohen'in

*Yrd. Doç. Dr., Muğla Sıtkı Koçman Üniversitesi, Fen Fakültesi, İstatistik Bölümü, Muğla, e-posta: gatilla@mu.edu.tr
**Arş. Gör., Muğla Sıtkı Koçman Üniversitesi, Fen Fakültesi, İstatistik Bölümü, Muğla, e-posta: bariskeskin@mu.edu.tr
***Arş. Gör., Muğla Sıtkı Koçman Üniversitesi, Fen Fakültesi, İstatistik Bölümü, Muğla, e-posta: selencak@mu.edu.tr

Kappa katsayısının bu bilgiye ne kadar paralel sonuçlar verdiğine bakılarak aralarında bir değerlendirme yapılmıştır. Çalışmada ilk olarak, LR ve DA'dan genel hatlarıyla kısaca bahsedilmiş, iki analize ilişkin tahmin yöntemleri karşılaştırılmış ve kullanılan indekslere ilişkin bilgiler verilmiştir. Sonrasında, yapılan benzetim çalışması anlatılmış ve elde edilen sonuçlar değerlendirilmiştir.

2. LOJİSTİK REGRESYON VE DOĞRUSAL DİSKRİMİNANT ANALİZİ

2.1 Lojistik Regresyon

Lojistik regresyon analizi; çok değişkenli verilerin sınıflandırılmasında, gerek bu verilere uygulanabilecek çeşitli istatistiksel yöntemler için gerekli bir ön analiz olarak, gerekse başlı başına bir analiz olarak sıkça kullanılan bir analizdir. Bu analizde bağımlı değişken kategorik, bağımsız değişkenler ise kategorik değişken, sürekli değişken veya kategorik ve sürekli değişkenlerin bir karması olabilir. Lojistik regresyon analizi normallik, ortak kovaryansa sahip olma, süreklilik varsayımı gibi ön koşullara gerek duymadığından, bu varsayımları gerektiren yöntemlere alternatif olarak geliştirilmiştir.

Lojistik regresyon analizinde kullanılan model, bağımlı değişkenin 0, 1 gibi iki ya da ikiden çok düzey içeren kesikli değişken olması durumunda normallik varsayımı gerektirmemesi nedeniyle kullanım rahatlığı sağlamaktadır (Gürcan, 1998). Yorum kolaylılığı ve kestirim güçlülüğü nedeniyle yaygın olarak kullanılmaktadır.

Bu analizde çok değişkenli istatistiksel verilerin sınıflandırılması, bağımlı değişkenin hesaplanan olasılık tahminleri yardımıyla yapılır. y bağımlı değişkenin sadece 0 ve 1 gibi iki değer aldığı durumda, $P(y_i = 1)$ *i.* gözlemin 1 değeri alması olasılığı, $P(y_i = 0)$ *i.* gözlemin 0 değeri alması olasılığı olmak üzere *i.* gözlemin beklenen değeri eşitlik 1 ile verilir.

$$E(y_i) = 1 \times P(y_i = 1) + 0 \times P(y_i = 0) = P(y_i = 1) \tag{1}$$

Analizde p adet bağımsız değişkenli n adet gözlem için kullanılan lojistik regresyon modeli ise β_k *k.* bağımsız değişkene ait katsayıyı göstermek üzere, gözlemin grup 1'de olma olasılığı ya da başka bir ifadeyle bağımlı değişkenin 1 değerini alması olasılığı eşitlik 2 ile verilmiştir.

$$P(y_i = 1) = \frac{1}{1+e^{-\left(\sum_{k=0}^{p} \beta_k x_{ik}\right)}} \qquad i = 1,2,\dots,n \tag{2}$$

Modelde yer alan parametre tahminleri, en çok olabilirlik, yeniden ağırlıklandırılmış en küçük kareler ve minimum lojit ki-kare yöntemleri ile hesaplanabilir (Gürcan, 1998). Bu çalışma kapsamında, en çok olabilirlik tahmin edicileri dikkate alınmıştır.

Bu analizin en önemli avantajı, bağımlı değişkeni etkileyebilecek önemli değişkenlerin belirlenebilmesi ve belirli karakteristiklere göre bir gözlemin bağımlı değişken kategorilerine düşmesi olasılıklarının hesaplanabilmesidir.

2.2 Doğrusal Diskriminant Analizi

DA, kelime ve genel anlamı itibariyle noktaları, bireyleri ya da gözlemleri ayırma ile ilgili bir analizdir. Bu analiz bireylere ait *p* tane özellikten yararlanarak ait oldukları grupları belirlemede veya mevcut grupları birbirinden ayıracak en iyi fonksiyonu bulmada kullanılan çok değişkenli istatistik tekniklerinden biridir. Hatalı sınıflandırma olasılığını en aza indirgeyerek, gözlemleri ait oldukları gruplara ayırma amacına yönelik olarak kullanılan istatistiksel bir karar verme yöntemi olarak da tanımlanmaktadır. Birden çok bağımsız değişkene göre gözlemleri gruplara atamada yaygın olarak kullanılan ve kabul görmüş olan bu yöntem, değişkenlerin çok değişkenli normal dağılım göstermesi, aralarında önemli ilişkilerin olmaması ve grup varyanslarının homojenliği varsayımlarına dayanır. İlk olarak Fisher tarafından gözlemleri iki farklı gruba atamak için önerilen yöntem, sonrasında daha fazla grup için de geliştirilmiş, genelleştirilmiştir. Yöntemin geliştirilmesi ile DA'nın kullanımı giderek yaygınlaşmıştır.

Bu analizde gözlemlerin sınıflandırılması işlemi, elde edilen diskriminant fonksiyonları ve kesim noktaları yardımıyla yapılır. Yöntemin başarısı, veri kümesindeki kaç tane gözlemi doğru biçimde sınıflandırdığı (kendi grubuna atadığı) ve bazı indeks ölçütleri ile ölçülmektedir.

DA'nın en basit hali iki grup söz konusu olduğunda elde edilir. Gruplara ait gözlemlerin birbirinden ayrılmasında ya da yeni bir gözlemin bir gruba atanmasında, bu grupların merkezleri arasından geçen doğrusal bir ayırma fonksiyonu kullanılır. DA modeli, LR'nin aksine, iki ayrı grubun sırasıyla μ_1 ve μ_2 ortalamaları ve aynı Σ kovaryans matrisi ile normal dağıldıklarını varsayar. Bu varsayım altında, bir bireyin grup 1'e atanması olasılığı ise aşağıdaki gibi ifade edilebilir.

$$P(y_i = 1) = \frac{1}{1+e^{-\left(\Sigma_{k=0}^{p} \beta_k x_{ik}\right)}} \qquad i = 1,2,\dots,n \tag{3}$$

Burada dikkat çeken nokta, LR ve DA modellerinin aynı fonksiyonel formda ifade edilebildikleridir. Aralarındaki fark ise katsayıların tahmininden kaynaklanmaktadır.

DA'da, π_1 ve π_2 bir gözlemin sırasıyla grup bir ve grup ikide olma önsel olasılıklarını temsil etmek üzere, β_0 ve $\beta' = (\beta_1, \beta_2, \dots, \beta_p)$ aşağıdaki gibi hesaplanır.

$$\beta_0 = -\log\frac{\pi_2}{\pi_1} + \frac{1}{2}(\mu_1 + \mu_0)' \Sigma^{-1}(\mu_1 - \mu_0)$$

$$\beta = (\mu_1 - \mu_0)' \Sigma^{-1}$$

Eğer $\pi_1, \pi_2, \mu_1, \mu_2, \Sigma$ parametreleri bilinmiyorsa örneklemden elde edilen tahmin değerleri kullanılır. n_1 grup 1'deki gözlem sayısı, n_2 grup 2'deki gözlem sayısı ve $n = n_1 + n_2$ toplam gözlem sayısı olmak üzere tahmin değerleri aşağıdaki eşitlikler ile elde edilir.

$$\hat{\pi}_1 = n_1/n, \qquad \hat{\pi}_2 = n_2/n$$

$$\bar{x}_{11} = \frac{1}{n_1}\sum_{y_i=0} x_{i1},\ \bar{x}_{12} = \frac{1}{n_1}\sum_{y_i=0} x_{i2},\ ...,\ \bar{x}_{1p} = \frac{1}{n_1}\sum_{y_i=0} x_{ip}$$

$$\bar{x}_{21} = \frac{1}{n_1}\sum_{y_i=0} x_{i1},\ \bar{x}_{22} = \frac{1}{n_1}\sum_{y_i=0} x_{i2},\ ...,\ \bar{x}_{2p} = \frac{1}{n_1}\sum_{y_i=0} x_{ip}$$

$$\hat{\mu}_1' = \bar{x}_1 = (\bar{x}_{11}, \bar{x}_{12}, ..., \bar{x}_{1p}), \qquad \hat{\mu}_2' = \bar{x}_2 = (\bar{x}_{21}, \bar{x}_{22}, ..., \bar{x}_{2p})$$

$$\hat{\Sigma} = \frac{(n_1 - 1)\Sigma_1 + (n_2 - 1)\Sigma_2}{n_1 + n_2 - 2}$$

2.3 Lojistik Regresyon ve Diskriminant Analizinde Kullanılan Tahmin Yöntemlerinin Karşılaştırılması

- Aynı kovaryans matrisiyle normal dağılıma sahip olma varsayımı sağlanmadığı zaman DA yöntemi ile elde edilen eğim katsayısı tahminleri tutarlı olmayabilir. Bu nedenle, örneklem büyük olsa dahi bu yöntemle elde edilen tahminlerin iyi tahminler olacağı ya da iyi bir uyum elde edileceği konusunda bir garanti yoktur. Pratikte bağımsız değişkenlerin sürekli olmadıkları, iki veya daha fazla kategoriye sahip kesikli değişken oldukları durumlarla sıklıkla karşılaşılmaktadır. Böyle durumlarda, DA ile, veri büyüklüğü sonsuz olsa dahi, bağımlı değişken için elde edilen olasılığın doğru olması beklenemez. Böyle durumlarda, LR ve parametre tahminleri için tutarlı yöntemlerden biri olan en çok olabilirlik yöntemi kullanılabilir.

- Diskriminant tahminleri, normallik koşulu ihlal edildiğinde, model katsayılarının önemi ile ilgili hatalı sonuçlar verebilir. Bu durumda, gerçekte sıfır olan eğim katsayısının büyük örneklemlerde en çok olabilirlik yöntemi ile sıfır olarak tahmin edilme eğilimi olmasına rağmen, diskriminant fonksiyonlarıyla tahmin edilen parametrelerle anlamsız değişkenlerde modelde yer alma eğiliminde olacaktır.

- Halperin, Blackwelder ve Verter (1971) en çok olabilirlik yöntemiyle elde edilen LR ve diskriminant fonksiyonlarıyla elde edilen DA tahminlerinin sayısal karşılaştırmasını yapmışlardır. Normal dağılım varsayımının sağlanmadığı koşullarda genellikle en çok olabilirlik yönteminin modele biraz daha iyi uyum sağladığını bulmuşlardır. Ayrıca, eşitlik 2 ile ifade edilen model geçerli olsa bile diskriminant tahminlerinin çok zayıf bir uyum vermesi ihtimali için teorik bir taban olduğunu bulmuşlardır (Press vd., 1978).

- Lojistik regresyon modeli yeterli istatistiklerin elde edilebildiği bir modeldir. En çok olabilirlik tahminleri yeterli istatistik fonksiyonlarıdır ve her zaman hata kareler ortalamasını daha küçük yapar. Ancak diskriminant fonksiyon tahmini yeterli istatistikler elde edemez.

- Lojistik regresyon modelin en çok olabilirlik kestirimleri, olayların gözlenen ve beklenen sayılarının eşit olmasını gerektirir ($\sum y_i = \sum P(x_{1i}, ..., x_{ki})$). Bu özellik, herhangi bir düzleştirme sürecinde de öncelikli olarak olması istenen bir özelliktir. Diskriminant fonksiyonu yaklaşımında ise tahmin toplamları gerçek gözlem sayılarından büyük olabilmektedir.

- Bazı uygulamalarda, diskriminant fonksiyonu tahmin edicilerinin önemli derecede yanlı olma eğilimi gösterdiğine dair kanıtlar vardır. McFadden (1996), Bayesci analiz çalışmasında, açıklayıcı değişkenlerin tipik önsel dağılımı için, diskriminant analizine dayanan seçim olasılıklarının tahminlerinin önemli derecede yanlı olacağını söylemiştir (Press vd., 1978).

3. LOJİSTİK REGRESYON (LR) VE DOĞRUSAL DİSKRİMİNANT ANALİZİNİN (DA) KARŞILAŞTIRILMASINDA KULLANILAN İNDEKSLER

LR ve DA yöntemlerinin karşılaştırılmasında kullanılan en basit ölçüt Doğrusal Sınıflandırma Oranı (DSO)'dır. Buna rağmen, yeterince hassas ve istatistiksel olarak etkin bir ölçüt değildir. Değişkenlerin kategorik olması durumunda doğrusal sınıflandırma oranı ile minimum düzeyde bilgi elde edilebilir. DSO'nun değeri 0 ile 1 arasında değişmekle birlikte genel olarak yorumlama aralığı 0.5 ile 1 arasındadır ve 1'e yakın olması istenir. 0.5'in altında elde edilen bir DSO değeri genellikle "şansa bağlı uyum"a atfedilir. DSO'nun zayıf yönü ise iki gruba ait toplam doğru tahmin sayılarının hangisinde yoğunlaşma olduğu hakkında bilgi vermemesidir.

Harell ve Lee (1985), yöntemlerin tahminlerdeki doğruluğunu kıyaslamak amacıyla, doğrusal sınıflandırma oranından daha güvenilir ölçüt olarak A, B, C ve Q indekslerini önermiştir (Pohar vd., 2004). Önerilen bu indeksler, yöntemlerin tahmin etmede ya da bireyleri gruplara ayırmada ne kadar iyi olduğunu daha etkili şekilde karşılaştıran kriterlerdir. C indeksi eşitlik 4 ile verilmiştir.

$$C = \sum_{\substack{i=1 \\ y_i=0}}^{n} \sum_{\substack{j=1 \\ y_j=1}}^{n} [I(P_j > P_i) + \frac{1}{2} I(P_j = P_i)] / n_1 n_2 \qquad (4)$$

Burada, P_i i. gözleme ait olasılık değeri, P_j j. gözleme ait olasılık değerini ve I işaret fonksiyonunu göstermektedir. Görüldüğü gibi bu indeks, gerçek grup üye değerlerinden bağımsızdır. Sadece gruplar arası ayırmanın bir ölçüsüdür. Tahminin doğruluğuna dair bilgi vermez. C indeksi 1 değerini aldığında "mükemmel ayırma"ya, 0.5 değerini aldığında "şansa bağlı ayırma" ya atfedilir.

C indeksinden farklı olarak, B ve Q indeksleri tahminlerin doğruluğunu değerlendirmede kullanılabilir. B indeksi, tahmini olasılık değeri ile gerçek değerler arasındaki farkların karelerinin ortalamasının bir ölçüsüdür.

$$B = 1 - \sum_{i=1}^{n} (P_i - y_i)^2 / n \qquad (5)$$

Burada, P_i *i.* gözleme ait olasılık değeri, y_i ise gerçek değeridir (1 ya da 0). B indeksi 0 ile 1 aralığında değerler alır. B indeksinin 1 değeri alması "mükemmel tahmin"e atfedilir. Örneklem büyüklükleri eşit iken rastgele tahmin durumunda, B indeksinin değeri 0.75 civarındadır.

B indeksi ile benzer başka bir indeks ise Q indeksidir. Tahminlerin doğruluğunun bir ölçüsü olarak aşağıdaki eşitlik 6 ile bulunur.

$$Q = \sum_{i=1}^{n} [1 + \log_2 (P_i^{Y_i} (1 - P_i)^{1-Y_i})] / n \qquad (6)$$

Q indeksinin 1 değerini göstermesi "mükemmel tahmin"e, 0 değerini göstermesi "rastgele tahmin"e atfedilir. 0 değerinden küçük çıkması ise rastgele tahminden de kötü olduğunu gösterir.

İki düzeyli verilerde, gözlemciler ya da aynı gözlemci tarafından yapılan iki ölçüm arasındaki uyum genellikle Cohen'in kappa katsayısı (κ) ile incelenir. DSO'ya göre κ katsayısının avantajı, uyumun şans ile ortaya çıkma durumunu düzeltmektedir. Tablo 1 yardımıyla hesaplanan Cohen'in kappa katsayısı, eşitlik 7 ile verilir (Alpar vd., 2010).

Tablo 1. 2×2 uyum tablosu

	Gözlemci A		
Gözlemci B	Var	Yok	Toplam
Var	a	b	(a+b)
Yok	c	d	(c+d)
	(a+c)	(b+d)	n

$$\kappa = \frac{P_0 - P_e}{1 - P_e} \qquad (7)$$

Burada, P_0, DSO'yu göstermektedir. P_e şansa bağlı uyumu gösterir ve eşitlik 8 ile verilmektedir:

$$P_e = \frac{[(a+c)\times(a+b)] + [(b+d)\times(c+d)]}{n^2} \qquad (8)$$

Kappa katsayısı -1 ile +1 arasında değerler alır. $\kappa = 1$ olması durumunda tam uyum söz konusudur. $\kappa \geq 0$ ise gözlenen uyum şansa bağlı uyuma eşit ya da büyüktür. Gözlenen uyumun şansa bağlı uyumdan küçük olması durumunda $\kappa < 0$ olur. Kappa katsayısının negatif değerleri güvenirlik açısından anlamlı değildir. Bu yüzden sadece pozitif değerleri dikkate alınır.

Landis ve Koch uyumun derecesini elde edilen kappa katsayısı 0.20'ye eşit yada küçük ise "zayıf uyum", 0.21-0.40 aralığında ise "ortanın altında uyum", 0.41-0.60 aralığında ise "orta düzeyde uyum", 0.61-0.80 aralığında ise "iyi düzeyde uyum" ve 0.81-1.00 aralığında ise "çok iyi düzeyde uyum" olarak tanımlamışlardır (Subhash, 1996).

İki gözlemci arasındaki uyumun şansa bağlı kısmını düzelten bir uyum ölçüsü olarak tanıtılan kappa katsayısının iki sorunu vardır. Birinci sorun; DSO yüksek olmasına rağmen, satır–sütun toplamlarındaki büyük dengesizlikler nedeniyle küçük bir κ değerinin elde edilmesidir. İkinci sorun ise, satır-sütun toplamlarında asimetrik

dengesizlik olduğunda κ'nın simetrik dengesizlik olması durumuna göre daha yüksek bulunabilmesidir. (a+c) ile (b+d) ve (a+b) ile (c+d) toplamlarının eşit ya da birbirine yakın olması durumunda satır-sütun toplamlarının dengeli olduğu söylenir. Bu toplamlar birbirinden uzaklaştıkça satır-sütun toplamları dengesizleşmeye başlar. (a+c) ile (a+b) ve (b+d) ile (c+d) toplamlarının eşit ya da birbirine yakın olması durumunda ise satır-sütun toplamlarının simetrik olduğu söylenir. Yine bu toplamların birbirinden uzaklaşması satır-sütun toplamlarının asimetrik olmasına yol açmaktadır (Subhash, 1996).

Bu çalışmada, yukarıda tanımlanan Cohen'in Kappa'sı, gerçek grup üye değerleri ile tahmin edilen üye değerlerinin uyumunu ölçmek için kullanılmıştır. Kappa'nın kullanımında karşılaşılabilecek iki sorun da grup büyüklüklerinin eşit alınması sayesinde ortadan kaldırılmıştır. Burada, "gözlemciler" yerine "gerçek grup üye değerleri" ve "tahmin edilen grup üye değerleri" kullanılmıştır. "Evet-Hayır" yerine ise grupları temsilen "0-1" gösterimi kullanılmıştır.

4. UYGULAMA

Verilerin normal dağılım göstermesi varsayımı, diskriminant analizinin temel varsayımlarından biridir. Bu varsayımın ihlal edilmesi durumunda, böyle bir varsayıma gerek duymayan lojistik regresyon analizinin daha iyi sonuçlar vermesi beklenir. Bu beklenti literatürde daha önce yapılan çalışmalarla da desteklenmiştir. Bu çalışmada ise, bir benzetim çalışmasıyla farklı örneklem büyüklüğündeki normal dağılmayan veriler üretilmiş ve sonrasında bu verilere uygulanan lojistik regresyon ve diskriminant analizlerinden elde edilen sonuçlara ilişkin indeks değerleri hesaplanmıştır. Normal dağılım varsayımının ihlali durumunda, lojistik regresyon analizinin diskriminant analizinden iyi sonuçlar vereceği bilgisinden yola çıkarak, hesaplanan bu indekslerin başarısı karşılaştırılmıştır.

Bu çalışmada, dağılımları sırasıyla üstel, poisson ve tekdüze olan üç adet bağımsız değişken için veri üretimi yapılmıştır. Bu dağılımlar normal dağılım gösterme varsayımını bozmak amacıyla, keyfi olarak tercih edilmiştir. Açıklayıcı değişkenlerin dağılımlarına ilişkin parametre değerlerinin değiştirilmesi sayesinde iki farklı grup oluşturulmuştur. Gruplar arasındaki farklılık, Mahalonobis uzaklığından faydalanılarak, "yok", "az", "orta" ve "çok" olarak düzeylere ayrılmıştır. Bu düzeylere ilişkin parametre değerleri Tablo 2'de yer almaktadır. Her iki grubun örneklem büyüklükleri aynı olmak üzere, örneklem büyüklüğünün indekslere etkisini ölçmek amacıyla, altı farklı örneklem büyüklüğü için veriler üretilmiştir. Örneklem büyüklüğü seçimi için ön çalışma yapılıp, yeterince büyük ve yeterince küçük olan örneklem büyüklüklerini de içerecek şekilde keyfi olarak belirlenmiştir. Bu büyüklükler 20, 40, 60, 80, 100 ve 120'dir. Böylece, düzeyler ve örneklem büyüklükleri dikkate alınarak 24 farklı kombinasyon oluşturulmuştur.

Tablo 2. Gruplara ilişkin parametre değerleri

BAĞIMSIZ DEĞİŞKENLERİN DAĞILIMI	GRUPLAR	PARAMETRELER	GRUPLAR ARASI FARK			
			YOK	AZ	ORTA	ÇOK
ÜSTEL DAĞILIM	Grup 1	Ortalama	1	1	1	1
	Grup 2		1.1	1.5	2	3
POİSSON DAĞILIMI	Grup 1	Ortalama	1	2	3	2
	Grup 2		1.1	2.5	4	4
TEKDÜZE DAĞILIM	Grup 1	Alt Sınır	0	0	0	0
		Üst Sınır	1	1	1	1
	Grup 2	Alt Sınır	0	0	0	0
		Üst Sınır	1.1	1.5	2	3

Bu çalışmada her bir örneklem büyüklüğü için 1000 tekrar yapılmış ve elde edilen sonuçların ortalamaları hesaplanmıştır. Üretilen verilere lojistik regresyon ve diskriminant analizleri uygulanarak, bireylerin hangi gruba ait olduklarına ilişkin olasılık tahminleri yapılarak, gruplara atama işlemi gerçekleştirilmiştir. Sonraki aşamada, elde edilen tahmini grup üyelikleri ile gerçek grup üyelikleri yardımıyla her iki analize ait B, C ve Q indeksleri ile Kappa katsayısı hesaplanarak karşılaştırılmıştır.

5. SONUÇLAR VE TARTIŞMA

İki grup arasında fark olmadığı durumda lojistik regresyon ve diskriminant analizlerinin her ikisi de atama işlemini rasgele olarak yapacaklardır. Bu nedenle Tablo E1'de yer alan indeks değerlerinin birbirine çok yakın olduğu görülmektedir.

Literatürde rasgele tahmin değerinin 0.5 olarak verildiği C indeksinin, fark "yok" düzeyi için aldığı değerler 0.65 civarındadır ve bu değer örneklem büyüklüğü arttıkça azalma eğilimi göstermektedir. Buna rağmen örneklem büyüklüğü 120 için aldığı en düşük 0.58 değeri, 0.5 değerinden büyük bir değerdir. Bu nedenle, C indeksinin rasgele tahmini belirlemede zayıf kaldığı söylenebilir. B indeksi için rasgele tahmin değeri 0.75, Q için 0 olarak verilir. Kappa için 0.20'den az değerler kötü uyumu ifade eder. Bu çalışmada, bu değer rasgele tahmine karşılık gelmektedir. Tüm indeks değerlerinde, C indeksinde olduğu gibi örneklem büyüklüğü arttıkça azalma söz konusudur.

Gruplar arası fark olmadığında, rastgele tahmin yapan lojistik regresyon ve diskriminant analizinden elde edilen sonuçlar çok benzerdir. Ancak gruplar arasındaki fark düzeyi arttıkça lojistik regresyonun diskriminant analizinden daha başarılı olduğu dikkat çeker. Bunun en önemli sebebi; LR'nin DA gibi açıklayıcı değişkenlere normal dağılma şartı koymamasıdır.

Tüm farklılık düzeylerinde en dikkat çeken nokta, indeks değerlerindeki en büyük değişimin örneklem büyüklüğü 20 ile 40 arasında iken olmasıdır. Bu nedenle indekslerin küçük örneklem büyüklüklerine karşı hassas olduğu söylenebilir. Örneklem büyüklüğü arttıkça tüm indeks değerlerindeki değişim de azalmaktadır. Örneklem büyüklüğünden en fazla etkilenen indeks, fonksiyonel yapısı sebebiyle C indeksi olmuştur. Çünkü C indeksinin payda kısmında iki grubun örneklem büyüklüklerinin çarpımı yer alır. Q indeksinde ise örneklem büyüklüğü 40'tan büyük olduğunda değişim azdır.

Fark düzeylerinin yüksek olması durumunda, beklendiği gibi tüm indeks değerleri iyi bir atama yapıldığını işaret etmektedir. Kappa'nın 0.61-0.80 arasında aldığı değerler iyi bir grup atamasının ifade eder. Lojistik regresyonun, diskriminant analizinden üstünlüğünü ifade eden en iyi indeks ise Q indeksidir. Bu indekslerle aynı amaçla kullanılabileceği iddia edilen Kappa katsayısı, gruplar arasındaki fark "orta" düzeydeyken (Tablo E3), diskriminant analizi için lojistik regresyon analizi değerlerinden daha yüksek değerler vermiştir. Beklenenin aksi yönünde verdiği bu sonuca rağmen, diğer indekslere nazaran daha kolay hesaplanabilir olması ve çoğu zaman iyi sonuçlar elde etmesi nedeniyle, iki yöntemi karşılaştıran bir indeks olarak kullanılabileceği düşünülmektedir.

Sonuç olarak, tahmin edilen olasılık değerlerini dikkate alarak hesaplanan Q indeksinin dikkate alınan indeksler arasında en iyi indeks olduğu görülmüştür. Q indeksinin ardından, yine olasılıkları kullanan B indeksi iyi bir indekstir denilebilir.

6. KAYNAKLAR

Efron, B., 1975. The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis , Journal of the American Statistical Association, Vol. 70, No. 352, 892-898.

Pohar, M., Blas, M., Turk, S., 2004. Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study, Metodoloski Zvezki, Vol. 1, No. 1, 143-161.

Press, S. J., Wilson, S., 1978. Choosing Between Logistic Regression and Discriminant Analysis, Journal of the American Statistical Association, Vol. 73, No. 364, 699-705.

Gürcan, M., 1998. Lojistik Regresyon ve Bir Uygulama, Yüksek Lisans Tezi, On Dokuz Mayıs Üniversitesi 63s., Samsun.

McFadden, D., 1976. "A Comment on Discriminant Analysis 'Versus' Logit Analysis", Annals of Economic, and Social Measure- ment, 5, 511-523.

Özdamar, K., 1997. Paket Programlar ile İstatistiksel Veri Analizi, Anadolu Üniversitesi Yayınları 512 s., Eşkişehir.

Subhash, S., 1996. Applied Multivariate Techniques, John Wiley & Son's, USA, p. 287-317.

Alpar, C. R., Gözükara, Bağ, H. G., Karabulut, E., 2010. 2x2 Tablolarda Gözlemciler / Gözlemler Arası Uyumun Değerlendirilmesi, Hacettepe Diş Hekimliği Fakültesi Dergisi Cilt: 34, Sayı: 1-2, Sayfa: 46-52.

A COMPARISON OF SOME INDEXES USED IN LOGISTIC REGRESSION AND LINEAR DISCRIMINANT ANALYSIS

ABSTRACT

Logistic regression and linear discriminant analysis are two widely used methods to classify individuals or observations. The use those analyses for the same goal brings in mind the question of which analysis present better results. A comparison has been made and a study has been presented on this matter. Assumptions that are necessary for discrimant analysis, which are not necessary for logistic regression, made preferences switch under different conditions. In this paper, a simulation study has been carried out to make a comparison of these two methods and the indexes that are used to evaluate these analyses when the variables do not satisfy the normal distribution assumption. It is found that in general logistic regression analysis presents better results in comparison with the discriminant analysis method for data generated from multivarite non-normal distribution. As the sample size increases, the diversity of the results obtained from both analyses are considerably decreased. It is found that no matter what the sample size is logistic regression analysis has always been better in classification than the discriminant analysis method according to any index criteria. Cohen's Kappa coefficient has been used as a new index. In addition, when the better model is known, the indexes used are evaluated in terms of their success of estimating the true model.

Keywords: Discriminant analysis, Cohen's Kappa, Logistic regression.

Ek : Tablolar

Tablo E1. İki Grup Arası Fark Yokken İndeks Sonuçları

Örneklem Büyüklüğü	Analiz	C	B	Q	Kappa
20	LR	0.652690	0.772009	0.068327	0.222400
	DA	0.652700	0.771857	0.067631	0.221850
40	LR	0.612874	0.761998	0.036343	0.162175
	DA	0.612838	0.761970	0.036181	0.162200
60	LR	0.595916	0.758619	0.025678	0.135533
	DA	0.595866	0.758606	0.025627	0.135267
80	LR	0.591763	0.757949	0.023716	0.127813
	DA	0.591769	0.757943	0.023681	0.128375
100	LR	0.583337	0.756516	0.019278	0.115680
	DA	0.583335	0.756513	0.019262	0.115210
120	LR	0.580103	0.755996	0.017722	0.111783
	LR	0.580086	0.755994	0.017710	0.111633

Tablo E2. İki Grup Arası Fark Azken İndeks Sonuçları

Örneklem Büyüklüğü	Analiz	C	B	Q	Kappa
20	LR	0.753422	0.804738	0.179261	0.380450
	DA	0.752115	0.804025	0.175499	0.378400
40	LR	0.734282	0.795829	0.146955	0.345100
	DA	0.733442	0.795514	0.145377	0.345300
60	LR	0.728986	0.793478	0.138392	0.337217
	DA	0.728134	0.793280	0.137380	0.336867
80	LR	0.725708	0.791935	0.133141	0.330775
	DA	0.724951	0.791810	0.132410	0.329913
100	LR	0.726404	0.792047	0.133451	0.331050
	DA	0.725529	0.791904	0.132687	0.331550
120	LR	0.723959	0.790927	0.129502	0.327800
	LR	0.723090	0.790803	0.128847	0.327300

Tablo E3. İki Grup Arası Fark Ortayken İndeks Sonuçları

Örneklem Büyüklüğü	Analiz	C	B	Q	Kappa
20	LR	0.845927	0.850054	0.346129	0.556700
	DA	0.842250	0.847358	0.330723	0.553800
40	LR	0.838228	0.843362	0.316924	0.541350
	DA	0.835191	0.841994	0.307992	0.543100
60	LR	0.833851	0.839913	0.303627	0.530433
	DA	0.831148	0.838865	0.296234	0.533233
80	LR	0.832706	0.839230	0.300703	0.528350
	DA	0.830234	0.838325	0.294164	0.531500
100	LR	0.831722	0.838650	0.298142	0.527240
	DA	0.829292	0.837861	0.292004	0.530770
120	LR	0.830432	0.837653	0.294297	0.524600
	LR	0.827956	0.836888	0.294297	0.528933

Tablo E4. İki Grup Arası Fark Çokken İndeks Sonuçları

Örneklem Büyüklüğü	Analiz	C	B	Q	Kappa
20	LR	0.948777	0.923619	0.644286	0.791550
	DA	0.943107	0.910164	0.574960	0.758850
40	LR	0.942987	0.914602	0.601025	0.765050
	DA	0.939308	0.905190	0.553419	0.744700
60	LR	0.942759	0.913193	0.593044	0.763300
	DA	0.939907	0.904643	0.550469	0.742717
80	LR	0.941085	0.911116	0.583672	0.757188
	DA	0.938178	0.902791	0.542595	0.738738
100	LR	0.941284	0.911257	0.583109	0.759110
	DA	0.938671	0.903337	0.543702	0.738910
120	LR	0.941504	0.911213	0.583334	0.757825
	LR	0.938975	0.903327	0.544222	0.739225