ANKARA ÜNİVERSİTESİ TIP FAKÜLTESİ MECMUASI Cilt 52, Sayı 4, 1999 191-199

LOJİSTİK REGRESYON ANALİZİ VE TIP ALANINDA KULLANIMINA İLİŞKİN BİR UYGULAMA

Yıldır Atakurt*

ÖZET

Bu çalışmada, çok değişkenli istatistiksel analiz yöntemlerinden Lojistik Regresyon Analizi'nin esasları ve ana basamakları açıklanmış ve kardiyolojik verilere uygulanarak yorumlanması yapılmıştır.

Çalışmada 500 bireyden elde edilen 34 değişken kullanılmış analiz sonucunda bağımlı değişken olarak alınan KOR (Koroner arter hastalığı) ile diğer bağımsız değişkenlerin ilişki dereceleri ve önemlilikleri test edilerek modele katkıları belirlenmiştir. Bulunan istatistikler doğrultusunda yorumlar yapılmıştır.

Anahtar Kelimeler: *Lojistik regresyon analizi, Lojit transformasyon, Odds oranı, Olabilirlik oran istatistiği.*

SUMMARY

Logistic Regression Analysis and Related Application to Usage in Medicine

In this study, the principles and basic steps of Logistic Regression Analysis, which is one of the means for multivariate analysis, are explained. Coronary Artery Disease data have been applied and the results of this application were evaluated.

Thirty-four independent variables were obtained from 500 patients as a candidate for multivariate model. The use of multivariate logistic regression analysis was permitted recognition of independent variables associated with dependent variable CAD (Coronary Artery Disease) and their significance tests. Interpretations were evaluated by using statistical results.

Key Words: *Logistic Regression Analysis, Logit transformation, Odds ratio, Likelihood ratio test.*

Tıp alanındaki araştırıcılar üzerinde çalıştıkları konuda çok etken olması durumunda etkenlerin tek tek bağımlı değişken üzerine etkisi yanında, bunların birlikte etkisini de öğrenmek ya da incelemek istemektedirler. Birlikte etkinin incelenmesinde kullanılan değişik istatistik yöntemler bulunmaktadır. Örneğin, bağımlı değişkenin sürekli, bağımsız değişkenlerin kesikli olması durumunda *varyans analizi*, hepsinin kesikli olması durumunda *log-lineer modeller*, hepsinin sürekli olması durumunda *regresyon analizi* gibi. Tıp alanındaki araştırmalarda çok zaman bağımlı ve bağımsız değişkenlerin tür ve yapıları yukarıda belirtilenlere benzemez, sürekli ve kesikli karışımı bağımsız değişkenlerle karşılaşılır. Üzerinde en çok durulan ve araştırıcı için önemli olan diğer bir konuda etken veya etkenlerle hastalık arasındaki ilişkinin risk yönünden incelenmesidir. Bu tip incelemelerde ağırlıklı olarak *Lojistik Regresyon Analizi* kullanılmaktadır. Lojistik regresyon analizi, temelde regresyon analizi olmakla birlikte bir ayırıcı analiz tekniği olma özelliğini de taşımaktadır. Ancak lojistik regresyon analizi, bağımsız değişken yapısı ve kombinasyonu yönünden diskriminant analizinden farklılık göstermektedir. Regresyon analizinden ise üç önemli farklılığı vardır.

1- Regresyon analizinde bağımlı değişken sayısal iken lojistik regresyon analizinde kesikli bir değer olmalıdır.
2- Regresyon analizinde bağımlı değişkenin değeri, lojistik regresyonda ise bağımlı değişkenin alabileceği değerlerden birinin gerçekleşme olasılığı kestirilir.
3- Regresyon analizinde bağımsız değişkenlerin çoklu normal dağılım göstermesi koşulu aranırken, lojistik regresyonun uygulanabilmesi için bağımsız de-

* Ankara Üniversitesi Tıp Fakültesi Biyoistatistik Anabilim Dalı Öğretim Üyesi

Geliş tarihi: 14 Aralık 1998 Kabul tarihi: 23 Şubat 1999

ğişkenlerin dağılımına ilişkin hiçbir koşul gerekmez.(1)

Çalışmamızda, lojistik regresyon analizinin esasları ve ana basamakları kısaca açıklandıktan sora, kardiyolojik veriler üzerinde uygulanması ve sonuçlarının değerlendirilmesi amaçlanmıştır.

YÖNTEM

Lojistik regresyon modelinde, bağımlı (sonuç) değişken ikili (binary) 0, 1 gibi kesikli bir değişken olup; risk belirten durum 1, diğer durum 0 ile gösterilir. Regresyon problemlerinde anahtar değer, verilen bir bağımsız değişken değerine bağlı olarak, bağımlı (sonuç) değişkenin ortalama değerini bulmaktır. Bu değer koşullu ortalama olarak adlandırılır ve E(Y\x) ile gösterilir. Burada, Y bağımlı değişkeni, x ise bağımsız değişkeni göstermektedir. Lineer regresyon analizinde, koşullu ortalamanın x'in lineer bir denklemi olduğu varsayılır;

$$E(Y\backslash x) = \beta_0 + \beta_1 x \qquad 1$$

bu eşitlik , x'in aralığının $-\infty$ ve $+\infty$ arasında değişmesinden dolayı E(Y\x)'in mümkün olan her değeri alabileceğini göstermektedir. Lojistik regresyon analizinde ise koşullu ortalama; 0'dan büyük, 1'den küçük ya da 1'e eşit olmak zorundadır.

$$0 \leq E(Y\backslash x) \leq 1 \qquad 2$$

Lojistik regresyon analizinde, $E(Y\backslash x) = \beta_0 + \beta_1 x$ eşitliğinin sol tarafı 0-1 arasında sınırlı olasılık değerleri aldığından ve bu değerler sonsuz değerler alabilen açıklayıcı değişkenlerle ilişkilendirildiğinden, söz konusu eşitlik her zaman sağlanamamaktadır. Böylesi bir durumla karşılaşılmaması için en iyi çözüm, sonuç değeri olarak ifade edilen olasılık değerinin çeşitli dönüşümlerle $-\infty$ ile $+\infty$ arasında tanımlı hale getirilmesidir (1,2).

İki düzey içeren bir sonuç değişkenin analizinde kullanılmak üzere önerilen birçok dağılım fonksiyonu vardır. En yaygın kullanılan iki tanesi lojit ve probit dönüşümleridir. Bunlardan lojistik dağılımı seçmek için de iki tane önemli neden vardır. İlk neden, lojistik regresyon analizinde varsayım kısıtlaması olmamasından dolayı kullanım rahatlığının yanı sıra, analiz sonucu elde edilen modelin matematiksel olarak çok esnek olması, ikinci neden ise biyolojik olarak kolay yorumlanabilir olmasıdır.

Gösterimi kolaylaştırmak için, lojistik dağılım kullanıldığında, x bilindiğinde Y'nin koşullu ortalamasını göstermek için $\pi(x) = E(Y\backslash x)$ değerini kullanabiliriz. Lojistik regresyon modelinin spesifik formu aşağıdaki gibidir;

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \qquad 3$$

Lojistik regresyon çalışmamıza merkez olacak $\pi(x)$'in bir transformasyonu yukarıda bahsedildiği gibi lojit transformasyondur. Bu transformasyonu $\pi(x)$ cinsinden aşağıdaki gibi tanımlarız;

$$g(x) = \left[\frac{\pi(x)}{1 - \pi(x)} \right] \qquad 4$$

Bu transformasyonun önemi, g(x)'in lineer regresyon modelinin istenen tüm özelliklerini taşımasıdır. Lojit g(x), parametreleri bakımından lineer, sürekli ve x'in aldığı değerlere bağlı olarak $-\infty$ ve $+\infty$ arasında değişebilmektedir (1).

Modelin Oluşturulması

Lojistik regresyon modelinde katsayıların kestirimi (tahmini) için lineer regresyonda olduğu gibi maksimum olabilirlik kestirimi yöntemi kullanılır. (x_i , y_i) gibi n tane bağımsız gözlem eşinin olduğu varsayıldığında, y_i iki düzeyli sonuç değişkenini, x_i'de i denek için bağımsız değişkenin değerini gösteriyorsa ve sonuç değişkeni için 0 ve 1 kodlarının belirli bir karakteristiğin yokluğunu ya da varlığını belirlediği kabul edildiğinde lojistik regresyon modelini uydurabilmek için bilinmeyen β_o ve β_1 parametrelerini kestirmemiz gerekir. Eğer Y, 0 ve 1 olarak kodlandıysa, $\pi(x)$ ifadesi x verildiğinde Y'nin 1'e eşit olma koşullu olasılığını vermektedir $\pi(x)$= P(Y= 1\x). [1-$\pi(x)$] değeri verilen herhangi bir x için Y'nin 0'a eşit olma koşullu olasılığını göstermektedir 1-$\pi(x)$= P(Y= 0\x). (x_i, y_i) çiftinin y_i= 1 olduğu zaman olabilirlik fonksiyonuna katkısı $\pi(x_i)$ iken, y_i= 0 olduğu zaman olabilirlik fonksiyonuna katkısı 1-p(x_i) kadardır. (x_i, y_i) çiftinin olabilirlik fonksiyonuna katkısını ifade etmenin güvenilir bir yolu da aşağıdaki gibidir;

$$\zeta(x_i) = \pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i} \qquad 5$$

Gözlemlerin birbirlerinden bağımsız olduklarını varsaydığımız için, olabilirlik fonksiyonu 5 numaralı denklemdeki terimlerin çarpılmasıyla elde edilir.

$$l(\beta) = \prod_{i=1}^{n} \zeta(x_i) \qquad 6$$

Matematiksel olarak 6 numaralı eşitliğin logaritmasıyla çalışmak daha kolay olacağından log olabilirlik fonksiyonu şöyle tanımlanmıştır;

$$L(\beta) = \ln\left[l(\beta)\right] = \sum_{i=1}^{n} \left\{ y_i \ln\left[\pi(x_i)\right] + (1-y_i) \ln\left[1-\pi(x_i)\right] \right\} \qquad 7$$

$L(\beta)$ maksimum yapan β değerini bulabilmek için $L(\beta)$'yı β_o ve β_1'e göre türevini alıp 0'a eşitleriz.

Sonuçta elde edilen eşitlikler aşağıdaki gibidir;

$$\sum_{i=1}^{n} \left[y_i - \pi(x_i) \right] = 0 \qquad 8$$

ve

$$\sum_{i=1}^{n} x_i \left[y_i - \pi(x_i) \right] = 0 \qquad 9$$

Bu eşitliklere olabilirlik eşitlikleri denir. Lineer regresyonda olabilirlik eşitlikleri kolay çözülebilen lineer denklemlerdir, fakat lojistik regresyonda bu ifadeler β_o ve β_1'e göre nonlineer denklemlerdir, bu denklemlerin çözümü için özel iterasyonla yapılan metotlar gerekir. 8 ve 9 nolu denklemlerden elde edilen β'nın değeri, maksimum olabilirlik kestirimi (tahmini) olarak adlandırılır ve $\hat{\beta}$ olarak gösterilir. Örnek olarak, $\pi(x_i)$'nin maksimum olabilirlik kestirimini $\hat{\pi}(x_i)$ ile gösterebiliriz. Bu değer x'in x_i gibi bir değere eşit olduğu bilindiği zaman, Y'nin 1'e eşit olma koşullu olasılığının kestirimini verir.

8 nolu denklemin sonucunda;

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\pi}(x_i) \qquad 10$$

y'nin gözlenen değerlerinin toplamının, kestirilen (tahmin edilen) değerlerin toplamına eşit olduğu görülmektedir.

Değişkenlerin Önemliliği

Regresyon analizi tekniğinin temel kavramlarından biri modele katılan değişkenlerin önemli olmasıdır. Modele katkısı olamayan değişkenler kullanarak kestirimde bulunmak hatalıdır. Aynı şekilde lojistik regresyon analizinde de modele katılacak olan değişkenlerin önemliliğin test edilmesi gerekir (3).

Katsayıları kestirdikten sonra, kestirilen modeldeki değişkenlerin önemlilikleri araştırılır. Bu test genelde, modelde bulunan bağımsız değişkenlerin "önemli" bir şekilde sonuç değişkeniyle ilişki içinde olup olmadığının testi şeklinde olmaktadır. Testin yapılış metotları bir modelden diğerine spesifik özelliklerine bağlı olarak farklılık göstermektedir.

Lojistik regresyonda katsayıların önem testi için ana prensip sorgulama altındaki değişkeni kapsayan ve kapsamayan modellerden elde edilen kestirim değerlerinin, sonuç değişkeninin gözlenen değerleriyle karşılaştırılmasıdır. Gözlenen ve kestirilen değerlerin karşılaştırma işlemi log-olabilirlik fonksiyonu ile yapılır. Olabilirlik fonksiyonlarını kullanarak gözlenen ve kestirilen değerleri karşılaştırmak aşağıdaki ifade ile olmaktadır;

$$D = -2\ln\left[\frac{\textit{Şu andaki modelin olabilirliği}}{\textit{Doymuş modelin olabilirliği}} \right] \qquad 11$$

Bu teste olabilirlik oranı testi adı verilir. (11) nolu denklemi kullanarak aşağıdaki denklemi elde ederiz.

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

D istatistiği sapma (deviance) olarak adlandırılır. Uyum iyiliğine karar verirken D istatistiği önemli bir rol oynamaktadır. Bağımsız bir değişkenin önemine karar vermek amacıyla, denklemde bağımsız değişkenin olduğu ve olmadığı durumlardaki D değerleri karşılaştırılır.

Bağımsız değişkeni kapsamasından dolayı ortaya çıkan D'deki değişim aşağıdaki gibidir;

G= D(Değişkensiz model için) –

D(Değişkenli model için) 13

Bu istatistik lineer regresyonda kullanılan F testindeki pay kısmı ile aynı rolü üstlenir.

$$G = -2\ln\left[\frac{\textit{Değişkensiz modelin olabilirliği}}{\textit{Değişkenli modelin olabilirliği}} \right] \qquad 14$$

Tek bağımsız değişkenli özel durumlarda, değişkenin modelde olmadığı zamanki β_o'ın maksimum olabilirlik kestirimi $\ln(n_1/n_0)$'dır $n_1 = \Sigma y_i$ ve $n_0 = \Sigma(1-y_i)$.

Kestirim değeri sabittir (n_1/n). G istatistiği de aşağıdaki gibidir;

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

ya da

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

Tüm değişkenleri içeren model ile kestirilen modele ilişkin olabilirlik oran değerlerinin farkına dayanan ölçütlerin ki-kare dağılımı göstereceği düşüncesinden hareketle kurulan modelin geçerliliği sınanmaktadır. Bu yolla modele girecek açıklayıcı değişkenlere karar verilmektedir. $\beta_1 = 0$ hipotezi altında, G istatistiği 1 serbestlik derecesinde ki-kare dağılım gösterir. Katsayıları kestirdikten sonra, kestirilen modeldeki değişkenlerin önemlilikleri araştırılır (1).

Çok Değişkenli Lojistik Regresyon

Birden çok bağımsız değişkenin yer aldığı lojistik modellere çok değişkenli lojistik regresyon adı verilir. Yapısal olarak bu modelin diğer çok değişkenli regresyon modellerinden farkı olmayıp regresyon katsayılarının yorumlanması farklıdır. Yorumlama bağımsız değişken türüne göre değişir. Çok değişkenli lojistik regresyonda sürekli olmayan değişkenler; nominal (sınıflandırılabilir) ve ordinal (sıralanabilir) değişkenler olabilir (2).

Lineer regresyonda olduğu gibi lojistik regresyonda da modellemenin gücü çok değişkenli modelleme yeteneğine bağlıdır. Değişkenlerden bazıları değişik ölçüm biçimlerinde olabilir. Çoklu lojistik regresyon modelinde genel eğilimimiz katsayıların tahmini ve onların önem testi şeklinde olacaktır. Kesikli ve nominal ölçekli bağımsız değişkenleri denkleme sokabilmek için dizayn değişkenleri kullanılacaktır (2,4).

$x' = (x_1, x_2, \ldots, x_p)$ vektörü ile gösterilen, p tane bağımsız değişken toplandığını varsayalım. Sonuç değişkeninin mevcut olduğu (Y=1) zamanki koşullu olasılık $P(Y=1\backslash x)=\pi(x)$'e eşittir. Çoklu lojistik regresyon modelinin lojiti aşağıdaki denklem ile verilmiştir;

$$g(x)= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \qquad 17$$

bu durumda

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}} \qquad 18$$

Eğer bazı bağımsız değişkenler kesikli, nominal ölçekli (ırk, cinsiyet, tedavi grupları v.b.) ise o zaman bu değişkenleri aralık değişkenleriymiş gibi denkleme sokmak yanlış olacaktır. Çünkü bu değişkenlere verilen kodların herhangi bir sayısal değerleri yoktur. Bu durumlarda dizayn değişkenleri ya da "dummy" değişkenleri (kukla değişkenleri) kullanılmalıdır.

Genel olarak, eğer nominal bir değişken k kategoriye sahipse o zaman $k-1$ dizayn değişkenine ihtiyaç vardır. $k-1$ dizayn değişkeni D_{ju} olarak ve katsayıları da β_{ju} olarak belirtilmiştir. Sonuç olarak, j. değişkeni kesikli olan p değişkenli model için lojit aşağıdaki gibidir;

$$g(x)= \beta_0 + \beta_1 x_1 + \ldots + \sum_{u=1}^{k_j-1}\beta_{ju}D_{ju} + \beta_p x_p \qquad 19$$

Birbirinden bağımsız n tane (x_i, y_i) değişkeni olduğunu varsayalım. Tek değişkenli modelde olduğu gibi modeli uydurmak için kestirim vektörünü $\beta'= (\beta_0, \beta_1, \ldots, \beta_p)$ elde etmemiz gerekir. Çok değişkenli durumda kestirim metodunun, tek değişkenlide olduğu gibi maksimum olabilirlik olduğunu söyleyebiliriz. Log olabilirlik fonksiyonunu $p+1$ katsayıya göre türevini alarak $p+1$ olabilirlik denklemi elde ederiz (5).

$$\sum_{i=1}^{n}\left[y_i - \pi(x_i)\right] = 0 \qquad 20$$

ve

$$\sum_{i=1}^{n} x_{ij}\left[y_i - \pi(x_i)\right] = 0,\ j=1, 2,\ldots,p \qquad 21$$

$\hat{\beta}$ bu denklemlerin çözümünü göstersin. Çoklu lojistik regresyon modeli için $\hat{\beta}$ ve x_i'yi kullanarak uydurulan değerler ile $\hat{\pi}(x_i)$ bulunur. Maksimum olabilirlik kestiriminin teorisi, log olabilirlik fonksiyonunun ikinci dereceden türevlerinden oluşan matristen kestirim değerlerinin elde edileceğini vurgular. Bu türevlerin genel şekli aşağıdaki gibidir;

$$\frac{\partial^2 L(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 \pi_i (1-\pi_i) \qquad 22$$

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_u} = -\sum_{i=1}^{n} x_{ij} x_{iu} \pi_i (1-\pi_i) \qquad 23$$

$j, u=0, 1, 2,\ldots,p.$

"Information" matrisi adı verilen $[(p+1) \times (p+1)]$ matrisi yukarıdaki denklemlerde verilen terimlerin negatiflerini kapsar. Kestirilen katsayıların varyans kovaryansları bu matrisin tersinden elde edilir $\Sigma\beta = I^{-1}(\beta)$. Çok özel durumların dışında bu matrisin açık şeklini yazmak mümkün değildir. $\sigma^2(\beta_j)$ ile bu matrisin j. diyagonal elementini gösterebiliriz, ki o da β_j'nin varyansıdır. Matrisin diyagonal olmayan elemanlarından $\sigma(\beta_j, \beta_u)$'de ve $\hat{\beta}_j$ ve $\hat{\beta}_u$'nun kovaryanslarını vermektedir. Varyans ve kovaryansların kestirimleri $\hat{\Sigma}(\hat{\beta})$ ile gösterilmiştir.

Matristeki elemanları $\hat{\sigma}^2(\hat{\beta}_j)$ ve $\hat{\sigma}(\hat{\beta}_j, \hat{\beta}_u)$ ile göstereceğiz. Kestirilen katsayıların standart hataları aşağıdaki gibidir;

$$\hat{SE}(\hat{\beta}_j) = \left[\sigma^2(\hat{\beta}_j)\right]^{1/2}, \quad j= 0, 1, 2, \ldots, p. \qquad 24$$

Yukarıdaki bu formülü, katsayıları test ederken ve kestirimlere ilişkin güven sınırlarını bulurken kullanacağız.

"Information" matrisinin aşağıdaki formu model uydururken ve uyumun iyiliği tartışılırken kullanılabilir.

$$I(\hat{\beta}) = X'VX \qquad 25$$

X matrisi $[n \times (p+1)]$ boyutunda bir matrisdir ve her bir denek için verileri kapsar.

$$X = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{bmatrix}$$

V matrisi $(n \times n)$ boyutunda genel elemanı $\pi_i(1-\hat{\pi}_i)$ olan $\hat{\pi}_i$ diyagonal bir matrisdir.

$$V = \begin{bmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & 0 & \ldots & 0 \\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \ldots & \ldots & \hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix}$$

Modelin Önemlilik Testi

Bu işlemdeki ilk adım modeldeki değişkenlerin önem kontrolünü yapmaktır. Modeldeki bağımsız değişkenler için p katsayının tümel olabilirlik oran testi tek değişkenli durumdakiyle aynıdır. Test 12 ve 14 nolu denklemlerde verildiği gibi *G* istatistiği temeline bağlıdır. "Modeldeki *p* tane "eğim" katsayısının sıfıra eşit olması" hipotezi altında, *G* istatistiği *p* serbestlik derecesinde ki kare dağılımı gösterir.

Değişkenleri tek tek Wald test istatistiği ile test edebiliriz $W_j = \hat{\beta}_j / \hat{SE}(\hat{\beta}_j)$. "Bir katsayının $(\hat{\beta}_j)$ sıfıra eşit olması" hipotezi altında Wald istatistiği standart normal dağılım gösterir. Bu istatistiğin önemi, modeldeki herhangi bir değişkenin önemli mi yoksa önemsiz mi olduğunu belirlemektir (1).

Göz önünde bulundurmamız gereken asıl nokta, en iyi uyum modelini en az parametre ile belirlemektir. Bundan sonraki ilk mantıklı adımımız, önemli olduğunu düşündüğümüz değişkenleri modele alarak yeni bir analiz yapmak ve bunu full modelle karşılaştırmaktır.

Kategorisel olarak ölçeklendirilmiş bağımsız değişkenler modelden çıkarıldığı (ya da girdiği) zaman, onun bütün dizayn değişkenleride modelden çıkarılmalıdır (ya da girmelidir). Eğer kategorik bir değişkenin *k* seviyesi varsa, serbestlik derecesine bu değişkenin katkısı *k-1* kadar olacaktır. Çoklu serbestlik derecesinden dolayı Wald istatistiğini kullanırken dikkatli olmamız gerekmektedir. Örnek olarak, eğer her iki katsayı için W istatistiği 2'yi geçerse, o zaman dizayn değişkeninin önemli olduğuna karar verebiliriz. Alternatif olarak, eğer katsayılardan birinin W istatistiği 3, diğerinin değeri 0.1 ise değişkenin modele katkısı hakkında kesin bir şey söyleyemeyiz (1,2).

Lojistik Regresyon Modelinde Katsayıların Yorumlanması

Karar vermemiz gereken ilk adım, "bağımlı değişkenin hangi fonksiyonu bağımsız değişkenler ile lineer bir fonksiyon oluşturmaktadır?" sorusudur. Bu fonksiyona link fonksiyonları adı verilir.(5)

Lineer regresyon modelinde link fonksiyonu *I* (identity) matrisidir, çünkü bağımlı değişken parametreleri ile lineerdir. Lojistik regresyon modelinde ise link fonksiyonu lojit transformasyondur.

$$g(x) = \ln\{\pi(x)/[1-\pi(x)]\} = \beta_0 + \beta_1 x \qquad 26$$

Lojistik regresyon katsayılarının yorumuna bağımsız değişkenin ikili olduğu zamanki durum ile başlayacağız. x'in 0 ve 1 ile kodlandığını varsayalım.

x=1 olan bireyler içinde, sonuç değişkeni görülme (y=1) odds değeri $\pi(1)/[1-\pi(1)]$ olarak tanımlanmıştır. Benzer şekilde $x = 0$ olan bireyler içinde, sonuç değişkeni görülme (y=1) odds değeri $\pi(0)/[1-\pi(0)]$ olarak verilmiştir. Odds değerlerinin logaritması lojit olarak adlandırılır.

$$g(1) = \ln\{\pi(1)/[1-\pi(1)]\}$$

$$g(0) = \ln\{\pi(0)/[1-\pi(0)]\}$$

Odds oranı, Ψ, bu odds değerlerinin oranı olarak tanımlanır.

$$\Psi = \frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]} \qquad 27$$

Odds oranının logaritması, log-odds, lojit farka eşittir.

$$\ln(\Psi) = \ln\left[\frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]}\right] \qquad 28$$

$$= g(1) - g(0)$$

Tablo 1'deki değerleri yukarıdaki denklemde yerine koyarsak odds oranı aşağıdaki gibi bulunur;

Tablo 1. İkili Bağımsız Değişkende Lojistik Regresyon Katsayıları

	Bağımsız değişken (X)	
	x = 1	x = 0
Sonuç değişkeni y = 1	$\pi(1)=\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$\pi(0)=\frac{e^{\beta_0}}{1+e^{\beta_0}}$
Sonuç değişkeni y = 0	$1-\pi(1)=\frac{1}{1+e^{\beta_0+\beta_1}}$	$1-\pi(0)=\frac{e^{\beta_0}}{1+e^{\beta_0}}$
Toplam	1.0	1.0

$$\Psi = \frac{\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\ \frac{1}{1+e^{\beta_0}}}{\frac{e^{\beta_0}}{1+e^{\beta_0}}\ \frac{1}{1+e^{\beta_0+\beta_1}}} \qquad 29$$

$$= \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

Lojistik regresyonda bağımsız değişken ikili ise odds oranı $\Psi = e^{\beta_1}$ ve lojit fark da $\ln(\Psi) = \beta_1$'e eşit olacaktır. Odds oranı çok yaygın kullanılan bir ilişki ölçüsüdür.

Sürekli değişkenlerde odds oranını yorumlarken, lojitin sürekli değişkenle lineer bir ilişkide olması gerekmektedir. Eğer lojitin, sürekli değişkenle lineer bir ilişki içinde olmadığı düşünülüyorsa, sürekli değişkenleri gruplandırmak ya da dizayn değişkenlerini kullanmak gerekecektir. Alternatif olarak, bu tür değişkenlere transformasyon yapılmalıdır.

Uyum İyiliği Testi

Doğrusal regresyon çalışmalarında modelin önemliliği için yapılan varyans analizi gibi lojistik regresyonda da modelin uyum iyiliği testi gereklidir. Uyum iyiliğinde kullanılan istatistik ve analizler çeşitlidir. Bunlardan bazıları regresyondan ayrılışın test edilmesi gibi klasik anlayışı taşırlar. Diğerleri ise lojistik mantıktan hareket ederek ki-kare uyum iyiliği istatistiğini kullanırlar. Regresyondan ayrılışı test etme amacı taşıyan uyum iyiliği test istatistiklerinde bağımsız değişken kavramı kullanılmaktadır. Genel olarak çok değişkenli regresyon analizlerinde birbiri ile aynı olan bağımsız değişken kombinasyonlarına bağımsız değişken deseni adı verilir. Bağımsız değişken deseni sayısı DS≤toplam gözlem sayısıdır. Bağımsız değişken desenleri regresyonda aynı davranışı göstermediklerinden özellikle regresyondan ayrılışların hesaplanmasında önemli rol oynarlar. Regresyondan ayrılışlar;

(Artık)Ai = y - P (y= 1 , x) (i = 1,2,...DS)

i' inci değişken deseni için;

Z_{Ai} = Standartlaştırılmış artık

olmak üzere Pearson İstatistiği;

$$Pearson\ \chi^2 = \sum_{i=1}^{DS} Z_{A_i}^2 \qquad 30$$

eşitliği ile bulunur. Pearson istatistiği, k bağımsız değişken sayısını göstermek üzere (DS-k-1) serbestlik derecesi ile khi-kare dağılımı gösterir.

Regresyondan ayrılışın incelenmesinde kullanılan diğer bir istatistik de Deviance İstatistiğidir. (y_i, x_i) çiftinin gözlem sayısı 1 olmak üzere Deviance artıkları;

$$y_i = 1 \quad \text{için} \quad d_i = \sqrt{2|\ln(P(y=1,x))|} \qquad 31$$

$$y_i = 0 \quad \text{için} \quad d_i = -\sqrt{2|\ln(1-P(y=1,x))|}$$

olmak üzere, bu özel durum dışında i' inci değişken desenindeki gözlem sayısı n_i olmak üzere

$$d_i = \sqrt{\left\{2\left[\left(\frac{y_i}{n_i P(y_i,x_i)}\right)+(n_i-y_i)\ln\left(\frac{(n_i-y_i)}{n_i[1-P(y_i,x_i)]}\right)\right]\right\}} \qquad 32$$

olmak üzere

$$D = \sum_{i=1}^{DS} d_i^2 \qquad 33$$

dir. D istatistiği de (DS-k-1) serbestlik derecesi ile khi-kare dağılımı gösterir.

Pearson ve Deviance istatistiklerinin DS = n olması durumunda dağılımları bozulduğundan kullanılması önerilmez.

MATERYAL

Uygulamamızda kullanılan veriler A.Ü. Tıp Fakültesi Kardiyoloji Anabilim Dalı kayıtlarından alınmıştır.(6) Çalışmaya dahil edilen 500 hastanın 356'sı koroner arter hastalığına sahip, 144 'ü ise koroner arter hastalığına sahip değildi. Çalışma grubundaki bireylerden elde edilen; YAŞ, CİNSİYET, D1Q, D1N, D2Q, D2N, D3Q, D3N, AVRQ, AVRN, AVLQ, AVLN, AVFQ, AVFN, V1Q, V1N, V2Q, V2N, V3Q, V3N, V4Q, V4N, V5Q, V5N, V6Q, V6N, KOR, KAH, AHIPO, AAKI, ADIS, IHIPO, IAKI, IDIS 34 adet değişken analizde kullanıldı. Bu değişkenlerden KOR (koroner arter hastalığı) bağımlı değişken olarak alındı. Değişkenler içerisinde tek sürekli değişken olan yaş, lojit ile lineer bir ilişki içinde olmadığından dolayı kategorik olarak bireyler 5 yaş grubunda toplandı (Tablo 2). Cinsiyet değişkeni, erkek=1 kadın=0 olarak, diğer değişkenler var=1 yok=0 olarak kodlandı.

BULGULAR VE TARTIŞMA

Çalışmamızda SPSS istatistik paket programının Lojistik Regresyon Analizi modülünden yararlanılmıştır (7). İlk aşamada her değişken için tek tek değişkenli analizler yapılarak analiz sonucunda $p<0.25$ değerine sahip olan ; YAŞ, CİNSİYET, D1Q, D1N, D2N, D3Q, AVLQ, AVLN, AVFQ, AVFN, V1Q, V1N, V2N,V3Q, V3N, V4N, V4N, V6N 18 adet değişken çok değişkenli regresyon analizine aday olarak seçildi. İkinci aşamada bağımlı değişken seçilen KOR ile en fazla ilişkili değişkenden başlanarak bağımsız değişkenler modele birer birer ilave edildi ve önemlilik testleri G istatistiği ile yapıldı. Önemliliği $p < 0.05$ 'in altında olan bağımsız değişkenler modele dahil edildiler (Tablo 3).

Modele alınan değişkenlerin kestirilen katsayıları, standart hataları ve odds oranları Tablo 4'de verilmiştir.

Aşağıdaki formül yardımıyla, verilen bir hastanın koroner damar hastalığına sahip olma olasılığı hesaplanabilir:

$$P(KAR)=\frac{e^{[-.07Y(1)+1.06Y(2)+1.66Y(3)+1.45Y(4)+1.75C+.92D2N+1.52D3Q+1.06AVLN+2.03AVFQ+3.55V1Q+2.20V1N+2.76V5N-3.07]}}{1+e^{[-.07Y(1)+1.06Y(2)+1.66Y(3)+1.45Y(4)+1.75C+.92D2N+1.52D3Q+1.06AVLN+2.03AVFQ+3.55V1Q+2.20V1N+2.76V5N-3.07]}}$$

Y(1)=YAŞ(1), Y(2)=YAŞ(2), Y(3)=YAŞ(3), Y(4)=YAŞ(4), C=CİNSİYET

Tablo 4'ün odds oranlarına göre irdelenmesi yapıldığında 34 ve altındaki yaş grubuna göre; diğer yaş grupları içerisinde en yüksek odds oranına 55-64 yaş grubunun sahip olduğu, bu yaş grubunun koroner ar-

Tablo 2. Yaş Gruplarına Göre Değişkenler

Yaş grupları	Grup no	Frekans	Yaşgr1	Yaşgr2	Yaşgr3	Yaşgr4
≤ 34	1	17	0	0	0	0
35 - 44	2	88	1	0	0	0
45 - 54	3	164	0	1	0	0
55 -64	4	168	0	0	1	0
≥ 65	5	63	0	0	0	1

Tablo 3. Bağımsız Değişkenler

Değişken	-2 Log Likelihood	G	s.d.	p
Sabit terim	600.351	–	–	–
CİNSİYET	520.536	79.815	< 0.001	1
D3Q	472.813	47.723	< 0.001	1
V1Q	428.287	44.526	< 0.001	1
V1N	394.161	34.126	< 0.001	1
YAŞ	370.079	24.082	< 0.001	4
D2N	357.729	12.350	< 0.001	1
V5N	338.514	19.125	< 0.001	1
AVLN	328.11	10.403	< 0.01	1
AVFQ	323.536	4.575	< 0.05	1

ter hastalığına sahip olma riskinin 5.3 kat daha fazla olduğu görülmektedir.

Cinsiyet değişkeninde ise erkeklerin kadınlara göre 5.7 kat daha fazla riske sahip oldukları görülmektedir.

Elektrokardiyografi dalga bulgularına göre ise;

- V1Q dalgasında bozukluğa sahip olanların, olmayanlara göre 34.8
- V5N dalgasında bozukluğa sahip olanların, olmayanlara göre 15.9
- V1N dalgasında bozukluğa sahip olanların, olmayanlara göre 9.0
- AVFQ dalgasında bozukluğa sahip olanların, olmayanlara göre 7.6
- D3Q dalgasında bozukluğa sahip olanların, olmayanlara göre 4.6
- AVLN dalgasında bozukluğa sahip olanların, olmayanlara göre 2.9
- D2N dalgasında bozukluğa sahip olanların, olmayanlara göre 2.5

kat daha fazla koroner arter hastalığına yakalanma riskine sahip oldukları görülmektedir.

Modeldeki bağımsız değişkenlerin uyum iyiliği testi Pearson ve Deviance istatistikleri ile yapıldı. Hesaplamalar sonucu Pearson istatistiği $\chi^2 = 140.65$, Deviance istatistiği D = 83.86 bulundu. Her iki değer de (186-9-1) 176 serbestlik derecesinde χ^2 dağılımı tablo değeriyle (207.96) karşılaştırıldı $p > 0.05$ bulundu. Modelin uyumlu olduğu görüldü.

Lojistik regresyon analizinin ayırım gücü Tablo 5'de verilmiştir. Tablonun incelenmesinde modelin ayırıcılık tümel gücünün % 90 olarak belirlendiği görülmektedir. Koroner damar hastalığına sahip olma-

Tablo 4. Değişkenlerin Katsayıları, Standart Hataları, Odds Oranları

Değişken	B	Standart Hata	Odds oranı
YAŞ(1)	-.0731	.8678	.9295
YAŞ(2)	1.0644	.8406	2.8992
YAŞ(3)	1.6639	.8407	5.2796
YAŞ(4)	1.4522	.9105	4.2724
CİNSİYET	1.7455	.3077	5.7287
D2N	.9228	.2918	2.5162
D3Q	1.5223	.7435	4.5830
AVLN	1.0560	.3293	2.8749
AVFQ	2.0257	.9865	7.5816
V1Q	3.5497	.7906	34.8019
V1N	2.2001	.5011	9.0258
V5N	2.7647	.7917	15.8750
Sabit terim	-3.0700	.8447	

Tablo 5. Lojistik Regresyon Analizinin Ayırım Gücü

GÖZLENEN (KOR)	KESTİRİLEN (KOR)		SINIFLAMA YÜZDESİ
	KOR (-)	KOR (+)	
KOR (-)	116	28	80.56
KOR (+)	22	334	93.82
		TÜMEL	90.00

yanları doğru olarak sınıflandırılma, seçicilik oranı (specificity) % 80.56, koroner damar hastalığına sahip olanların doğru olarak sınıflandırılma, duyarlılık oranı (sensitivity) % 93.82 olarak saptandı.

KAYNAKLAR

1. Hosmer D.W., Lemeshow S.: Applied Logistic Regression, J.Wiley&Sons, New York, 1989.
2. Agresti A. : Categorical Data Analysis, J.Wiley&Sons, New York, 1990.
3. Schlesselman J.J.: Case Control Studies, Oxford University Press, New York, 1982
4. Aggarwal A.R., Singh P.: Estimation of Relative Risk Control Studies Thrrough Logistic Regression Analysis, Biom.J.Vol:35 No.4,479-485, 1993.
5. Christensen R.: Log-Lineer Models, Springer-Verlag, New York, 1990.
6. Alpman A. et all : Importance of Notching and Slurring of the Resting QRS complex in the Diagnosis of Coronary Artery Disease, Journal of Electrocardiology Vol:28 No.3 ,199-209,1995.
7. Norusis M.J.: SPSS For Windows 6.0 Advences Statistics, SPSS Inc. Chicago, 1993.