İstatistik Araştırma Dergisi Cilt 03 No 02 Sayfa 23-38 Ağustos 2004

İKİLİ LOJİSTİK REGRESYON ANALİZİ VE BİR UYGULAMA

Özgül VUPA * Serdar KURT **

ÖZET

Regresyon yanıt değişkeni ile bir ya da birden fazla açıklayıcı değişkenler arasındaki ilişkiyi bulan veri analiz yöntemlerinden biridir. Lojistik Regresyon analizi ikili yanıt değişkeni ile hem sürekli hem de kesikli değişkenlerden oluşan bağımsız değişkenler kümesi arasındaki ilişkiyi tanımlar. Doğrusal regresyonda hata teriminin bütün gözlemler için sabit varyansla normal dağılır. Fakat yanıt değişkeni ikili olduğu zaman kullanılan lojistik regresyonda ise hata teriminin sabit varyansla dağılmadığı bazı özel durumlar ortaya çıkabilir. Bu lojistik regresyon modeli için genel kestirim yöntemi en çok olabilirliktir. En çok olabilirlik yöntemi gözlenen veri kümesini elde etmenin olasılığını maksimum yapan bilinmeyen parametrelerin değerlerini verir. Bu yöntem için ilk önce en çok olabilirlik fonksiyonunun bulunması gerekmektedir. Modeldeki herhangi bir bağımsız değişkenin önemine karar vermek için model denkleminde o bağımsız değişkenin bulunduğu ve bulunmadığı durumlardaki sapma (Deviance) değerleri karşılaştırılır. Sapma içindeki bu değişim G istatistiği olarak adlandırılır. Farklılıkların oranı (Odds Ratio, Ω) katsayılarının yorumlanması için kullanılır. İnsanlarda akciğer kanseri olmayı etkileyen birçok faktör vardır. Lojistik regresyon analizi kanser olmanın oranını azaltmak için modelde yer alan değişkenleri seçmek amacıyla kullanılmıştır. Bu uygulamada bunun sonucunun elde edilmesi için uygulanan lojistik regresyon analizinin adımları SPSS paket programı ile yapılmıştır ve hemen arkasından sonuçlar yorumlanmıştır.

Anahtar Kelimeler: *En Çok Olabilirlik, Farklılıkların Oranı (Ω) İkili Lojistik Regresyon, Olabilirlik Oran Testi,*

1. GİRİŞ

Lojistik regresyonun kullanım amacı diğer model oluşturma teknikleri ile aynıdır. Mümkün olan en az sayıda değişkeni kullanarak sonuç değişkeni ile bağımsız değişken(ler) arasındaki ilişkiyi doğru bir şekilde tanımlayan, en iyi uyuma sahip ve aynı zamanda da biyolojik olarak anlamlı bir model oluşturmaktır.

* Dokuz Eylül Üniversitesi Fen Edebiyat Fakültesi, İstatistik Bölümü, Tınaztepe Kampüsü 35160 Buca/İzmir, Türkiye, e-mail:ozgul.vupa@deu.edu.tr

Lojistik regresyon bağımlı değişkenin durumundan dolayı normal dağılıma sahip olmama ve ortak kovaryansa sahip olmama gibi çeşitli varsayımların bozulmalarına karşı literatürde çok sık kullanılan ayrımsama analizi (discriminant analysis) ve çapraz tablolara (contingency table) alternatif olarak kullanılan bir analizdir. Ayrıca bağımlı değişkenin ikili veya daha fazla düzeyini içeren durumlarda normallik varsayımının bozulması nedeni ile doğrusal regresyon analizine de alternatif olur. Lojistik regresyonda hataların binom dağılıma sahip olduğu varsayılır.

Bağımlı değişkenin binom dağılıma uyduğu durumda, lojistik regresyon modelinin kullanımı özellikle biyolojik alanlarda yaygın hale gelmiştir. Biyolojik alanda özellikle akciğer kanseri olmayı etkileyen faktörleri bulmak için yapılmış bazı çalışmalar vardır.

Lee ve arkadaşları (2000) 713 Tayvanlı kadın üzerinde yaptıkları araştırmaya göre 20 yıldan veya 40 yıldan fazla sigara içmenin 1.8 ve 2.2 odds oranları ile akciğer kanseri olma riski taşıdığını bulmuşlardır. Ayrıca bu çalışmayla Tayvan kadınlarının eşi sigara içiyorsa, bu içen kadınların içmeyenlere göre 3.3 kat daha fazla akciğer kanseri olma risklerinin olduğunu da belirtmişlerdir.

Schneider ve arkadaşları (2004) 1068 hastada yaptıkları çalışmaya göre, insanda varolan GSTM1 enziminin bozukluğu ile birlikte sigara içilmesinin insanda akciğer kanseri yapma riskini artırdığı yönünde bir çalışma elde etmişlerdir. Bu çalışmaya göre gende GSTM1 enzimini çalıştıran parçanın bozulması, insanın akciğer kanseri olmasını sigara içmeyenlere göre 158.49 kat daha fazla artırdığını belirtmişlerdir.

Darby ve arkadaşları (2004 radyoaktif bölünmenin akciğer kanseri yaptığına ilişkin bir çalışma yapmışlardır. Bu çalışmaya göre 13 Avrupa ülkesindeki 5 ile 34 yaşları arasındaki 21,356 kişide bu radyoaktif bölünmenin akciğer kanser olma riskini % 8.4 kat artırdığı yönündedir.

Vineis ve arkadaşları (2005) yaptıkları kohort çalışmasına göre 10 yıldan fazla sigara içenlerin, içmeyenlere ya da ilk 10 yılda bırakanlara göre akciğer kanseri olma risklerini daha fazla bulmuşlardır.

2. LOJİSTİK REGRESYON MODELİ

Regresyon modellerinde verilen bir bağımsız değişken değerine bağlı olarak bağımlı değişkeninin ortalama değeri $E(Y|x)$ ile gösterilir ve koşullu ortalama olarak adlandırılır. Doğrusal regresyon analizinde, koşullu ortalamanın x'in doğrusal bir denklemi olduğu varsayılır ve $E(Y|x)=\beta_0+\beta_1 x$ ile ifade edilir. Burada x'in aralığının $-\infty$ ve ∞ arasında değişmesinden dolayı $E(Y|x)$'in olası her değeri alabileceği görülür. Bağımlı değişken ikili olduğu zaman koşullu ortalama, 0 ile 1 arasında değişir ve gösterimi $0 \le E(Y|x) \le 1$ şeklindedir. Lojistik regresyon analizinde, $E(Y|x)=\beta_0+\beta_1 x$ 'in sol tarafı 0-1 arasında sınırlı olasılık değerleri alırken eşitliğin sağ tarafı sonsuz değerler alabilen bağımsız değişkenden oluşur. Bu sorunun üstesinden gelmek için olasılık değerlerinin çeşitli dönüşümlerle $-\infty$ ve ∞ arasında tanımlı hale getirilir.

$(\tilde{x}_i, y_i)$ gösterimli n tane birbirinden bağımsız gözlem eşinin olduğu varsayılır. $(i = 1,2,K,n)$ $\tilde{x}_i = (x_1, x_2, K, x_p)$ vektörü ile gösterilen p tane bağımsız değişken içersindekilerin bazıları kategorik bazıları da sürekli olabilir. Kesikli ve nominal ölçekli bağımsız değişkenleri modele dahil etmek için dizayn değişkenlerinin kullanımı gereklidir. Modeldeki bağımsız değişkenler ile bağımlı değişken arasındaki doğrusal ilişkiyi veren fonksiyona link fonksiyon adı verilir. Doğrusal regresyonda link fonksiyon birim matrisi (Identity Matrix = I) iken, lojistik regresyonda lojit ya da probit dönüşümdür (Logit or Probit Transformation). Buna göre β_p katsayısı x_p'deki bir birim artışın lojit içersinde sağlayacağı değişim demektir. Çoklu lojistik regresyon modelinin ve buradaki $\pi(\tilde{x})$'in lojit dönüşümün gösterimleri sırasıyla denklem (1) ve (2)'de verilir.

$$\pi(\tilde{x}) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \Lambda + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \Lambda + \beta_p x_p)} = \frac{\exp(g(\tilde{x}))}{1 + \exp(g(\tilde{x}))} \tag{1}$$

$$g(\tilde{x}) = \ln\left[\frac{\pi(\tilde{x})}{1 - \pi(\tilde{x})}\right] = \ln\left\{\frac{\dfrac{\exp(\beta_0 + \beta_p x_p)}{(1 + \exp(\beta_0 + \beta_p x_p))}}{\dfrac{1}{(1 + \exp(\beta_0 + \beta_p x_p))}}\right\} = \ln\left(e^{\beta_0 + \beta_p x_p}\right) = \beta_0 + \beta_p x_p$$

(2)

Lojistik regresyonda katsayıların yorumlanması için farklılıklardan (odds) ve farklılıkların oranı'ndan (odds ratio) yararlanılır. Farklılıklar, lojit'in doğal logaritması alınmamış halidir. $x = 1$ ve $x = 0$ için farklılıkların değerleri denklem (3)'de verilir.

$$\frac{\pi(x=1)}{(1-\pi(x=1))}, \qquad \frac{\pi(x=0)}{(1-\pi(x=0))} \tag{3}$$

Farklılıkların oranı ise $x = 1$ için hesaplanan farklılıkların değerinin $x = 0$ için hesaplanan farklılıkların değerine oranı şeklindedir. Farklılıkların oranının doğal logaritması log odds oranı veya log odds şeklinde ifade edilir ve bu da lojit farka eşittir. Buna göre lojistik regresyonda bağımsız değişkenin ikili olması yani 0 ile 1 şeklinde kodlanması durumunda farklılıkların oranı $\Omega = \exp(\beta_p)$ şeklinde ifade edilir. Farklılıkların oranının tahmini, $\hat{\Omega}$, eğik bir dağılıma sahiptir. Örnek genişliği yeteri kadar büyük olduğu zaman $\hat{\Omega}$'nın dağılımı normal olur. Farklılıkların oranı, lojit fark ve farklılıkların oranı için $\%100(1-\alpha)$ güven aralığının tahmini sırasıyla denklem (4), (5) ve (6)'da verilir.

$$\hat{\Omega} = \frac{\dfrac{\pi(1)}{(1-\pi(1))}}{\dfrac{\pi(0)}{(1-\pi(0))}} = \exp(\hat{\beta}_p) \tag{4}$$

$$\ln(\hat{\Omega}) = \hat{\beta}_p \tag{5}$$

$$\exp\left[\hat{\beta}_p \pm Z_{1-\alpha/2}\, S\hat{E}(\hat{\beta}_p)\right] \tag{6}$$

Modelde ikiden fazla sınıflı bağımsız değişkenin olduğu durumda farklılıkların oranının hesaplanması aynı iki sınıflı bağımsız değişkendeki gibi lojit fark ile yapılır. Modelin sürekli bağımsız değişken içermesi durumunda tahmin edilen katsayıların nasıl yorumlanacağı değişkenin modele nasıl gireceğine bağlıdır. Bunun için kartil bölünme ya da x(log(x)) değişkenini modele ekleme yöntemleri kullanılır. Kartil bölünme yönteminde kartiller kullanarak bağımsız değişken dört gruba ayrılır ve bu gruplar küçükten büyüğe ya da büyükten küçüğe doğru sıralanır. Bunlar modele dizayn değişkeni şeklinde girer. Grupların farklılıkların oranları arasında doğrusal bir artış veya azalış varsa, incelenen sürekli bağımsız değişken lojitle doğrusal olduğu varsayılır ve modele sürekli değişkenmiş gibi girer.

Bağımsız değişkenler sayısal olarak sınıflandırıldığı zaman çeşitli dizayn değişkenlerinin kategorik olan bu değişkenleri temsil etmesi için kullanılması gerekir. Nominal değişken k kategoriye sahipse, o zaman $k-1$ dizayn değişkeni kullanılır. Eğer j'inci bağımsız değişken olarak ifade edilen x_j, k_j kategoriye sahipse $k_j - 1$ dizayn değişkeni D_{ju} ve katsayılarında $u = 1,2,K,k_j - 1$ olarak belirtilir. j'inci bağımsız değişkeni kesikli olan p değişkenli model için lojit biraz daha farklı olur ve gösterimi denklem (7)'deki gibidir.

$$g(\tilde{x}) = \beta_0 + \beta_1 x_1 + K + \sum_{m=1}^{k_j-1} \beta_{jm} D_{jm} + \beta_p x_p \tag{7}$$

(1)'deki lojistik regresyon modelindeki bilinmeyen parametrelerin tahmin edilmesi en çok olabilirlik yöntemi ile (Maximum Likelihood Method) yapılır. Bu yöntem gözlenen veri kümesini elde etme olasılığını maksimum yapan bilinmeyen parametrelerin değerlerini verir. Bunun için en çok olabilirlik fonksiyonunun (Maximum Likelihood Function) oluşturulması gerekir. Sonuç değişkeninin 1'e eşit olduğu zaman olabilirlik fonksiyonuna katkısı $\pi(\tilde{x})$, 0'a eşit olduğu zamanki katkısı ise $1-\pi(\tilde{x})$'dir. $(\tilde{x}_i, y_i)$ çiftindeki birbirinden bağımsız olabilirlik fonksiyonunun gözlem eşleri çarpımla ifade edilir ve gösterimi aşağıdaki gibidir.

$$L(\beta_0, \beta_p) = \prod_{i=1}^{n} \pi(\tilde{x}_i)^{y_i} (1 - \pi(\tilde{x}_i))^{1-y_i} \tag{8}$$

En çok olabilirlik yöntemindeki amaç $\tilde{\beta}$ kestiriminin (8)'deki denklemi maksimum yapmasıdır. (8)'deki denklemin logaritmasıyla çalışmak matematiksel

olarak daha kolay olacağından log olabilirlik fonksiyonu denklem (9)'daki gibi elde edilir.

$$\ln L(\beta_0,\beta_1,K,\beta_p)=\ln L(\tilde{\beta})=\sum_{i=1}^{n}\{y_i(\beta_0+\beta_1x_{1i}+K+\beta_px_{pi})-\ln(1+\exp(\beta_0+\beta_1x_{1i}+K+\beta_px_{pi}))\} \quad (9)$$

Yukarıdaki denklemi maksimum yapan $\tilde{\beta}$ değerlerinin bulabilmek için $\ln L(\tilde{\beta})$'nın $\tilde{\beta}$'lara göre başka bir deyişle $p+1$ katsayıya göre türevi alınarak bu $p+1$ tane olabilirlik eşitlikleri 0'a eşitlenir. Sonuçta elde edilen eşitliklere olabilirlik eşitlikleri denir ve denklem (10) ile ifade edilirler.

$$\sum_{i=1}^{n}[y_i-\pi(x_i)]=0, \qquad \sum_{i=1}^{n}x_{ij}[y_i-\pi(x_{ij})]=0 \qquad j=1,2,K,p \quad (10)$$

Bu denklemlerden elde edilen $\tilde{\beta}$ değerleri, en çok olabilirlik kestirimi olarak adlandırılırlar ve $\hat{\tilde{\beta}}$ ile gösterilir. $\pi(\tilde{x}_i)$'nin en çok olabilirlik kestirimi $\hat{\pi}(\tilde{x}_i)$ ile gösterilir ve bu değer x'in x_i gibi değere eşit olarak verildiği zaman, Y'nin 1'e eşit olma koşullu olasılığının kestirimini verir. Burada kestirilen katsayıların varyans ve kovaryanslarının kestirim yöntemi log olabilirlik fonksiyonlarının ikinci derecede kısmi türevlerinden oluşan matristen elde edilir. Ama bunun elde edilmesi uzun ve karışık işlemler gerektirdiğinden paket programlardan yararlanılır. (SPSS, Minitab, SAS)

3. LOJİSTİK REGRESYON KATSAYILARININ ÖNEM TESTİ

Katsayıların kestiriminden sonra modeldeki değişkenlerin önemliliğine bakılır. Katsayıların önemlilik testleri en iyi modeli mümkün olan en az değişkenle oluşturmada yardımcıdır.

Katsayıların önemi olabilirlik oran testi (Likelihood Ratio Test), Wald testi ve score testi olmak üzere üç farklı yöntemle yapılabilir. Buradaki asıl sorun incelenecek olan değişkeni kapsayan modelin (Full Model) sonuç değişkeni hakkında o değişkeni kapsamayan modelden (Saturated Model) daha çok bilgi içerip içermediğidir. Bu sorun sonuç değişkeninin gözlenen değerlerini, her iki modelden elde edilen kestirilen değerlerle karşılaştırılarak cevaplanır. Eğer değişkenli modelin kestirilen değerleri değişkeni içermeyen modelden daha iyi ise o zamam incelediğimiz değişkenin önemli olduğu sonucuna varırız. Bu karşılaştırma işlemi log olabilirlik fonksiyonu ile yapılır ve gösterimi denklem (11)'deki gibidir.

$$D=-2\ln\left[\frac{\text{şu andaki modelin olabilirliği}}{\text{doymuş modelin olabilirliği}}\right] \quad (11)$$

Burada parantez içindeki ifade olabilirlik oranı olarak ifade edilirken doğal logaritmanın (–2) katının alınması ile de dağılımı bilinen bir değer elde edilir ve bu da hipotez testinde kullanılır. Bu eşitlik log olabilirlik cinsinden yazılacak olursa denklem (12) elde edilir.

$$D = -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\hat{\pi}(\tilde{x}_i)}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\hat{\pi}(\tilde{x}_i)}{1-y_i}\right)\right] \tag{12}$$

Sapma (Deviance) olarak adlandırılan D istatistiği doğrusal regresyondaki hata kareler toplamı ile aynı rolü üstlenmesinin yanında uyum iyiliğine karar verilirken de kullanılır. Bağımsız bir değişkenin önemine karar vermede, model denkleminde bu değişkenin olduğu ve olmadığı D değerleri karşılaştırılır. D'deki bu değişim, doğrusal regresyonda kullanılan F testindeki pay kısmı ile aynı rolü üstlenen G istatistiği olarak adlandırılır ve gösterimi denklem (13)'deki gibidir.

$$G = -2\ln\left[\frac{\text{değeğişkenz modelin olabilirliği}}{\text{değeğişkenmodelin olabilirliği}}\right] \tag{13}$$

Modelde tek bağımsız değişken varsa, değişkenin modelde olmadığı zamanki β_0'ın maksimum olabilirlik tahmini ve G istatistiği sırasıyla denklem (14), (15) ve (16)'da verilir.

$$\hat{\beta}_0 = \ln\frac{n_1}{n_0}, \qquad n_1 = \sum_{i=1}^{n} y_i, \qquad n_0 = \sum_{i=1}^{n}(1-y_i), \tag{14}$$

$$G = -2\ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n}\hat{\pi}_i^{\,y_i}\left(1-\hat{\pi}_i\right)^{(1-y_i)}}\right] \tag{15}$$

$$G = 2\left\{\sum_{i=1}^{n}[y_i \ln(\hat{\pi}_i) + (1-y_i)\ln(1-\hat{\pi}_i)] - [n_1\ln(n_1) + n_0\ln(n_0) - n\ln(n)]\right\} \tag{16}$$

$H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$ hipotezi altında, G istatistiği 1 serbestlik dereceli ki-kare dağılımına sahiptir. Eğer hesaplanan G değeri, ki-kare değerinden büyükse hipotezi reddederiz. Başka bir yol olarak incelenen değişkeni içeren ve içermeyen modelin log olabilirlik değerine bakılır. İncelenen değişkenin modele eklenmesi log olabilirlik

değerinde artışa neden olur. Log olabilirlik oran testi her iki modelin log olabilirlik değerleri arasındaki farkın (–2) katına eşittir. Çoklu lojistik regresyon modelinde birlikte değişenler için p tane eğim katsayısının sıfıra eşit olması hipotezi altında G istatistiği $(\nu_2 - \nu_1)$ serbestlik derecesiyle ki-kare dağılımı gösterir (ν_2= tüm modeldeki değişken sayısından 1 fazla, ν_1= indirgenmiş modeldeki değişken sayısından 1 fazla). $(\nu_2 - \nu_1)$ serbestlik derecesinde bulunan yanılma olasılığı 0.05'den büyük olursa, indirgenmiş model tüm model kadar iyidir. Burada dikkat edilmesi gereken nokta kategorik olarak ölçeklendirilmiş bağımsız değişken modele girdiği ya da modelden çıktığı zaman onun bütün dizayn değişkenlerinin eklenmesi ya da çıkartılmasıdır.

Wald testi eğim parametresinin en çok olabilirlik kestirimiyle onun standart hatasının kestiriminin karşılaştırılması ile elde edilir. Elde edilen oran (W), $H_0 : \beta_1 = 0$ hipotezi altında standart normal dağılım gösterir. Wald testinin çok değişkenli olduğu durumdaki karşılığı vektör-matris hesaplamalarından elde edilir. p+1 katsayının her birinin sıfıra eşit olması hipotezi altında W istatistiği p+1 serbestlik derecesiyle ki-kare dağılır ve gösterimleri sırasıyla denklem (17) ve (18)'deki gibidir.

$$W = \frac{\hat{\beta}}{S\hat{E}\left(\hat{\beta}\right)} \tag{17}$$

$$W = \hat{\beta}'\left[\sum \hat{\beta}\right]^{-1}\hat{\beta} \tag{18}$$

Score testi ise log olabilirliğin türevlerinin dağılım teorisine bağlıdır. Yani $\tilde{\beta}$'ya göre $L(\tilde{\beta})$'nın p tane türevinin koşullu dağılımı üzerine kurulmuştur. Score testi aslında matris hesapları gerektiren çok değişkenli bir test olup standart normal dağılım gösterir ve hesaplanışı denklem (19)'daki gibidir.

$$ST = \frac{\sum_{i=1}^{n} x_i(y_i - \bar{y})}{\sqrt{\bar{y}(1-\bar{y})\sum_{i=1}^{n}(x_i - \bar{x})^2}} \tag{19}$$

4. LOJİSTİK REGRESYON ANALİZİNDE DEĞİŞKEN SEÇİMİ

Bağımsız değişken sayısı az olduğunda model oluşturulması kolay olmaktadır. Ancak değişken sayısı artıkça modele girecek değişkenlerin seçilmesi ve oluşturulması zorlaşır. Modele ne kadar çok değişken girerse tahmin edilen standart hata artar ve gözlenen veri kümesine daha çok bağımlı hale gelir. Lojistik regresyonda değişken

seçimimde uygulanan bazı yöntemler vardır. Bunlar basit lojistik regresyonla katsayıların tek tek testi (Univariate Analysis), adımsal lojistik regresyon (Stepwise Logistic Regresyon) ve en iyi alt küme (Best Subsets) olmak üzere üç tanedir.

Değişken seçme işlemine her bir değişken için ayrı ayrı tek değişkenli basit lojistik regresyon analizi yapılır. Olabilirlik oran ki-kare testi $k-1$ serbestlik derecesiyle, tek bağımsız değişkeni kapsayan lojistik regresyon modelindeki $k-1$ dizayn değişkeninin katsayılarının önemi için olabilirlik oran testinin değeriyle tamamen birbirine eşittir Orta derecede ilişki gösteren değişkenler için düzeylerden birini referans grup olarak kullanarak bireysel farklılıkların oranlarını güven aralıklarıyla birlikte kestirim de yapılabilir. Eğer gözlenen frekanslar içinde sıfırlı hücre ya da hücreler varsa onlara dikkat edilmelidir. Bu durum farklılıkların oranlarının biri için tek değişkenli nokta tahminini sıfır ya da sonsuz yapar. Bunu önlemek için bağımsız değişkenin kategorileri birleştirilir, iptal edilir ya da eğer değişken sıralı ölçekliyse sürekliymiş gibi modele alınır. Sürekli değişkenler için tek değişkenli lojistik uyuma karar verilmesi için lojit ile doğrusallığına bakılır. Tek değişkenli analizlerin tamamlanmasından sonra çok değişkenli analiz için değişkenler seçilir. Olasılık değeri 0.25'den küçük olan değişkenler çok değişkenli model için aday değişkenlerdir. Değişkenler tamamlandıktan sonra, bu kriterle seçilen değişkenlerin tümünü içeren model oluşturulur. Şüpheli değişkenlerin modele dahil edilmesinin mümkün olabilmesi için önemlilik düzeyi yeterince büyük seçilir. Çok değişkenli modelin kurulmasıyla, modele dahil edilen her bir değişkenin önemi doğrulanır. Bu Wald testi ve olabilirlik oran testi ile yapılır. Bu kriterlere göre modele katkıda bulunmayan değişkenler model dışı bırakılarak kalan değişkenlerle yeni bir model kurulur. Yeni model olabilirlik oran testi kullanarak eski modelle karşılaştırılır. Özellikle katsayıları önemli derecede değişen değişkenler üzerinde durulur. Bu durum model dışı kalan değişkenler açısından önem taşır. En son aşamada modele giren sürekli değişkenlerin doğru ölçekle girip girmediğine bakılır.

Eğer değişken sayısı fazla ise adımsal lojistik regresyon analizine başvurulabilir. Bu analiz ileriye doğru seçim (Forward Selection) ve geriye doğru eleme (Backward Elimination) yöntemleri olmak üzere iki tanedir.

5. UYGULAMA

Bu çalışmada bağımlı değişken kanser olma (Ca) ya da olmama (Co) durumudur. İnsanlarda kanser olmayı etkileyen bir çok faktör vardır. Lojistik regresyon analizi için yarısı kanserli (Ca) yarısı da kontrol grubundan (Co) oluşan 1200 hastalık veri seti, İzmir ilindeki Ege Ünv. Tıp Fakültesi Göğüs Hastalıkları Anabilim Dalından elde edilmiştir. Bu veriler içinde kanser olmayı etkileyen 7 tane bağımsız değişken vardır. Bunlardan yaş (YAS) değişkeni sürekli iken cinsiyet (C), eğitim (EGT), sigara içilen yıl (SIY), sigaraya başlama yaşı (SBY), bir yılda içilen paket sayısı (YIPS) ve sigarayı bıraktığı süre (SBS) kategorik değişkenlerdir. Aşağıdaki tabloda kategorik olan değişkenlerin kodlaması verilir.

Tablo 1 . Kategorik değişkenlerin kodlanması

EGT	Okur yazar değil	1	**YIPS**	**İçmemiş**	**Referans (0)**
	İlkokul	2		1 ile 10 paket arasında	1
	Ortaokul	3		11 ile 20 paket arasında	2
	Lise ya da Üniversite	**Referans (0)**		21 ile 30 paket arasında	3
SIY	**İçmemiş**	**Referans (0)**		30 paketten fazla	4
	20 yıldan az	1	**SBS**	İçiyor	1
	21 ile 30 yıl arasında	2		1 ile 5 yıl arasında	2
	31 ile 40 yıl arasında	3		6 ile 11 yıl arasında	3
	40 yıldan fazla	4		11 yıldan fazla	4
SBY	**Başlamamış**	**Referans (0)**		**İçmiyor**	**Referans (0)**
	10 yıldan az	1	C	**Erkek**	**Referans (0)**
	11 ile 15 yıl arasında	2		Kadın	1
	16 ile 19 yıl arasında	3			
	20 yıldan fazla	4			

İlk olarak çok değişkenli lojistik regresyon modeline girecek olan değişkenleri belirlemek amacıyla, aday değişkenlerin her biri için tek tek yapılan tek değişkenli basit lojistik regresyon analizi Tablo 2'de verilir.

Tablo 2. Değişkenlerin tekli lojistik regresyon analizi

Değişken	$\hat{\beta}$	StdHata	Wald	sd	p-değeri	Exp ($\hat{\beta}$)	Güven Aralığı	G	p-değeri
C (1)	-0.334	0.275	**1.480**	1	**0.224 ***	0.716	**0.418-1.227**	1.498	**0.221**
EGT			**37.420**	3	**0.000 ***	8.958		53.819	**0.000**
1	2.193	0.391	**31.477**	1	0.000	8.127	**4.165-19.270**		
2	2.095	0.384	**29.746**	1	0.001	4.448	**3.828-17.256**		
3	1.492	0.444	**11.302**	1	0.000	1.050	**1.863-10.617**		
YAS	0.049	0.006	**59.434**	1	**0.000 ***	1.050	**1.037-1.063**	65.135	**0.000**
SIY			**196.073**	4	**0.000 ***			273.580	**0.000**
1	0.789	0.357	**4.879**	1	0.027	2.201	**1.093-4.432**		
2	1.876	0.261	**51.600**	1	0.000	6.527	**3.912-10.888**		
3	2.657	0.250	**112.588**	1	0.000	14.258	**8.727-23.293**		
4	3.001	0.247	**147.959**	1	0.000	20.110	**12.399-32.616**		
SBY			**134.226**	4	**0.000 ***			202.272	**0.000**
1	3.043	0.300	**103.108**	1	0.000	20.978	**11.658-37.748**		
2	2.666	0.248	**115.263**	1	0.000	14.375	**8.831-23.385**		
3	2.405	0.267	**81.264**	1	0.000	11.081	**6.568-18.693**		
4	2.167	0.245	**78.261**	1	0.000	8.731	**5.402-14.111**		
YIPS			**231.148**	4	**0.000 ***			312.626	**0.000**
1	0.914	0.408	**5.016**	1	0.025	2.495	**1.121-5.552**		
2	0.890	0.347	**6.588**	1	0.010	2.436	**1.234-4.808**		
3	1.780	0.266	**44.721**	1	0.000	5.928	**3.519-9.987**		
4	2.989	0.236	**160.060**	1	0.000	19.861	**12.500-31.556**		
SBS			**138.993**	4	**0.000 ***			206.137	**0.000**
1	2.635	0.234	**126.814**	1	0.000	13.943	**8.814-22.056**		
2	2.499	0.286	**76.389**	1	0.000	12.176	**6.951-21.326**		
3	1.951	0.346	**31.895**	1	0.000	7.038	**3.576-13.853**		
4	1.689	0.300	**92.992**	1	0.000	5.414	**3.009-9.740**		

Olabilirlik oran testi sonucunda p değeri 0.25'den küçük olan değişkenler çok değişkenli modele girecek olan aday değişkenlerdir. Buna göre bütün değişkenler çoklu lojistik regresyon analizi için aday değişkenlerdir. Sadece cinsiyet değişkeninin önem düzeyi 0.25'e yakındır. Zaten bu değişkenin Wald istatistik değerinin küçük olmasından da rahatlıkla görülebilir. Tüm değişkenlerin modele girdiği çoklu lojistik regresyon analizi Tablo 3'de kurulur. Tablo 3'deki analize göre SBY değişkeni ile SBS değişkeninin p değerleri 0.10 değerinden büyük olmasından dolayı istatistiksel olarak bir öneme sahip olmadığı sonucuna varılır ve incelenmesi için o değişkenleri içeren ve içermeyen modeller olabilirlik oran test istatistiği ile karşılaştırılır. Bunun sonucunda bu iki değişken model içinde kalır ya da model dışı bırakılır. Bunların gösterimi sırasıyla tablo 4 ve tablo 5'de gösterilir.

Tablo 3. Çoklu lojistik regresyon analizi

Değişken	$\hat{\beta}$	StdHata	Wald	sd	p-değeri	Exp ($\hat{\beta}$)	Güven Aralığı	G	p-değeri
C(1)	1.892	0.407	**21.568**	1	**0.000**	6.631	2.984-14.735	411.411	**0.000 ***
EGT			**15.255**	3	**0.002**				
1	1.557	0.439	12.566	1	0.000	4.747	2.006-11.230		
2	1.660	0.426	15.194	1	0.000	5.258	2.282-12.112		
3	1.576	0.500	9.926	1	0.000	4.834	1.814-12.882		
YAS	0.060	0.11	**29.223**	1	0.000	1.062	1.039-1.086		
SIY				4	**0.000**				
1	2.605	0.567	21.071	1	0.000	13.527	4.448-41.132		
2	3.054	0.471	41.979	1	0.000	21.203	8.417-53.413		
3	2.237	0.421	28.232	1	0.000	9.367	4.104-21.378		
4	1.857	0.465	15.918	1	0.000	6.402	2.572-15.939		
SBY			**8.755**	3	**0.033**				
1	0.711	0.272	6.864	1	0.009	2.037	1.196-3.469		
2	0.437	0.187	5.464	1	0.019	1.549	1.073-2.235		
3	0.289	0.214	***1.822*** **	1	**0.177** **	1.335	**0.877-2.032**		
YIPS			**39.350**	3	**0.000**				
1	-1.907	0.492	15.028	1	0.000	0.149	0.057-0.390		
2	-1.805	0.366	24.382	1	0.000	0.164	0.080-0.337		
3	-1.524	0.301	25.609	1	0.000	0.218	0.121-0.393		
SBS			**29.468**	3	**0.000**				
1	1.374	0.293	21.918	1	0.000	3.949	2.222-7.019		
2	1.059	0.331	10.273	1	0.000	2.884	1.509-5.513		
3	0.246	0.386	***0.406*** **	1	**0.524** **	1.279	**0.600-2.723**		
Sabit	-7.960	0.853	**87.070**	1	**0.000**	0.000			
-2LL=1252.142									

Tablo 4 . SBY Değişkenini içermeyen çoklu lojistik regresyon analizi

Değişken	$\hat{\beta}$	StdHata	Wald	sd	p-değeri	Exp ($\hat{\beta}$)	Güven Aralığı	G	p-değeri
C(1)	1.866	0.406	**21.104**	1	**0.000**	6.463	2.915-14.330	402.575	**0.000 ***
EGT			**15.080**	3	**0.002**				
1	1.568	0.437	12.842	1	0.000	4.796	2.035-11.306		
2	1.651	0.425	15.069	1	0.000	5.210	2.264-11.990		
3	1.559	0.500	9.718	1	0.002	4.754	1.784-12.669		
YAS	0.051	0.010	**23.622**	1	**0.000**	1.052	1.031-1.074		
SIY			**53.404**	4	**0.000**				
1	2.736	0.563	23.635	1	0.000	15.430	5.120-46.503		
2	3.261	0.462	49.835	1	0.000	26.065	10.542-64.449		
3	2.546	0.404	39.614	1	0.000	12.751	5.771-28.171		
4	2.369	0.430	30.394	1	0.000	10.691	4.605-24.821		
YIPS		-	**40.473**	3	**0.000**				
1	-1.967	0.490	16.090	1	0.000	0.140	0.054-0.366		
2	-1.854	0.366	25.653	1	0.000	0.157	0.076-0.321		
3	-1.520	0.300	25.654	1	0.000	0.219	0.121-0.394		
SBS			**26.347**	3	**0.000**				
1	1.248	0.285	19.162	1	0.000	3.483	1.992-6.091		
2	0.925	0.325	8.117	1	0.004	2.523	1.335-4.768		
3	0.186	0.384	***0.236*** **	1	0.627	1.205	0.568-2.556		
Sabit	-7.342	0.806	**82.947**	1	**0.000**	0.001			
-2LL=1260.978									

SBY değişkenini içeren ve içermeyen modelleri karşılaştıran olabilirlik oran test istatistiği ve serbestlik derecesi sırasıyla $G=[1260.978-1252.142]=8.836$ ve $(v_{tüm}-v_{indirgenmiş})=19-16=3$ şeklinde hesaplanır. Hesaplanan bu G istatistiği 3 serbestlik dereceli ki-kare dağılımından büyük olduğu için $(\chi^2_{3,0.95}=7.81)$, SBY değişkeninin modelde kalmasında bir sakınca yoktur.

Tablo 5 . SBS Değişkenini içermeyen çoklu lojistik regresyon analizi

Değişken	$\hat{\beta}$	StdHata	Wald	sd	p-değeri	Exp ($\hat{\beta}$)	Güven Aralığı	G	p-değeri
C(1)	1.848	0.401	**21.218**	1	**0.000**	6.348	2.891-13.937	381.068	**0.000** *
EGT			**17.120**	3	**0.001**				
1	1.605	0.433	13.761	1	0.000	4.979	2.132-11.627		
2	1.729	0.420	16.986	1	0.000	5.635	2.476-12.822		
3	1.615	0.494	10.704	1	0.001	5.026	1.910-13.221		
YAS	0.033	0.009	**12.375**	1	**0.000**	1.034	1.015-1.053		
SIY			**101.927**	4	**0.000**				
1	3.010	0.540	31.104	1	0.000	20.294	7.046-58.456		
2	3.739	0.433	74.479	1	0.000	42.073	17.996-98.456		
3	3.273	0.341	92.320	1	0.000	26.391	13.536-51.453		
4	3.208	-0.353	82.451	1	0.000	24.730	12.374-49.426		
SBY			**5.329**	3	**0.149**				
1	0.553	0.264	4.377	1	0.036	1.738	1.036-2.916		
2	0.294	0.182	2.623	1	0.105	1.342	0.940-1.915		
3	0.285	0.211	1.824	1	0.177	1.330	0.879-2.012		
YIPS			**34.079**	3	**0.000**	0.164	0.063-0.426		
1	-1.809	0.488	13.756	1	0.000	0.187	0.093-0.377		
2	-1.676	0.358	21.983	1	0.000	0.260	0.146-0.466		
3	-1.346	0.297	20.579	1	0.000	1.034	1.015-1.053		
Sabit	-6.315	0.746	**71.584**	1	**0.000**	0.002			
-2LL=1282.485									

Tablo 5'de SBS değişkenini içeren ve içermeyen modelleri karşılaştıran olabilirlik oran test istatistiği ve serbestlik derecesi sırasıyla $G=[1282.485-1252.142]=30.343$ ve $(\nu_{tüm}-\nu_{indirgenmiş})=19-16=3$ şeklinde hesaplanır. Hesaplanan bu G istatistiği 3 serbestlik dereceli ki-kare dağılımından büyük olduğu için $(\chi^2_{3,0.95}=7.81)$, SBS değişkeninin de modelde kalmasında bir sakınca yoktur.

Sorunlu olan bu iki bağımsız değişkenin modelde kalmasında sakınca bulunmadıktan sonra modelde bulunan sürekli değişkenin lojit ile doğrusal bir ilişki içinde olup olmadığını ve bu arada modele doğru ölçekle girip girmediğine bakılması gerekir. Bunun için kartil bölünme yöntemi kullanılır. Buna göre elde edilen yeni modelin analizi Tablo 6'da verilir.

Tablo 6 : Sürekli olan YAS değişkeni için çoklu lojistik regresyon analizi

Değişken	$\hat{\beta}$	StdHata	Wald	sd	p-değeri	Exp ($\hat{\beta}$)	Güven Aralığı	G	p-değeri
C (1)	1.740	0.408	**18.198**	1	**0.000**	5.697	2.561-12.670	418.023	**0.000 ***
EGT			**17.003**	3	**0.001**				
1	1.763	0.442	15.936	1	0.000	5.830	2.453-13.854		
2	1.754	0.429	16.714	1	0.000	5.776	2.492-13.390		
3	1.620	0.505	10.307	1	0.001	5.054	1.880-13.590		
YAS			**36.168**	3	**0.000**				
1	1.120	0.221	25.651	1	0.000	**3.063**	1.986-4.724		
2	1.449	0.256	31.99	1	0.000	**4.259**	2.576-7.041		
3	1.331	0.284	21.999	1	0.000	**3.786**	2.171-6.603		
SIY			**47.106**	4	**0.000**				
1	2.537	0.566	20.089	1	0.000	12.640	4.168-38.327		
2	3.097	0.468	43.726	1	0.000	22.130	8.837-55.415		
3	2.224	0.420	28.009	1	0.000	9.243	4.056-21.063		
4	1.976	0.467	17.921	1	0.000	7.217	2.890-18.019		
SBY			**8.726**	3	**0.033**				
1	0.716	0.269	7.080	1	0.008	2.047	1.208-3.470		
2	0.420	0.186	5.072	1	0.024	1.521	1.056-2.192		
3	0.303	0.217	1.948	1	0.163	1.354	0.885-2.071		
YIPS			**36.454**	3	**0.000**				
1	-1.909	0.492	15.031	1	0.000	0.148	0.056-0.389		
2	-1.711	0.367	21.748	1	0.000	0.181	0.088-0.371		
3	-1.444	0.299	23.263	1	0.000	0.236	0.131-0.424		
SBS			**25.453**	3	**0.000**				
1	1.213	0.284	18.268	1	0.000	3.364	1.929-5.867		
2	0.920	0.324	8.062	1	0.005	2.509	1.330-4.735		
3	0.135	0.382	0.125	1	0.723	1.145	0.542-2.420		
Sabit	-5.428	0.556	**95.449**	1	**0.000**	0.004			
-2LL=1245.530									

Eğer lojit, YAS değişkeni ile doğrusal ise, bu durumda kategorik olarak sınıflandırılmış YAS değişkeninin farklılıkların oranının değerlerinde doğrusal artan ya da azalan bir eğilimim olması beklenir. Burada böyle bir ilişki görülmediğinden YAS değişkenin modele sürekli olarak girmesinde sakınca bulunmaz. (3.063, 4.259, 3.786)

6. SONUÇ

Sonuç olarak bu model tüm bağımsız değişkenleri içerir. Son modelimiz Tablo 3'deki gibidir. Değişkenlerin yorumu farklılıkların oranlarına bakılarak yapılır. Cinsiyeti kadın olanların erkeklere göre kanser olma riski 6.631 kat daha fazladır. EGT değişkenine bakıldığında sırasıyla okur-yazar olmayanların, ilkokul ve ortaokul mezunu olanların, lise ve üniversite mezunlarına göre kanser olma riski 4.747, 5.258 ve 4.834 kat daha fazladır. YAS değişkeni için de yaştaki 10 birimlik artış kanser olma riskini %0.06 oranında artırır. Sigara içme yılındaki artışın kanser olmayı artırdığı SIY değişkeninin farklılıkların oranlarına bakılarak anlaşılır. Sigara içme yılına bakıldığında sırasıyla 21 yıldan daha az sigara içenlerin, 21 ile 30 yıl arasında sigara içenlerin, 31 ile 40 yıl arasında sigara içenlerin ve 40 yıldan daha fazla sigara içenlerin sigara içmeyenlere göre kanser olma riski 13.527, 21.203, 9.367 ve 6.402 kat daha fazladır. Sigaraya başlama yaşı ne kadar küçük olursa kanser olma riski o kadar artar. Bu SBY değişkenindeki farklılıkların oranı değerlerinin 1.335'den 2.037'e çıkmasıyla da görülür. Yıl içersinde harcanan paket sayısının artmasıda kanser olmayı artıran bir etken olduğu kolaylıkla tespit edilebilir.

Değişken seçiminde göz önüne alınan birçok kriter vardır. Bunların teker teker incelenmesi değişken seçimi açısından önemlidir. Bu çalışmada model yapılandırmada değişken seçim prosedürlerine yer verildi ve modele giren yedi bağımsız değişkeninde önemli olduğu sonucuna varıldı. Ayrıca daha önce akciğer kanseri olmayı etkileyen faktörler başka araştırmacılar tarafından da ele alınmıştır. Bu araştırmacılardan Lee ve arkadaşları 713 Tayvanlı kadın üzerinde yaptıkları araştırmada sigara içmenin akciğer kanseri olmada etkili olduklarını bulmuşlardı. Bizim araştırmamızda da bunu destekleyen sonuçlar elde edilmiştir. Lee odds değerlerini 1.8 ve 2.2 olarak bulurken bu çalışmada yaklaşık olarak 6.4 ve 21.2 olarak bulunmuştur. Schneider ve arkadaşları ise GSTM1 enziminin gendeki bozukluğunun sigara içme ile etkileşim oluşturduğunun ve bunun da akciğer kanseri olmada etkili olduğunu bulmuşlardır. Bizim uygulamamızda ise genlerle ilgili bir çalışma yer almamaktadır. Vineis ve arkadaşlarının yaptıkları çalışmaya göre 10 yıldan fazla sigara içenlerin sigara içmeyenlere ya da 10 yıldan az sigara içenlere göre akciğer kanseri olma risklerinin daha fazla olduğunu bulmuşlardır. Bizim çalışmamızda da bu verilenleri desteklemektedir.

KAYNAKLAR

DARBY S., HILL D., AUVINEN A., DIOS J., BAYSSON H. and so on (2004), *Radon in Homes and Risk of Lung Cancer : Collaborative Analysis of Individual Data from 13 European Case Control Studies, BMJ.*

ELHAN A., (1997), *Lojistik Regresyon Analizinin İncelenmesi ve Tıpta Bir Uygulaması,* Ankara Üniversitesi Sağlık Bilimleri Enstitüsü, Yüksek Lisans Tezi.

GROUVEN U. & BENDER R., (1998), *Using Binary Logistic Regression Models for Ordinary Data with Non-proportional Odds.* J. Clin. Epidemiol, 51,809-816.

HARRELL F., (2001), *Regression Modeling Strategies. Springer- Verlag* ,New York.

HOSMER D. & LEMESHOW S., (1989), *Applied Logistic Regression. John Wiley & Sons.*

KLEINBAUM D., (1994), *Logistic Regression A Self-Learning Text. Springer- Verlag* New York.

LEE C., KO Y., GOGGINS W., HUANG J. And so on (2000), *Lifetime Environmental Exposure to Tobacco Smoke and Primary Lung Cancer of Non-Smoking Taiwanese Women,* International Journal of Epidemiology, 29,224-231.

MENDENHALL W. & SINCIEH, T., (1996), *A Second Course in Statistics.* (5th ed.). Prentice Hall.

NETER J., KUTNER M.H., NACHSHEİM C.J. and WASSERMAN W., (1996), *Applied Linear Regression Methods.* (4th ed.). The McGraw-Hill Irwin.

SCNEIDER J., BERNGES U., PHILIPP M. AND WOITOWITZ H., (2004), *GSTM1, GSTT1 and GSTP1 Polymorphism and Lung Cancer Risk in Relation to Tobacco Smoking,* Cancer Letters, 208, 65-74.

TATLIDİL H., (1996), *Uygulamalı Çok Değişkenli İstatistiksel Analiz,* Cem Ofset.

VINEIS P., AIROLDI L., VEGLIA, F., OLGIATI, L., PASTORELLI, R. and so on (2005), *Environmental Tobacco Smoke and Risk of Respiratory Cancer and Chronic Obstructive Pulmonary Disease in Former Smokers and never Smokers in the EPIC Prospective Study*, BMJ.

BINARY LOGISTIC REGRESSION AND AN APLICATION

ABSTRACT

Regression methods are one of any data analysis concerned with describing the relationship between a response variable and one or more explanatory variables. The logistic regression analysis defines the relation between dichotomous outcome variable and the set of independent variables that contains both continuous and discrete variables. There are some special problems when the response variable is dichotomous. In linear regression model, the error terms are assumed to have a normal distribution with a constant variance for all observations. But in logistic regression model, the error terms are not normal nor a constant variance when the response variable is dichotomous. The general method of estimation for logistic regression model is maximum likelihood. The method of maximum likelihood yields values for the unknown parameters which maximize the probability of obtaining the observed set of data. In order to apply this method, it is necessary to construct the likelihood function firstly. In order to determine whether the parameter is significant to the model or not, Deviance of the model containing the independent

variable must be compared with Deviance of the model without the independent variable. This change in D is called G statistic. Odds ratio (Ω) is used to construe the coefficients. There are many factors for patients with lung cancer. The logistic regression method is used for reducing the ratio of cancerous patients and selecting the variables in the model. In order to obtain a solution in this study, univariate analysis of each variable is applied to cancer data. The SPSS software package is used and results are evaluated.

Key Words: *Binary Logistic Regression, Likelihood Ratio Test, Maximum Likelihood, Odds Ratio (Ω)*