LOJİSTİK REGRESYON ANALİZİ: ÖĞRENCİLERİN SİGARA İÇME ALIŞKANLIĞI ÜZERİNE BİR UYGULAMA

Yrd. Doç. Dr. Cengiz AKTAŞ
Eskişehir Osmangazi Ünv. Fen-Ed.Fak. İstatistik Böl.
caktas@ogu.edu.tr

Öz

Sigara, tüm dünyada korunulabilir hastalıklar arasında ölüm oranı en yüksek olan sağlık riskidir. Öğrencilik dönemi sigaraya başlamak için riskli bir dönemdir. Bu yönüyle öğrencilerin sigara içme davranışlarının bilinmesi önemlidir. Bu çalışmada önce, bağımlı değişkenin iki düzeyli olması durumunda demografik, davranış ve risk faktörüyle ilgili tahmin çalışmalarında olduça sık kullanılan lojistik regresyon analizi teorik olarak kısaca incelenmiştir. Daha sonra, Eskişehir Osmangazi Üniversitesi (ESOGÜ) öğrencileri arasında sigara içme alışkanlığını etkileyen faktörleri belirlemek için lojistik regresyon ve diskriminant denklemi belirlenmiştir.

Anahtar Kelimeler : Lojistik Regresyon, Sigara İçme, Sınıflama, Odds

THE LOGISTIC REGRESSION ANALYSIS AND ITS APPLICATION ON THE SMOKING PREVALENCE OF STUDENTS

Abstract

Smoking is a habit risk with highest mortality among the worldwide preventable diseases. While student period is a risky period for starting smoking. At this point of view it is important to know the smoking behaviour of students. In this study, firstly, it was briefly examined which logistic regression analysis are frequently used in studies for estimating associations that demographic, behavioral, and risk faktor variables have on a dichotomous outcome. Afterwards, for to determine factors on the habit of smoking among Eskisehir Osmangazi University (ESOGU) students, a logistic regression and discriminant equations is developed.

Keywords : Logistic Regression, Smoking, Classify, Odds

I. GİRİŞ

Bir gözlemi birkaç kütleden birine atamak, sınıflamadır. Eğer kütleler ortak varyans-kovaryans matrisine sahip ve normal dağılmışsa, diskriminant analizi kestiricileri, diskriminant analizi problemleri için lojistik regresyon kestiricilerine tercih edilebilir. Bununla birlikte pek çok diskriminant analizi uygulamasında değişkenlerden en az birinin kategorik değişken olması nedeniyle çok değişkenli normallik varsayımı geçerli olmayacaktır. Böyle durumlarda bağımsız değişkenlerin kategorik ve sürekli olmaları konusunda bir kısıt getirmeyen,

gözlemlerin atanması amacıyla kullanılabilen lojistik regresyon analizi önerilmektedir (Press ve Wilson, 1978: 2).

Gordon ve Kannel'in (1968) kardiyolojik hastalıklarla ilgili yaptıkları çalışma ikili lojistik regresyon analizinin başlangıcı olmuştur (Carroll ve diğerleri 1984). Bu dönüm noktası niteliğindeki çalışmadan sonra da biyoistatistik (Finney (1971)), müşteri seçim analizleri (Maddala (1983)) ve kriminoloji (Larntz (1980)) alanlarında yapılan çalışmalarla uygulama alanları genişlemiştir (Dufffy ve Santner, 1989). Lee (1984) basit dönüşümlü (cross-over) deneme planları için doğrusal lojistik modeller üzerinde durmuştur.

Lojistik regresyon modelleri, son yıllarda biyoloji, tıp, ekonomi, tarım, veterinerlik ve taşıma sahalarında yaygın olarak kullanılmaktadır. Breslow ve Day (1980), Pastides ve diğerleri (1985) halk sağlığı alanında, Abbott (1985), Efron (1988) yaşam analizi ile igili uygulamalı çalışmalar yapmışlardır. Gardside ve Glueck (1995) insanlarda beslenme şekli, sigara ve alkol kullanımı, fiziksel aktivite gibi risk faktörlerinin kalp hastalığı üzerindeki etkilerini incelemiştir. Bonney (1987) lojistik regresyon modelinin kullanımı ve geliştirilmesi üzerinde çalışmıştır. Duffy (1990) lojistik regresyonda hata terimlerinin dağılışı ve parametre değerlerinin gerçek değerlere yaklaşımını incelemiştir. Kloiber ve ark (1996), Peoples ve ark. (1991), Buescher ve ark. (1993) kadınlarda düşük doğum ağırlığını etkileyen risk faktörlerini; Santos ve ark. (1998) kafein tüketimi ve düşük doğum ağırlığı arasındaki ilişkiyi; Sable ve Herman (1997) erken doğum ve düşük doğum ağırlığı arasındaki ilişkiyi incelemişlerdir (Bircan, 2004: 186-187).

Türkiye'de de bu konuda çeşitli alanlarda çalışmalar yapılmıştır. Bunlardan bazıları ise şunlardır: Vupa ve Çelikoğlu (2006) akciğer kanseri hastaları için lojistik regresyon modeli önermişlerdir. Ünsal ve Güler (2005) Türk bankacılık sektörünü lojistik regresyon analiziyle incelerken, Tatlıdil, Başarır ve Hökmen (1990) ülkelerin sosyo ekonomik gelişmişliklerine göre sınıflandırılmasına ilişkin çalışma yapmışlardır. Ayrıca, Aktaş ve Yılmaz (2001) LPG kullanan özel araç sürücülerinin sınıflandırılımasını, Çolak ve Özdamar (2004) ölümle sonuçlanan trafik kazalarında risk faktörlerini, lojistik regresyon analiziyle incelemişlerdir.

Tütün genellikle sigara şeklinde tüketilen, birey ve toplum sağlığına son derece zararlı maddelerden biridir. Asıl etken maddesi fiziksel ve psikolojik bağımlılık yapan nikotindir. Sigara kullanımının başta neoplastik hastalıklar, kronik obstrüktif akciğer hastalığı, kardiyovasküler sistem hastalıkları olmak üzere birçok değişik hastalığın etiolojisinde doğrudan veya dolaylı olarak etkili olduğu bilinmektedir.

Sigaranın kısa sürede alışkanlık yapabilmesi, dünyanın her yerinde kolayca temin edilebilir olması, sadece sigara içenleri değil, çevrede bulunanların sağlığını da tehdit etmesi gibi nedenler, sigaranın halk sağlığı açısından önemini belirten nedenlerden bazılarıdır. Sigaranın bir diğer özelliği; zararlı etkilerinin hemen ya da kısa sürede ortaya çıkmaması nedeniyle sigara içenlerin konuyu

önemsememeleridir. Sigara kullanma sıklığı, ülkeden ülkeye ve yıldan yıla değiştiği gibi aynı toplumun değişik kesimlerinde farklılıklar göstermektedir. Ancak, son yıllarda özellikle, ülkemizdeki gençler arasında sigara içme alışkanlığında önemli bir artış sözkonusudur (Tekbas v.d. 2006: 106).

Düzenli bir şekilde sigara içmeye başlayıp, içmeyi sürdürenlerin yarısı sigara nedeniyle yaşamlarını kaybetmektedir. Sigara nedeniyle 35-69 yaş arasında ölenlerin yaşamlarından kaybettikleri süre 20-25 yıl olarak hesaplanmıştır. Dünya'da 2000 yılında sigara nedeniyle öldüğü tahmin edilen insan sayısı, yarısı gelişmekte olan ülkelerden olmak üzere, 4 milyon olarak tahmin edilmektedir (Demirel ve Sezer, 2005: 1).

Amerika Birleşik Devletleri'nde sigara tüketimi 1981 yılında 640 milyar adetken, tüketimi sürekli azalarak 2000 yılında 430 milyar adete düşmüştür. Bu 20 yıllık dönemde düşme oranı %32.8'dir. Türkiye'de 1985'de yaklaşık 64.8 milyar adet olan yıllık sigara satışı, 2000'de yaklaşık 122.6 milyar adete ulaşmıştır, yani söz konusu dönemde %89.2 oranında artmıştır (Demirel ve Sezer, 2005: 1). Sigara içme alışkanlığı yaklaşık %40 oranında 15-19 yaşlarında başlamakta; dünyada ve Türkiye'de 15 yaşın üzerindeki nüfusun %45'inin sigara bağımlısı olduğu varsayılmaktadır (İlhan v.d, 2005: 189).

Türkiye'de acele olarak müdahale edilmesi gerekli bir sigara salgını yaşanmaktadır. Yapılacak müdahaleler başlamayı önleme, bırakmayı destekleme ve sigara dumanının kontrol altına alınması öğelerini içermek durumundadır. Bu müdahalelerin çevresel planlanmasında müdahale öncesi durumun tanımlanması önemlidir. Türkiye'nin sigara salgınına yönelik mücadelesinde üniversiteler önemli alanlardan biri olarak düşünülebilir (Demirel ve Sezer, 2005: 2). Dolayısıyla gençlerin neden sigaraya başladıklarının bilinmesi, sigarayı bıraktırma mücadelesinde oldukça yararlı olacaktır. Bundan dolayı da bu çalışmanın amacı, diskriminant ve lojistik regresyon analizleri yardımıyla, Eskişehir Osmangazi Üniversitesi öğrencilerinin sigara içme alışkanlığını etkileyen faktörleri ortaya koymak ve bunu sağlayacak en uygun denklemi belirlemektir.

Bu makalede önce, lojistik regresyon ve diskriminant analizine ilişkin teorik bilgiler kısaca sunulduktan sonra, Eskişehir Osmangazi Üniversitesi öğrencileri arasında sigara içme alışkanlığını etkileyen faktörleri belirleyebilmek için ampirik bulgulara yer verildi.

2. GÖZLEMLERİN VAROLAN GRUPLARA ATANMASI

Değişkenler arası ilişkileri incelemede en çok kullanılan istatistik yöntemlerinden biri, regresyon analizidir. Regresyon analizi, çözümüne başlamadan yapılması gereken değişkenlerin niteliklerinin bilinmesi ve bağımlı değişken ile bağımsız değişkenin en iyi şekilde tayin edilmesidir. Genelde bilinen bağımlı değişken ölçülebilir nitelikte olup, sürekli bir değişkendir. Ancak, her zaman bağımlı değişken sürekli değişken niteliğinde olmayabilir. Örnegin, öğrencilerin sigara içip

içmediğinin belirlenmesi amaçlandığında, öncelikle belirtilmesi gereken bağımlı değişkenin sürekli bir değişken olmayıp, kategorik bir değişken olduğudur.

Lojistik regresyon analizinin kullanım amacı, istatistikte kullanılan diğer model yapılandırma teknikleriyle aynıdır. En az değişkeni kullanarak en iyi uyuma sahip olacak şekilde bağımlı (sonuç) değişkeni ile bağımsız değişkenler kümesi (açıklayıcı değişkenler) arasındaki ilişkiyi tanımlayabilen ve genel olarak kabul edilebilir modeli kurmaktır.

Lojistik regresyonu, doğrusal regresyondan ayıran en belirgin özellik ise, lojistik regresyonda bağımlı değişkeninin kategorik değişken olmasıdır. Lojistik regresyon ve doğrusal regresyon arasındaki bu fark, hem parametrik model seçimine, hem de varsayımlara yansımaktadır. Lojistik regresyonda da, doğrusal regresyon analizinde olduğu gibi bazı değişken değerlerine dayanarak kestirim yapılmaya çalışılır, ancak iki yöntem arasında üç önemli fark vardır: (Çoşkun v.d, 2004: 43).

1-Doğrusal regresyon analizinde tahmin edilecek olan bağımlı değişken sürekli iken, lojistik regresyonda bağımlı değişken kesikli bir değer olmalıdır.

2-Doğrusal regresyon analizinde bağımlı değişkenin değeri, lojistik regresyonda ise bağımlı değişkenin alabileceği değerlerden birinin gerçekleşme olasılığı kestirilir.

3-Doğrusal regresyon analizinde bağımsız değişkenlerin çoklu normal dağılım göstermesi koşulu aranırken, lojistik regresyonun uygulanabilmesi için bağımsız değişkenlerin dağılımına ilişkin hiçbir ön koşul yoktur.

Gözlemleri verilerin yapısında bulunan olası gruplara atamak için;

i) Kümeleme analizi,
ii) Diskriminant analizi,
iii) Lojistik regresyon analizi tekniklerinden yararlanılır.

Kümeleme analizinde, verilerin yapısındaki grup sayısı bilinmemekte, gözlemler uzaklık ya da benzerlik ölçütlerine göre kümelenmektedir. Burada amaç, yanlızca gözlemlerin oluşturduğu kümenin yapısını bulmaktır.

Diskriminant ve lojistik regresyon analizinde ise, yapısındaki grup sayısı bilinmekte ve bu verilerden faydalanarak bir ayrımsama modeli elde edilmektedir. Kurulan bu model yardımı ile veri kümesine yeni alınan gözlemlerin gruplara atanması yapılmaktadır (Başarır, 1990:1). Çalışmamızda diskriminant ve lojistik regresyon analizi uygulandığından, sadece bu iki teknik incelenmiştir.

2.1. Diskriminant Analizi

Kütlelerin normal dağılımlı ve ortak varyans-kovaryans matrisine sahip olmaları durumunda gözlemlerin varolan gruplardan birine atanması amacıyla kullanılan tekniklerden birisi de, diskriminant (ayırma) analizidir. Diskriminant anal-

izi, istatistiksel bir karar vermedir. Yani hatalı sınıflandırma olasılığını en aza indirgeyerek gözlemleri (birimleri) ait oldukları gruplara ayırmak, çekilmiş oldukları kütleleri belirlemektir.

İki grup sözkonusu olduğunda (g=2) bir gözlemi A_1 ve A_2 kütlelerinden birine atamak için çoğunlukla kullanılan kural

$$v=(\underline{\overline{X}}^{(1)} - \underline{\overline{X}}^{(2)})^T S^{-1}(\underline{X} - \frac{1}{2})(\underline{\overline{X}}^{(1)} + \underline{\overline{X}}^{(2)})^* \quad (1)$$

değerini hesaplamaktır. Eğer v≥c ise gözlem A_1 içine, aksi takdirde A_2 içine sınıflanır. Buradaki $\underline{\overline{X}}^{(1)} ve \underline{\overline{X}}^{(2)}$ n_1 ve n_2 büyüklüğünde iki bağımsız örneklemin ortalama vektörleri, S ise örneklem varyans-kovaryas matrisini göstermektedir ve bu matrisin elemanları

$$s_{ii'} = \sum_{g=1}^{2}\sum_{\alpha=1}^{n_g} \frac{(x_{i\alpha}^{(g)} - \overline{X}_i^{(g)})(x_{i'\alpha}^{(g)} - \overline{X}_{i'}^{(g)})}{n_1 + n_2 - 2} \quad (2)$$

dir. c'nin seçimi ise çeşitli şekillerde yapılabilir. Eğer k bağımsız değişken normal dağılımlı ve iki kütlenin kovaryans matrisleri aynı (ortak) ise o zaman c değeri aşağıdaki gibi belirlenebilir:

$$c = \ln(\frac{\hat{A}(2)}{\hat{A}(1)}) \quad (3)$$

Buradaki $\hat{A}$(g), A(g)'nin bir kestiricisidir. A(g) ise bir birimin A(g)'den seçilmesinin önsel (priori) olasılığıdır. Ancak çoğunlukla

$$\hat{A}(1) = \hat{A}(2) = \frac{1}{2} \quad \text{kabul edilerek}$$

$$c = \ln(\frac{\hat{A}(2)}{\hat{A}(1)}) = 0 \quad (4)$$

olarak alınır (Blasfield vd, 1989: 68).

2.2 İki Düzeyli Lojistik Regresyon Analizi

k bağımsız değişken ve N gözlem olduğunda doğrusal regresyon modelinin genel formu i.gözlem için

$y_i=\beta_0+\beta_1x_{1i}+\beta_2x_{2i}+.....+\beta_kx_{ki}+\varepsilon_i$'dir. (5)

Örneklem büyüklüğü n olduğunda ise doğrusal regresyon modeli

[*] Altı çizili harfler matris ya da bir vektörü gösterecektir

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \ldots\ldots\ldots + \hat{\beta}_k x_{ki} + e_i \quad (6)$$

şeklinde yazılır.

Bağımlı değişkenin alabileceği değerlerin 0-1 arasında olmasını sağlamak için bağımsız değişken ve bağımlı değişken arasında eğrisel bir ilişkiyi sağlayan modeli kullanmak daha uygundur. β_1'in işaretine göre S veya ters S şeklinde olan eğrileri sağlayan

$$E(y_i)=\pi_i=\frac{\exp(\beta_o + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots\cdots\cdots + \beta_k x_{ki})}{1+\exp(\beta_o + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots\cdots\cdots + \beta_k x_{ki})} \quad (7)$$

formundaki bu fonksiyona "Lojistik Fonksiyon" adı verilir. Bu lojistik fonksiyonlar genellikle S şeklinde fonksiyon olarak isimlendirilir. Bunlar 0 ve 1 asimtotlarına sahiptir ve böylece E(y), 0 ile 1 sınırları arasında kalır.

Lojistik fonksiyonun diğer bir özelliği de kolayca doğrusallaştırılabilir olmasıdır ve

$$\eta=\ln\left(\frac{\pi_i}{1-\pi_i}\right) \quad (8)$$

dönüşümü yapılarak bağlantı fonksiyonu elde edilir. Eşitlik (4)'deki $\pi_i/(1-\pi_i)$ oranı ise "Odds Oranı" olarak nitelendirilir. Ln odss dönüşümü ise "Lojit" olarak isimlendirilir ve ln odds için elde edilen

$$E(y_i)=\eta=\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\ldots\ldots+\beta_k x_{ki} \quad (9)$$

modele "Lojistik (ya da lojit) Regresyon Modeli " denir. $E(y_i)$ ise $-\infty$, ∞ arasında değer almaktadır (Agresti, 1990: 106).

2.2 1. Lojistik Regresyon Analizinde Parametre Tahmini

Bağımlı değişkeni (0,1) gibi iki düzey içeren bir lojistik regresyon modelindeki parametrelerin kestirimi genellikle diskriminant fonksiyonu, iteratif ağırlıklı enküçük kareler ve enbüyük olabilirlik kestirim tekniklerinden biriyle yapılır. Lojistik regresyon analizi için literatürde en çok kullanılan teknik ise, enbüyük olabilirlik (maximum likelihood) tekniğidir. Bu nedenle, çalışmamızın uygulama bölümündeki çözümleme için enbüyük olabilirlik tekniği kullanıldığından bundan sonraki kısımda sadece bu teknik ele alınmıştır.

Hata terimlerinin normal dağılım gösterdiği durumlarda doğrusal regresyon modelinin katsayılarının kestiriminde, enküçük kareler fonksiyonunun temelini oluşturan genel kestirim tekniği "enbüyük olabilirlik" tekniğidir. Bu teknik, lojistik regresyon modelinin katsayı kestirimlerini elde etmek için temel oluşturur. Çok genel bir anlamda enbüyük olabilirlik tekniği, gözlemlenmiş veri kümesinden elde edilmenin olasılığını enbüyük yapacak bilinmeyen parametrelerin değerlerini verir.

Bu tekniği uygulamak için ilk olarak olabilirlik (likelihood) fonksiyonu olarak isimlendirilen bir fonksiyon kurulur.

$P(y_i=1)= \pi_i$

$P(y_i=o)=(1-\pi_i)$

de her bir y_i gözlemi bir Bernoulli tesadüfi değişkenidir. Bunun olasılık dağılımı da

$$P(y_i/x_i)=f_i(y_i)= \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \qquad i=1,2.....n \qquad (10)$$

dir.

Bu durumda $f_i(y_i)$, $y_i=1$ ya da $y_i=0$ olmasının basit bir olasılığıdır. y_i gözlemleri bağımsızdır ve bunların birleşik olsılık fonksiyonu

$$L(\underline{y}/\underline{X})=P(\underline{y}/\underline{X})=\prod_{i=}^{n} f_i(y_i) = \prod_{i=1}^{n} \pi_i^{y_i} (1-\pi_i)^{1-y_i} \qquad (11)$$

dir. Bu fonksiyon n birimlik bir örneklemde ($\underline{y},\underline{X}$) gözlemlerinin bir fonksiyonu olduğundan buna "olabilirlik fonksiyonu " denir (Neter vd, 1989: 73).

Enbüyük olabilirlik tekniğinde ya olabilirlik fonksiyonu ya da olabilirlik fonksiyonunun logaritması enbüyük yapılır. Ancak, olabilirlik fonksiyonunun logaritmasını en büyük yapmak daha kolaydır. Bu durumda olabilirlik fonksiyonunun logaritması

$$LnL(\underline{y}/\underline{X},\underline{\beta})=\sum_{i=1}^{n} [y_i Ln\pi_i+(1-y_i)Ln(1-\pi_i)] \qquad (12)$$

ya da

$$LnL(\underline{y}/\underline{X},\underline{\beta})=\sum_{i=1}^{n} [y_i \underline{X}^T{}_i\underline{\beta}-ln(1+exp(\underline{X}^T{}_i\underline{\beta})] \qquad (13)$$

$\underline{X}^T=(1,x_{1i},x_{2i},x_{3i},.... x_{ki})$'dir.

Lojistik regresyon modelinde $(\beta_0,\beta_1,\beta_2,...,\beta_k)$'nın enbüyük olabilirlik kestiricileri, eşitlik (10)'daki olabilirlik fonksiyonunun logaritmasını enbüyük yapacak $(\beta_0,\beta_1,\beta_2,...,\beta_k)$ değerleridir. Eşitlik (10)'u enbüyük yapmak için $\beta_0,\beta_1,\beta_2,...,\beta_k$'ya göre kısmi türevleri alınıp sıfıra eşitlenirse aşağıdaki "olabilirlik denklemleri" elde edilir:

$$\sum_{i=1}^{n} (y_i-\pi_i)=0 \quad ve$$

$$\sum_{i=1}^{n} x_{ij}(y_i-\pi_i)=0 \quad j=1,2.....k \qquad (14)$$

(Hosmer ve Lemeshow, 1989: 9-10).

Bu denklemlerin çözümü ile $\underline{\beta}$'nın kestirim değerleri elde edilir. Ancak eşitlik (5)'deki π_i'nin üstel olması nedeniyle, bu denklemler doğrusal değildir ve bunların çözümü için Newton-Raphson iteratif işlemleri (çözümleme) önerilmiştir. Newton-Raphson iteratif işlemleri, $\underline{\beta}$'nın enbüyük olabilirlik kestiricisini elde etmek ($\underline{\hat{\beta}}_{ML}$) için $\underline{\hat{\beta}}$ gibi başlangıç bir kestiriciyi temel alır. Ancak, bu başlangıç değerleri için çeşitli alternatifler öne sürülmüştür. Bunlardan bir tanesi Newton Raphson'un önerilerinden biri olan, başlangıç kestiricisi için ortak bir seçim olarak enküçük kareler kestiricisini almaktır.

Newton-Raphson tarafından önerilen ve (t+1)'inci iterasyonda $\underline{\hat{\beta}}$ değerini bulan iterasyon

$$\underline{\hat{\beta}}(t+1)=\underline{\hat{\beta}}(t)+(\underline{X}^T\underline{V}(t)\underline{X})^{-1}\underline{X}^T\underline{r}(t) \qquad (15)$$

dir. Eşitlikteki

$$\underline{V}=\text{diag}[\pi_i(1-\pi_i)] \quad \underline{r}=\underline{Y}-\underline{P}_i \qquad (16)$$

olarak hesaplanır.

İterasyon işlemlerine yakınsama sağlanıncaya kadar devam edilir. Yakınsama ise iterasyonlar arasında fark olmaması durumunda sağlanmaktadır.

Lojistik modelin en büyük olabilirlik kestirimlerini bulmak için iterasyona başlarken eşitlik (13)'deki başlangıç değerlerini vermenin çeşitli yolları vardır. Bunlardan iki tanesi diskriminant fonksiyonunun katsayılarını kullanmak ve grafiksel gösterimlerden gözle kestirimde bulunmaktır. Başlangıç değerlerinin doğruluğu iterasyon sayısı ve kestirimlerin doğruluğu üzerinde önemli etkiye sahiptir. İyi bir başlangıç değeri ile az sayıda iterasyon sonucu optimum çözüme ulaşılabilmektedir (Aktaş, 1995: 32).

2.2.2. Lojistik Sınıflandırma ve Katsayıların Yorumlanması

Bir gözlemi birkaç gruptan birine atamak, sınıflamadır. Genellikle $P(y_i=1/x_i)$ değerini belirlemek amacıyla kullanılan lojistik regresyon modeli aynı zamanda bir sınıflandırma modeli olarak da kullanılır. Seçilen n birimlik bir örneklem sonucu elde edilen

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots\cdots + \hat{\beta}_k x_{ki} \qquad (17)$$

lojistik regresyon denklemi yardımıyla bulunan

$$P_i=\frac{e^{\hat{y}_i}}{1+e^{\hat{y}_i}} \qquad (18)$$

değerinin ≥ 0.5 olması durumunda $y_i=1$, π_i,$P_i<0.5$ ise $y_i=0$ biçiminde sınıflanır (Aktaş ve Yılmaz, 2001:253).

Lojistik regresyon fonksiyonunda tahmin edilen regresyon katsayılarının yorumlanması, doğrusal regresyon modelindeki kadar kolay değildir. x eksenindeki başlangıç noktasına göre hazırlanan lojistik regresyon modelinde x değişkenindeki bir birimlik artışın tesirini ölçmek zordur. β_1 katsayısı yorumlanırken x'deki bir birimlik artış için $\pi_i/(1-\pi_i)$ odds tahmini ile $\exp(\beta_1)$ çarpılarak elde edilen lojistik regresyon fonksiyonundan yararlanılır.

Lojistik modeldeki etkiler odds'a dayanır. x'in bir değerinde kestirilen odds'un, diğer değerinde kestirilen odds'a oranı olarak verilmektedir. Bu istatistik x=1 olan bireylerin x=0 olan bireylere nazaran bağımlı değişkenin kaç kat daha fazla 1 olarak görüldüğü sonucunu verir (Bircan, 2004: 29).

III. VERİ VE AMPİRİK BULGULAR

Bu kısımda öğrencilerin sigara içmesine neden olabilecek bağımsız değişkenler yardımıyla, Eskişehir Osmangazi Üniversitesindeki öğrencilerin sigara içmesinde etkili olan faktörlerin belirlenmesi için bir analiz yapılmıştır. Çalışmamızda bağımlı değişken iki düzeyli kategorik değişken olduğundan, bu tür verilerin analizinde uygulanan lojistik regresyon analizi kullanılarak, sigara içmede etkili olan en önemli değişkenlerin belirlenmesine çalışılmıştır.

Çalışmada kullanılan sigara içmede etkili olduğu düşünülen bağımsız değişkenler aşağıdaki gibidir:

x_1: Yaş,

x_2: Cinsiyet (0-Bay, 1-Bayan olarak kodlanmıştır),

x_3: Toplam aylık gelir,

x_4: Aylık harçlık,

x_5: Barınma şekli (0-Ev, 1-Özel Yurt, 2-Devlet Yurdu),

x_6: Babanın sigara içme durumu (0- Kullanıyor, 1-Kullanmıyor),

x_7: Annenin sigara içme durumu (0- Kullanıyor, 1-Kullanmıyor),

x_8: Ailede alkol kullanma durumu (0- Kullanıyor, 1-Kullanmıyor),

x_9: Öğrencinin alkol kullanma durumu (0-Kullanıyor , 1-Kullanmıyor),

x_{10}: Arkadaş çevresinin sigara kullanma durumu (0- Kullanıyor, 1-Kullanmıyor),

x_{11}: Spor yapma durumu (0-Evet, 1-Hayır),

x_{12}: Sigaranın sıkıntı, stress ve yanlızlığı giderdiğini düşünmesi (0-Evet, 1-Hayır),

x_{13}: Sigaranın statü kazandırdığını düşünmesi (0-Evet, 1-Hayır).

Bağımlı değişken y ise

0- Sigara kullanıyor,
1- Sigara kullanmıyor

olarak kodlanmıştır.

Yukarıda belirtilen değişkenlere ilişkin veriler, Eskişehir Osmangazi Üniversitesi Meşelik Kampüsü'nde okuyan lisans öğrencileri arasından, basit tesadüfi örnekleme uygulanarak seçilen 600 öğrenciye anket yapılarak elde edilmiştir. (Örneklem hacmi Meşelik Kampüsünde birinci ve ikinci öğretimde okuyan, yaklaşık 12000 öğrencinin %5'i olarak belirlenmiştir.)

SPSS paket programı kullanılarak, ileri doğru değişken seçme tekniğiyle, lojistik regresyon analizi sonucu elde edilen enbüyük olabilirlik katsayı kestirimleri ve diğer çıktı sonuçları, Tablo 1'de verilmiştir:

Tablo 1. İleriye Doğru Değişken Seçme Tekniğine Göre Analiz Sonuçları

Değişken	$\hat{\beta}_i$	$S.\hat{E}(\hat{\beta}_i)$	Wald	s.d	p	Exp($\hat{\beta}_i$)
Sabit	,7363	1,9854	,1375	1	,7108	
X1	-,2310	,0763	9,1510	1	,0025	,7938
X5	,4899	,2027	5,8405	1	,0157	1,6322
X6	,6005	,3083	3,7948	1	,0514	1,8230
X9	1,0358	,3209	10,4209	1	,0012	2,8175
X10	1,6527	,7983	4,2862	1	,0384	5,2209
X12	1,8521	,3844	23,2119	1	,0000	6,3730
X13	2,2390	1,1543	3,7622	1	,0524	9,3837

Bu sonuçlara göre, öğrencilerin sigara kullanmasındaki etkili faktörlerin, yaş, barınma şekli, babanın sigara içme durumu, alkol kullanma durumu, arkadaş çevresinin sigara kullanma durumu, sigaranın sıkıntıyı, stresi, yanlızlığı giderdiğinin ve statü kazandırdığının düşünülmesi, olduğu sonucunu ortaya koymuştur.

Dolayısıyla sınıflama için kullanılacak denklem;

$$\hat{y}_i = 0{,}7363\text{-}0{,}2310*X_1 + 0{,}4899*X_5 + 0{,}6005*X_6 + 1{,}0358*X_9 + 1{,}6527*X_{10} + 1{,}8521*X_{12} + 2{,}2390*X_{13} \quad (19)$$

olacaktır.

Çoklu doğrusal regresyonda katsayıların anlamlılığına ilişkin genel anlamlılık sınaması, F testine karşılık gelebilecek benzer bir test lojistik regresyon analizi

için geliştirilmiştir. L_0 sadece sabit terimden oluşan modelin olabilirlik değeri, L_1 elde edilen modelin olabilirlik değeri olmak üzere

$$C=-2\log(L_0/L_1)=-2(\log L_0-\log L_1) \quad (20)$$

olarak tanımlanan ölçüt (k-1) serbestlik derecesiyle Ki-kare dağılımı göstermektedir (Coşkun, 2004: 43).

Denklemin anlamlılığı için C=124,051 olarak bulunmuştur. α=0,05 ve 6 serbestlik dereceli Ki-kare tablo değeri 12,59'dan daha büyük olduğundan model, anlamlı bulunmuştur. Bu model için elde edilen sınıflandırma tablosu da, Tablo 2'de verilmiştir.

Tablo 2. İleriye Doğru Değişken Seçme Tekniğine Göre Sınıflandırma Sonuçları

Gözlemlenen	Kestirim		Doğruluk Yüzdesi(%)
	0	1	
0	58	45	56,31
1	24	173	87,82
		Tamamı	77,00

Çoklu doğrusal regresyonda, regresyon katsayılarının yorumu açıktır. Diğer bağımsız değişkenlerin değerleri aynı kalmak koşuluyla bir bağımsız değişkendeki bir birimlik değişimin bağımlı değişkende yarattığı değişim miktarını ifade eder. Oysa lojistik regresyondaki katsayı kestirimlerinin yorumu çoklu doğrusal regresyondaki gibi değildir. Sigara kullanma olasılığı kestiriminin sigara kullanmama olasılığı kestirimine oranı olan odds'lar ile yorum yapılmaktadır. Bu değerler de $e^{\hat{\beta}_i}$ sütunundaki değerlerdir. Bu durumda diğer değişkenlerin değeri, aynı kalmak koşuluyla örneğin, X_9 değişkeninin değeri bir birim arttırıldığında odds 2,8175 kat artacaktır. Diğer değişkenlerle ilgili yorumlar da aynı şekilde yapılır.

Lojistik regresyon modeliyle elde edilen sonuçların, gözlemlerin varolan gruplardan birine atanması için kullanılan diskriminant analizi sonuçlarıyla bir karşılaştırmasını yapmak amacıyla elde edilen değişken seçme tekniğiyle uygulanan diskriminant analizi katsayı kestirimleri, Tablo 3'te ve diskriminant analizi sonuçlarıyla oluşturulan sınıflandırma tablosu da, Tablo 4'te gösterilmiştir.

Tablo 3. Standartlaştırılmamış Kanonik Diskriminant Fonksiyonu Katsayıları

Değişkenler	Standartlaştırılmamış Katsayılar
X1	-0,1535463
X5	0,3219453
X6	0,4059184
X9	0,7952574
X12	1,5913482
X13	1,2474903
Sabit	-,1264290

Tablo 3'te katsayıları verilen ayırma fonksiyonu ise,

f_1=-0,1264290-0,1535463*X_1+0,3219453*X_5+0,4059184*X_6+0,7952574*X_9 +1,5913482*X_{12} +1,2474903*X_{13} (21)

şeklinde yazılır.

Tablo 4. Diskriminant Analizi Sonuçlarına Göre Sınıflandırma Tablosu

Gözlemlenen	Kestirim		Doğruluk Yüzdesi(%)
	0	1	
0	71	32	68,9
1	43	154	78,2
		Tamamı	75,00

Lojistik regresyondaki enbüyük olabilirlik kestiricisi sonuçlarına göre, toplam doğru atama yüzdesi 77 iken, bağımsız değişkenler arasında kategorik değişkenler olması nedeniyle, bu oran diskriminant analizi sonuçlarına göre 75'e düşmüştür.

IV. SONUÇ

Çalışmamızda, bağımlı değişkenin iki düzeyli, bağımsız değişkenler arasında da kategorik değişken(ler)in olduğu durumlarda, gözlemlerin gruplara atanmasında bir ayrımsama modeli olarak kullanılan ve son yıllarda diskriminant analizine alternatif olarak geniş bir uygulama alanı bulan, Lojistik Regresyon Analizi, kısaca incelenmiştir.

Günümüzde özellikle, öğrenciler arasında sigara içme oranlarında giderek bir artış sözkonusudur. Bu nedenle, sigara içmede etkili olabilecek bağımsız

değişkenler yardımıyla, Eskişehir Osmangazi Üniversitesi öğrencilerinin sigara içmelerine neden olan faktörlerin belirlenmesi için bir uygulama yapılmıştır. Yapılan analizler sonunda da lojistik regresyon analizine göre doğru sınıflandırma oranı %77 olarak bulunurken, diskriminant analizine göre doğru sınıflandırma oranı %75 olarak elde edilmiştir. (19) nolu denklem için elde edilen doğru sınıflama yüzdesi oldukça yüksektir. Ayrıca, Ki-kare testi sonucuna göre de, anlamlı olduğu tespit edilmiştir. Dolayısıyla lojistik regresyon için belirlenen (19) nolu denklem en uygun ayrımsama denklemi olarak kullanılabilecektir.

Lojistik regresyon analizi sonuçlarına göre, Eskişehir Osmangazi Üniversitesi lisans öğrencilerinin sigara içmelerindeki önemli olan değişkenlerin "yaş", "barınma şekli", "babanın sigara içmesi", "kendisinin alkol kullanması", "arkadaş çevresinin sigara kullanması", "sigaranın sıkıntı, stress ve yanlızlığı giderdiğini düşünmesi" ve "sigaranın statü kazandırdığına inanılması" olarak tespit edilmiştir. Odds oranlarına göre, öğrencilerin sigara içmelerindeki en önemli faktörün "sigaranın statü kazandırdığına inanılması" olduğu görülmüştür. Bu katsayı "sigaranın statü kazandırdığına inanmanın" sigara içme olasılığını "statü kazandırdığına inanmama" durumuna göre 9,3837 kat daha yüksek olduğunu gösterir. Bu faktörü sırasıyla, "sigaranın sıkıntı, stress ve yanlızlığı giderdiğinin düşünülmesi", "arkadaş çevresinin sigara kullanması", "kendisinin alkol kullanması", "babanın sigara içmesi" ve "barınma şekli"nin izlediği $\text{Exp}(\hat{\beta}_i)$ değerlerinden görülmektedir. Yaş değişkenine ilişkin odss değeri ise, yaş ilerledikçe sigara içmede azalma olduğunu belirtmektedir. Bu da üniversiteye başlayan öğrencilerin son sınıf öğrencilerine göre daha fazla sigara içtiğini göstermektedir. Dolayısıyla birçok ülkede olduğu gibi, bizde de sigaraya başlama yaşı 18 yaş öncesine kaymıştır.

Sigara ile mücadelede üniversitelerle birlikte, liselerin de seçilmesi uygun olacaktır. Gençlerin sigaraya karşı korunması ve bağımlılığın gelişmesini engellemek için lise yıllarında, hatta ortaokul yıllarında başlayan ve üniversite yıllarında yoğunlaşarak devam eden görsel (film, afiş vs.) ve eğitsel önlemlerin alınması ve eğitim programların uygulanması gerekmektedir. Görsel ve eğitsel programlarda da özellikle, sigaranın kişiye herhangi bir statü kazandırmadığı, sigaranın sıkıntı, stress ve yanlızlığı gidermediği anlatılmalıdır. Yine kendilerine, sigara içen arkadaşlar edinmemeleri özellikle belirtilmelidir. Ayrıca anne ve babaların da çocuklarının yanında sigara içmemeleri, alkol ve sigaranın zararları konusunda çocuklarını bilgilendirmeleri, öğrencilerin sigara içmemeleri konusunda yararlı olacaktır.

KAYNAKÇA

Abbott, R.D. (1985), "Logistic Regression in Survival Analysis**," American Journal of Epidemiology**, 121, 465-471.

Agresti, A. (1990), Analysis of Ordinal Categorical Data, **John Wiley and Sons, New York.**

Aktaş, C. ve Yılmaz V**.** (2001), "Eskişehir'de Lpg Kullanan Özel Araç Sürücülerinin Sınıflandırılımasında Lojistik Regresyon Analizi," **İstanbul Kent İçi Ulaşım Sempozyumu**, İstanbul, 251-256.

Başarır, G. (1990), Çok Değişkenli Verilerde Ayrımsama Sorunu ve Lojistik Regresyon Analizi, **Doktora Tezi (Yayımlanmamış).**

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

Blashfield, R. K. Breiman, L. and et all., (1989), "Discriminant Analysis and Clustering", **Statistical Science**, 4, 1, 34-69.

Breslow, N. E. and Day, N. E. (1980), "Statistical Methods Is Cancer Research," Vol. 1. **The Analysis Of Case-Control Studies. International Agency Of Cancer**, Lyon, France.

Bonney, G. E. (1987), "Logistic Regression for Dependent Binary Observations," **Biometrics**, 43, 951-973.

Carroll, R. J., Spiegelman, C. H., Gordon K. K., Bailey, K. T. and Abbott, R. D., (1984), "On Errors-in-Variables for Binary Regression Models", **Biometrika**, 71, 1, 19-25.

Coşkun, S., ve diğerleri (2004), "Lojistik Regresyon Analizinin İncelenmesi ve Diş Hekimliğinde Bir Uygulaması," **Cumhuriyet Üniversitesi Diş Hekimligi Fakültesi D**ergisi , Cilt:7, Sayı: 1, 42-50.

Cox, D. R. and Snell, E. J., (1970), "Analysis of Binary Data", (Second Edition), **Chapman and Hall.**

Çolak, E. ve Özdamar, K. (2004), "Ölümle Sonuçlanan Trafik Kazalarında Risk Faktörlerinin Koşullu ve Sınırlandırılmış Lojistik Regresyon Yöntemleri ile İncelenmesi," OGÜ **Tıp Fak. Dergisi**, 26, 1, 7-14.

Demirel, Y. ve Sezer, E. (2005), "Sivas Bölgesi Üniversite Öğrencilerinde Sigara Kullanma Sıklığı," **Erciyes Tıp Dergisi**, 27 (1), 1-6.

Dobson A. J., (1990), " An Introduction Generalized Linears Models", Chapman and Hall, New York.

Duffy, D. E. (1990), "On Continuity-corrected Residuals in Logistic Regression," **Biometrika**, 77, 287-293.

Duffy, D. E., Santner, T. J., (1989), "On the Small Sample Properties of Norm-Restricted Maximum Likelihood Estimators for Logistic Regression Models", Commun. **Statist.--Theory Meth**, 18, 959-980.

Gardside, P. S. and Glueck, C. J. (1995), "The Important Role of Modifiable Dietary And Behaviour Characteristic in The Causation And Prevention of Coronary Heart Disease Hospitalization and Mortality," **Journal of American College of Nutrition**, 14, 71-79.

Hosmer, D. W. and Lemeshow, S., (1989), "Applied Logistic Regression", John Wiley and Sons, New York,

İlhan, F. v.d., (2005), "Gazi Üniversitesi Tıp Faültesi Öğrencilerinin Sigara İçme Durumu," **TSK Koruyucu Hekimlik Bülteni**, 4, 4, 188-198.

Landwehr, J. M., Pregibon, D. and Shoemaker**,** A. C., (1984), "Graphical Methods for Assessing Logistic Regression Models", **Journal of American Statistical Association**, 79, 385, 61-83.

Lee, C. T. (1984), "Logistic Models for Cross-over Designs", **Biometrik**a, 71, 216-217.

Neter, J. Wasserman, W. and Kutner, M. H**.,** (1989), "Applied Linear Regression Models", (Second Edition), **Irwın, Boston.**

Pastides, H. and et all. (1985), "The Epidemiology of Fibrocystic Breast Disease," **American Journal of Epidemilogy**, 121, 440-447.

Press, S. J. and Wilson, S**.,** (1978), "Choosing Between Logistic Regression and Discriminant Analysis", **Journal of American Statistical Association**, 73, 364, 699-705.

Qu, Y., Williams, G. W., Beck, G. J. and Goormastic, M., (1987), "A Generalized Model of Logistic Regression For Clustered Data", **Commun. Statist.-Theory Meth**., 16,12, 3447-3476.

Tatlıdil, H., Başarır, G. ve V. Hökmen (1990), "Ülkelerin Sosyo Ekonomik Gelişmişliklerine Göre Kümelenmesine ve Sıralanmasına Yeni Yaklaşımlar," **Planlama Dergisi,** 26, 103-120.

Tekbaş, Ö. v.d. (2006), "Genç Erişkin Erkekler Arasinda Nikotin Bağımlılığı, Sigara İçme Sıklığı ve Bunları Etkileyen Faktörler," **TSK Koruyucu Hekimlik Bülteni**, 5 (2), 105-117.

Ünsal, A. ve Güler, H. (2005), "Türk Bankacılık Sektörünün Lojistik Regresyon Ve Diskriminant Analizi ile İncelenmesi", **VII. Ulusal Ekonometri ve İstatistik Sempozyumu,** İstanbul Ünv.

Vupa, Ö. ve Çelikoğlu, C. (2006), "Model Building in Logistic Regression Models About Lung Cancer Data", **Anadolu Ü. Bilim ve Teknoloji Dergisi**, cilt: 7, 1, 127-141.