www.istatistikciler.org
İstatistikçiler Dergisi 1 (2008) 135-143

İstatistikçiler Dergisi

İki düzeyli bağımlı değişken modelinin yarı parametrik tahmini

Özge Akkuş
Muğla Üniversitesi, Fen-Edebiyat Fakültesi, İstatistik Bölümü, 48170, Kötekli, Muğla
ozgeucar@hacettepe.edu.tr

Serdar Demir
Pamukkale Üniversitesi, İktisadi ve İdari Bilimler Fakültesi, Ekonometri Bölümü, 20020, Kınıklı, Denizli
sdemir@pau.edu.tr

Durdu Karasoy
Hacettepe Üniversitesi, Fen Fakültesi, İstatistik Bölümü, 06800, Beytepe, Ankara
durdu@hacettepe.edu.tr

Özet

İki düzeyli bağımlı değişken durumunda model tahmini için üç temel yaklaşım vardır. Bunlar, parametrik, parametrik olmayan ve yarı parametrik yaklaşımlardır. Parametrik yaklaşımda çok fazla varsayıma ihtiyaç duyulduğundan dolayı sonuçların güvenilirliği de giderek azalmaktadır. Parametrik olmayan yaklaşımda hiçbir varsayım yapılmamakta fakat açıklayıcı değişken sayısının fazla olması durumunda model tahmini zorlaşmaktadır. Bu çalışmada ise parametrik model varsayımlarının sağlanmaması durumunda parametrik ve parametrik olmayan yaklaşımların en iyi yönlerini alan yarı parametrik yöntem tanıtılmıştır. Bu yöntemde model tahmini iki aşamada elde edilmektedir. İlk aşamada parametreler, Klein ve Spady'nin yarı parametrik en çok olabilirlik tahmin edicisi ile, ikinci aşamada ise ilgilenilen olayın gerçekleşme olasılıkları, parametrik olmayan Nadaraya-Watson çekirdek kestirim yöntemi ile tahmin edilmiştir. Bu modelin uygulanabilirliği sayısal bir örnek ile gösterilmiştir.

Anahtar Kelimeler: *Yarı parametrik model; Klein ve spady tahmin edicisi; Nadaraya-Watson çekirdek kestiricisi.*

Abstract

The semiparametric estimation of the binary dependent variable model

There are three main approaches for the model estimation when the dependent variable is binary. These are the parametric, the nonparametric and the semiparametric approaches. The reliability of the results gradually decreases due to the fact that more assumptions are needed in the parametric approach. Any assumption is made in the nonparametric approach but model estimation becomes hard in the event that there are too many explanatory variables. In this study, the semiparametric method that combines the best features of the parametric and the nonparametric approaches are introduced when the parametric model assumptions are violated. In this method, the model estimation is obtained in two steps. In the first step, the parameters are estimated by the Klein and Spady's semiparametric maximum likelihood estimator while the realization probabilities of the concerned event are estimated by the nonparametric Nadaraya-Watson kernel estimation method in the second step. The applicability of this method is shown by a numerical example.

Keywords: *Semiparametric model; Klein and Spady estimator; Nadaraya-Watson kernel estimator.*

1. Giriş

Uygulamalı istatistiğin çoğu alanında amaç, Eşitlik (1) ile verilen koşullu ortalama fonksiyonunu modellemektir. Bir rastgele değişken olan bağımlı değişken Y'nin iki düzeyli bir değişken olması durumunda, X açıklayıcı değişkenlere bağlı olarak koşullu ortalama fonksiyonu yazıldığında, bir olasılık [P (.)] ifadesine ulaşılır.

$$E(Y|X=x)=P[Y=1|X=x]=G(X^{T}\beta+\varepsilon) \qquad (1)$$

Eşitlik (1)'de, X, açıklayıcı değişkenler vektörünü; β, açıklayıcı değişkenler ile ilgili parametreler vektörünü; T, transpozu; ε, hata terimini ve G, hata teriminin dağılım fonksiyonunu göstermektedir. Model tahmini için üç temel yaklaşım vardır. Bunlar, parametrik, parametrik olmayan ve yarı parametrik yaklaşımlardır [5,6].

Parametrik yaklaşım tümüyle varsayımlara dayalıdır. Eşitlik (1)'de, G'nin bilinen bir fonksiyon olduğu ve X açıklayıcı değişkenler arasındaki fonksiyonel yapının da $X^T\beta$ biçiminde doğrusal olduğu ve β parametrelerinin sonlu sayıda olduğu varsayılmaktadır [4,10]. İki düzeyli lojistik regresyon ve probit regresyon modelleri bu kategoridedir ve G'nin sırasıyla lojistik ve standart normal dağıldığı varsayımı üzerine kuruludur [1,3].

Parametrik olmayan yaklaşımda, Eşitlik (1)'de, G hata dağılımı ile ilgili herhangi bir varsayım yapılmamakta, β parametreler vektöründen bahsedilmemekte ve açıklayıcı değişkenler arasındaki fonksiyonel yapı da bilinmemektedir. Hiçbir varsayım gerektirmediğinden dolayı bu yaklaşımın uygulamalarda daha fazla tercih edilmesi gerekirken, birçok problem ile karşılaşılmasından dolayı nadiren kullanılmaktadır. Bu problemler içinde en fazla dikkat çekeni, açıklayıcı değişken sayısının fazla olması durumunda tahmin ve yorumlamada güçlük çekilmesidir. Bu sorun, "Boyutluluk Sorunu" olarak adlandırılmaktadır [5,6].

Eşitlik (1)'de G dağılım fonksiyonu ile ilgili herhangi bir dağılım varsayımı yapılmadığında ve açıklayıcı değişkenler arasındaki fonksiyonel yapı doğrusal varsayıldığında, başka hiçbir varsayım gerektirmeyen yarı parametrik yöntemler kullanılmaktadır [5].

Bu çalışmada iki düzeyli bağımlı değişken modelinin tahmininde yarı parametrik yaklaşım tanıtılmış ve sayısal bir örnek olarak mide kanseri verisi kullanılarak belirli bir süreç sonunda hastalığın nüksedip nüksetmemesine etki eden faktörler ve etki etme olasılıkları araştırılmıştır.

2. Yarı parametrik yaklaşım

Parametrik yaklaşımda çok fazla varsayım yapıldığından dolayı sonuçların güvenilirliği de giderek azalmaktadır. Parametrik olmayan yaklaşımda ise hiçbir varsayım yapılmamakta fakat açıklayıcı değişken sayısının fazla olması durumunda model tahmini zorlaşmaktadır. Yarı parametrik yaklaşım, parametrik ve parametrik olmayan yaklaşım arasında bir orta yol bulmayı amaçlamaktadır [11,12].

Yarı parametrik yaklaşımda, açıklayıcı değişkenler arasındaki fonksiyonel yapının parametrik yaklaşımda olduğu gibi $X^T\beta$ biçiminde doğrusal olduğu ve bilinmemesi durumunda "g" ile gösterilen hata dağılımı G'nin, modelden tahmin edildiği varsayılmaktadır. Model aşağıdaki biçimde ifade edilmektedir:

$$E(Y|X = x) = P[Y = 1|X = x] = g(X^T\beta) \qquad (2)$$

$x^T\beta$ doğrusal indeks varsayımından dolayı parametrik modelin bir özelliğini ve bilinmeyen "g" hata dağılım fonksiyonundan dolayı parametrik olmayan modelin bir özelliğini alan yarı parametrik model uygulamaları son yıllarda giderek yaygınlaşmaktadır [5,6].

Parametrik olmayan model tahmininde yorumlanabilir sonuçlar elde etmek için en fazla iki açıklayıcı değişken ile çalışmak olası iken, yarı parametrik yöntemde k tane açıklayıcı değişkenin bağımlı değişken üzerindeki etkisini incelemek mümkündür. Ayrıca parametrik modeldeki kadar varsayım yapılmaması nedeniyle bu yaklaşımın uygulamalı çalışmalarda kullanılması önerilmektedir [7].

2.1. Tahmin yöntemleri

Yarı parametrik model tahmini iki aşamadan oluşmaktadır. İlk aşamada uygun bir yarı parametrik yöntem ile β parametreler vektörü tahmin edilmekte ($\hat{\beta}$) ve her bir gözlem için $X^T\hat{\beta}$ değerleri hesaplanmakta, ikinci aşamada ise bağımlı değişken Y'nin $X^T\hat{\beta}$ üzerine parametrik olmayan regresyonu uygulanarak gözlemlerin bağımlı değişkende "1" olarak kodlanan düzeye ait olma olasılıkları tahmin edilmektedir.

$\hat{\beta}$ parametre tahminleri, parametrik yaklaşımda En Çok Olabilirlik Tahmin Edicisi (EÇOTE) ile; yarı parametrik yaklaşımda ise açıklayıcı değişkenlerin karma (kesikli-sürekli) olması durumunda Klein ve Spady (KS) (1993) tarafından geliştirilen Yarı Parametrik En Çok Olabilirlik Tahmin Edicisi (YPEÇOTE) ile elde edilmektedir.

Bu çalışmada yarı parametrik yaklaşım üzerine yoğunlaşıldığından YPEÇOTE yöntemi aşağıda ayrıntılı olarak verilmiştir.

2.1.1. Klein ve Spady'nin yarı parametrik en çok olabilirlik tahmin edicisi

KS tahmin edicisi, bağımlı değişken Y'nin sadece 0-1 gibi iki değer aldığı durumda kullanılmaktadır. Y, iki düzeyli bir değişken olduğundan dolayı, bu model için logaritmik olabilirlik fonksiyonu,

$$\log L_N(b) = N^{-1}\sum_{n=1}^{N}\{y_n \log G(x_n\beta) + (1-y_n)\log[1-G(x_n\beta)]\} \quad (3)$$

biçimindedir. Parametrik yaklaşımda G bilinen bir fonksiyon olduğundan dolayı (lojistik ya da normal dağılım) olabilirlik fonksiyonunun açık ifadesini elde etmek kolaydır. Ancak, yarı parametrik yaklaşımda G ile ilgili herhangi bir dağılım varsayımı yapılmadığından dolayı yerine geçebilecek bir tahmin ediciye ihtiyaç duyulmaktadır. Böyle bir tahmin edici parametrik olmayan yöntemler ile elde edilebilmektedir. G_N, G 'nin parametrik olmayan tahmini olmak üzere, Klein ve Spady, G_N'in, y'nin xb_{ks} üzerine parametrik olmayan regresyon tahmini ile elde edilebileceğini göstermiştir.

$$P_N = \frac{\sum_{n=1}^{N} y_n}{N} \quad (4)$$

bağımlı değişkende "1" cevabını verenlerin oranı

$$g_N(\upsilon|y=1) = \frac{1}{(N P_N h_N)}\sum_{n=1}^{N} y_n K\left[\frac{(\upsilon - x_n b_{ks})}{h_N}\right] \quad (5)$$

bağımlı değişkende "1" cevabını veren kişiler için elde edilen, $\upsilon = x_n b_{ks}$'nin çekirdek yoğunluk fonksiyonu tahmini,

ve

$$g_N(\upsilon|y=0) = \frac{1}{[N(1-P_N)h_N]}\sum_{n=1}^{N}(1-y_n) K\left[\frac{(\upsilon - x_n b_{ks})}{h_N}\right] \quad (6)$$

bağımlı değişkende "0" ı tercih eden kişiler için elde edilen, $\upsilon = x_n b_{ks}$'nin çekirdek yoğunluk fonksiyonu tahmini olmak üzere, G_N, aşağıdaki biçimde tahmin edilebilmektedir.

$$G_N(\upsilon) = \frac{P_N\, g_N(\upsilon|y=1)}{P_N\, g_N(\upsilon|y=1) + (1-P_N)\, g_N(\upsilon|y=0)} \tag{7}$$

Eşitlik (7) ile verilen $G_N(\upsilon)$'nin, Eşitlik (3)'de yerine koyulup olabilirlik fonksiyonunun maksimize edilmesi ile bilinmeyen β parametreler vektörü tahmin edilir.

Yarı parametrik tahminde dikkat edilmesi gereken en önemli nokta, parametrelerin tanımlanabilirlik koşullarının sağlanabilmesi ve tek bir $\hat{\beta}$ vektörünün elde edilebilmesi için çalışmada mutlaka sürekli bir açıklayıcı değişkenin olması koşuludur.

β parametreler vektörü tahmin edildikten sonra ikinci aşamada Eşitlik (2)'deki bilinmeyen g fonksiyonunun tahmin aşamasına geçilir. Bu tahminler, gözlemlerin bağımlı değişkende "1" olarak kodlanan düzeye ait olması olasılıklarını vermektedir. İlk olarak tüm gözlemler için $X^T\hat{\beta}$ indeks değerleri hesaplanır. Böylece yarı parametrik tahminin ilk aşaması tamamlanmış olur. Daha sonra bağımlı değişken Y'nin $X^T\hat{\beta}$ indeksi üzerine tek değişkenli parametrik olmayan regresyonu ile olasılık tahminlerine ulaşılır [8].

Bu çalışmada, yarı parametrik yaklaşımın ikinci aşamasında, parametrik olmayan regresyon tahmin edicilerinden Nadaraya-Watson (NW) tahmin edicisi kullanılmıştır. Aşağıda bu yöntemin ayrıntıları yer almaktadır.

2.2. Tek değişkenli parametrik olmayan Nadaraya-Watson tahmin edicisi

Koşullu ortalama ve koşullu ortalama fonksiyonu kavramları regresyon modellerinin merkezini oluşturmaktadır. X ve Y, $f(x,y)$ bileşik olasılık yoğunluk fonksiyonuna sahip iki rastgele değişken olsun. Y'nin verilen $X = x$ 'e göre koşullu beklenen değeri,

$$E(Y|X=x) = \int y\, f(y|x)\, dy = \int y \frac{f(x,y)}{f_X(x)} dy = m(x) \tag{8}$$

biçiminde ifade edilmektedir. Burada $f(y/x)$, Y'nin verilen $X = x$ 'e göre koşullu olasılık yoğunluk fonksiyonu ve $f_X(x)$, X'in marjinal olasılık yoğunluk fonksiyonudur.

Veri kümesindeki her bir x_i için koşullu beklenen değer $m(x_i)$ elde edilir ve toplam n tane ($i = 1,\ldots,n$) değerden oluşan koşullu beklenen değerler kümesi oluşturulur. Böylece, Y ve X'in "ortalama olarak" nasıl ilişkili olduğu ortaya çıkmaktadır. Bu nedenle regresyonda aşağıda verilen $m(.)$'in tahmini ilgilenilen temel noktadır.

$$m(x) = E(Y|X=x) = \int y \frac{f(x,y)}{f_X(x)} dy = \frac{\int y\ f(x,y)\, dy}{f_X(x)} \tag{9}$$

$\{X_i, Y_i\}$, $(i = 1,\ldots,n)$ biçiminde verilen gözlemler için, Eşitlik (9)'da $f(x,y)$ ve $f_X(x)$ bilinmemektedir. Tek bir değişkenin olasılık yoğunluk fonksiyonu olmasından dolayı $f_X(x)$'in tahmini kolaydır. İki değişken X ve Y'nin bileşik olasılık yoğunluk fonksiyonu olan $f(x,y)$'nin tahmini için, "Çarpımsal Çekirdekler ile Çekirdek Yoğunluk Fonksiyonu" özelliği kullanılmaktadır. h ve g sırasıyla X ve Y değişkenlerinin yoğunluklarının tahmininde kullanılan bant genişlikleri olmak üzere, bu özellik kullanılarak elde edilen yoğunluk fonksiyonu tahmini aşağıdaki biçimde ifade edilmektedir.

$$\hat{f}_{h,g}(x,y) = \frac{1}{n}\sum_{i=1}^{n} K_h\left(\frac{x - X_i}{h}\right) K_g\left(\frac{y - Y_i}{g}\right) \tag{10}$$

Çekirdek fonksiyonlarının integralinin 1'i verdiği ve 0 etrafında simetrik olduğu bilgisinden yararlanılarak veri kümesindeki herhangi bir x için karşılık gelen Y değerinin parametrik olmayan NW tahmini, aşağıdaki formül ile yapılmaktadır.

$$\hat{m}_h(x) = \frac{n^{-1}\sum_{i=1}^{n} K_h(x - X_i)\, Y_i}{n^{-1}\sum_{i=1}^{n} K_h(x - X_i)} \tag{11}$$

$\hat{m}_h(x)$ değerleri daha önce de belirtildiği gibi olasılık tahminlerini vermektedir. Böylece yarı parametrik tahmin süreci tamamlanır ve yoruma geçilir [9,13].

3. Sayısal örnek

Bu çalışmada, Hacettepe Üniversitesi Biyoistatistik Bölümü'nden temin edilmiş ve Cox regresyon çözümlemesi yapılmış olan mide kanseri hastalarına ait veriler kullanılmıştır [2]. Burada amaç, bu verilerin tıbbi yorumları değil, yarı parametrik yaklaşımın veriye uygulanabilir olduğunu göstermektir. Bu amaçla mide kanseri olan 95 hasta bilgisi kullanılarak, hastalığın nüksetmesine etki eden önemli faktörleri belirlemek ve hastalığın nüksetme olasılıklarını ortaya çıkarabilmek amacıyla yarı parametrik tahmin yöntemi uygulanmıştır.

İlgilenilen temel değişken (bağımlı değişken), nüks değişkenidir. Bu değişken ise, hastanın belirli bir tedavi süreci sonunda hastalığının nüksetmesi ya da nüksetmemesi olarak iki kategorilidir. Bu bağımlı değişkeni etkileyebilecek açıklayıcı değişkenler ise düzeyleri ve tanımları ile birlikte Çizelge 1'de verilmektedir.

Kategorik açıklayıcı değişkenlerin varlığında ilk ya da son düzeyler referans olarak alınmakta, bu düzeylere ilişkin parametre tahmini yapılmamakta ve yorumlar bu referans düzeylere göre elde edilmektedir. Bu çalışmada, "0" olarak kodlanan ilk düzeyler referans; "1" olarak kodlanan düzeyler ise indikatör düzeyler olarak alınmıştır. Modelde, açıklayıcı değişkenlerin sadece referans dışındaki düzeyleri için parametre tahminleri yer almaktadır.

Analizin ilk aşamasını oluşturan β parametreler vektörünün tahmini için gerekli olan KS kodları Nlogit 4.0 yazılımı ile ikinci aşamadaki hastalığın nüksetme olasılıklarının tahminini veren parametrik olmayan NW tahmin edicisi için ilgili program kodları ise Delphi 6.0 programlama dili ile oluşturulmuştur.

Çizelge 1. Kullanılan değişkenler ve düzeyleri

Değişken	Değişken Düzeyleri ($\bar{x} \pm$ standart hata)	n (%)	Nüksetmemiş Olay Sayısı (%)	Nüksetmiş Olay Sayısı (%)
METASTAZ	0. Yok (referans) 1. Var (METASVAR)	67 (70.5) 28 (29.5)	53 (79.1) 8 (28.6)	14 (20.9) 20 (71.4)
ALKOL	0. Yok (referans) 1. Var (ALKOLVAR)	84 (88.4) 11 (11.6)	54 (64.3) 7 (63.6)	30 (35.7) 4 (36.4)
KİLOKAYBI	0. Yok (referans) 1. Var (KKAYBIVAR)	53 (55.8) 42 (44.2)	38 (71.7) 23 (54.8)	15 (28.3) 19 (45.2)
RADYOTERAPİ	0. Yok (referans) 1. Var (RTERAPIVAR)	30 (31.6) 65 (68.4)	18 (60.0) 43 (66.2)	12 (40.0) 22 (33.8)
CEA	4.9237 ± 1.2493			
Bağımlı değişken	0. Nüksetmemiş 1. Nüksetmiş	61 (64.2) 34 (35.8)		

Yarı parametrik tahmininin ilk aşamasını oluşturan $\hat{\beta}$ parametre tahminleri için KS sonuçları Çizelge 2'de verilmektedir. Bu aşama için optimal bant genişliği, $h = 0.46814$ olarak bulunmuştur.

Çizelge 2. KS parametre tahminleri

Değişken	Parametre tahmini ($\hat{\beta}$)	Standart hata (Sh)	$\hat{\beta}/Sh$	p	Odds oranı
METASVAR	**3.1730**	**1.0871**	**2.919**	**0.0035***	**23.8787**
ALKOLVAR	-1.8331	1.2382	-1.480	0.1388	0.1599
KKAYBIVAR	**3.4655**	**1.3151**	**2.635**	**0.0084***	**31.9937**
RTERAPIVAR	**-3.3731**	**1.3461**	**-2.506**	**0.0122***	**0.0343**
CEA	1.0	Sabit parametre	-	-	-
Sabit terim	0.0	Sabit parametre	-	-	-

* α=0.05 yanılma düzeyinde anlamlı

Çizelge 2 incelendiğinde, yarı parametrik modelin tanımlanabilirliği için gerekli olan sürekli bir değişkenin katsayısının bire normalleştirilmesi koşulunun gerçekleştirildiği görülmektedir. "CEA" sürekli değişkeninin katsayısı 1 alınmış ve diğer değişkenlerin katsayıları bu normalleştirmeye bağlı olarak tahmin edilmiştir. Yarı parametrik yöntemin parametrik alternatifinden bir diğer farklılığı ise herhangi bir dağılım (merkezlenme) varsayımı yapılmadığından dolayı sabit terimin tahmin edilmemesidir.

3.1. Parametre tahminleri

Eşitlik (2) ile verilen modeldeki $X^T\hat{\beta}$ doğrusal indeksi, Çizelge 2'deki tahmin değerleri kullanılarak oluşturulmaktadır. Bu durumda yarı parametrik modelin ilk aşamasını oluşturan doğrusal indeks,

$$X^T\hat{\beta} = 3.1730\,\text{METASVAR} - 1.8331\,\text{ALKOLVAR} + 3.4655\,\text{KKAYBIVAR} - 3.3731\,\text{RTERAPIVAR} + \text{CEA}$$

biçiminde tahmin edilmektedir.

METASVAR, KKAYBIVAR ve RTERAPIVAR değişkenlerinin 0.05 yanılma düzeyinde istatistiksel olarak önemli bulunduğu görülmektedir (sırasıyla $p = 0.0035$; $p = 0.0084$; $p = 0.0122$). Parametre tahminlerinin negatif olması, hastalığın nüksetme olasılığında azalışa neden olurken, pozitif olması nüksetme ihtimalini artırmaktadır. Buna göre, metastaz'ın var olması (3.1730) ve kilo kaybındaki bir

birimlik bir artışın (3.4655) hastalığın nüksetme ihtimalini artırdığı; Radyoterapi uygulamasının ise (– 3.3731) hastalığın nüksetme ihtimalini azalttığı sonucuna ulaşılmaktadır.

3.2. Odds oranları

Parametrik lojistik regresyon analizinde olduğu gibi yarı parametrik alternatifinde de odds oranları yorumlanabilmektedir. Odds oranı, değişkenlerin referans düzeylerine oranla diğer düzeylerinin varlığında, bağımlı değişkende "1" olarak kodlanan (nüksetme) düzeyin ortaya çıkması ihtimalini vermektedir. Çizelge 2'nin son kolonunda odds oranları yer almaktadır. İstatistiksel olarak önemli bulunan değişkenler için bu değerler aşağıdaki biçimde yorumlanmaktadır.

Metastaz değişkeni için odds oranı 23.8787 bulunmuştur. Buna göre, metastaz olan mide kanseri hastalarının olmayanlara oranla hastalıklarının nüksetme ihtimalinin yaklaşık 23.88 kat daha fazla olduğu söylenebilir.

Kilo kaybı değişkeni için odds oranı 31.9937 bulunmuştur. Buna göre, kilo kaybı olanların olmayanlara oranla hastalıklarının nüksetme ihtimali yaklaşık 31.99 kat daha fazla olduğu söylenebilir.

Radyoterapi değişkeni için odds oranı 0.0343 bulunmuştur. Yorumu daha anlaşılır yapmak için bu değerin tersi alınırsa 29.16828 değeri elde edilir. Yorum yapılırken radyoterapi değişkeninin referans ve indikatör kategorilerinin de yer değiştirmesi gerekir. Bu durumda, radyoterapi almayan hastaların alanlara oranla hastalıklarının nüksetme ihtimalinin yaklaşık 29.17 kat daha fazla olduğu sonucuna ulaşılır.

Yarı parametrik modelde parametre tahminleri ve odds oranlarının yorumu parametrik alternatifinden farklı değildir. Buradaki amaç, doğru istatistiksel modeli kurabilmek, doğru tahminleri yorumlayabilmek ve bu modelin uygulanabilirliğini göstermektir.

3.3. Olasılık tahminleri

Bu bölümde yarı parametrik tahminin ikinci aşamasını oluşturan parametrik olmayan regresyon tahmin edicilerinden NW tahmin sonuçları verilmektedir. Bu sonuçlar belirli hasta karakteristiklerine göre hastalığın nüksetme olasılıklarını ortaya koymaktadır. NW tahmin yönteminde, Gaussian ve Epanechnikov çekirdek fonksiyonlarının her ikisi için de tahminler elde edilmiştir. Böylece farklı çekirdek fonksiyonlarına göre sonuçların nasıl değiştiği de ortaya çıkarılmış olmaktadır.

Çizelge 3'de modelin uyum kalitesini gösteren ölçütler verilmektedir. Hata Kareler Ortalaması (HKO), doğru tahminden sapmaların bir göstergesidir ve olabildiğince küçük olması istenir. Doğru Sınıflama Oranı (DSO) ise modelden doğru tahmin edilme yüzdelerini vermekte ve olabildiğince büyük olması beklenmektedir.

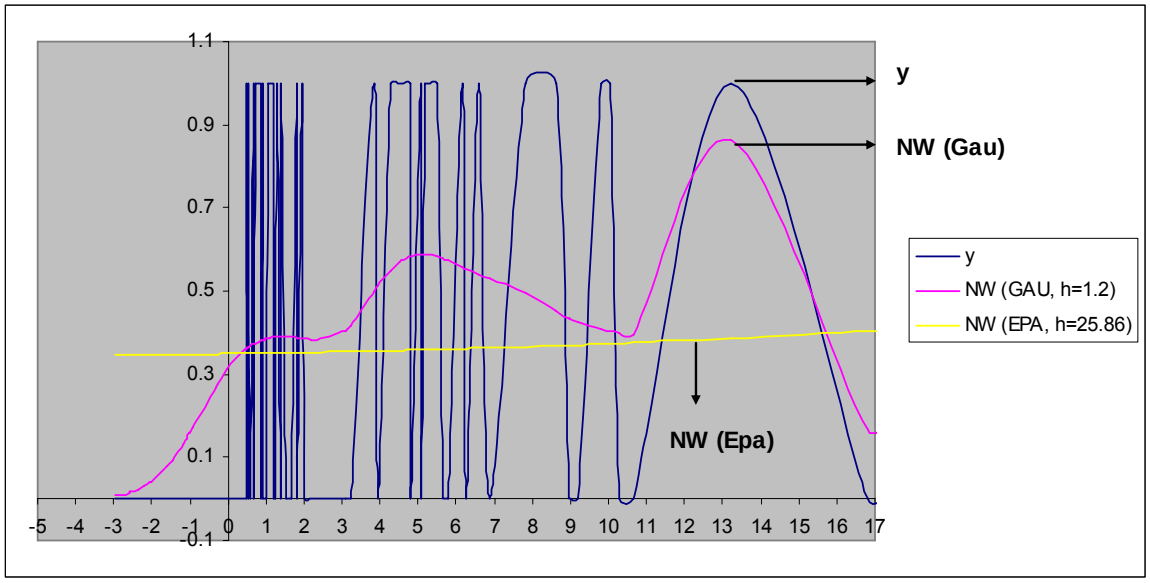
Çizelge 3. HKO ve DSO sonuçları

Çekirdek Fonksiyonu		HKO ve DSO değerleri	Optimal bant genişliği (h)
GAUSSIAN	HKO	0.156501	1.20
	DSO	0.694737	
EPANECNIKOV	HKO	0.221749	25.86
	DSO	0.652632	

Sonuçlar incelendiğinde, Gaussian çekirdek fonksiyonu kullanılarak yapılan parametrik olmayan regresyon fonksiyonu tahmin sonuçlarının, Epanechnikov çekirdek fonksiyonu sonuçlarına göre HKO'larının daha düşük olduğu (Gau_0.156501; Epa_0.221749); DSO'lar incelendiğinde ise yine Gaussian çekirdek fonksiyonu kullanılarak yapılan tahminlerin doğru sınıflama yüzdesinin, Epanechnikov sonuçlarına oranla daha yüksek olduğu görülmüştür (Gau_0.694737; Epa_0.652632).

Örnek olması bakımından veri kümesindeki ilk hasta için aşağıda olasılık tahminleri verilmiş ve yorumlar yapılmıştır. Bu tahminler daha önce de belirtildiği gibi NW tahmin edicisinden elde edilen sonuçlardır. Buna göre, hastalığında metastaz ve kilo kaybı görülen, alkol kullanmayan, radyoterapi almayan ve cea serum miktarı 2 olan bir hastanın hastalığının nüksetmesi olasılığı Gaussian çekirdek fonksiyonuna göre 0.450176887; Epanechnikov çekirdek fonksiyonuna göre ise 0.368595121'dir.

Toplam 95 hasta için hesaplanan doğrusal indeks değerlerinin olasılık değerlerine karşı grafiği, hem Gaussian hem de Epanechnikov çekirdek fonksiyonları için Şekil 1'de verilmiştir. Yatay eksenler indeks değerlerini, dikey eksenler ise karşılık gelen olasılık tahminlerini göstermektedir. Şekilden de anlaşıldığı gibi sonuçlar çekirdek fonksiyonlarına göre büyük ölçüde farklılık göstermektedir. Epanechnikov çekirdeği için tahmin edilen bant genişliği parametresinin çok büyük olduğu (25.86) ve aşırı düzleştirme yaptığı; Gaussian çekirdeği için tahmin edilen bant genişliğinin çok daha düşük olduğu (1.2) ve dağılımı çok fazla bozmadan yakın tahminler verdiği gözlenmiştir. Her iki çekirdek fonksiyonuna göre elde edilen tahminlerin HKO ve DSO'ları dikkate alındığında bu veri kümesi için Gaussian çekirdeği ile elde edilen sonuçların daha doğru olacağı sonucuna ulaşılmıştır.



Şekil 1. İki farklı çekirdek fonksiyonu için NW olasılık tahminleri

4. Sonuç

Bu çalışmada, bağımlı değişkenin iki düzeyli kategorik bir değişken olması durumunda kullanılan parametrik, parametrik olmayan ve yarı parametrik yaklaşımlar kısaca tanıtılmış ve özellikle yarı parametrik yöntem ile ilgili ayrıntılı teorik bilgilere yer verilmiştir.

Çalışmanın uygulama bölümünde 95 hasta bilgisini içeren mide kanseri verisi kullanılarak yarı parametrik modelin uygulanabilirliği gösterilmiştir. Tahminin ilk aşamasında belirli karakteristiklere göre hastalığın nüksedip etmemesine etki eden önemli faktörler belirlenmiştir. İlk olarak doğrusal indeks fonksiyonu KS yaklaşımı ile tahmin edilmiş ve sonuçlar yorumlanmıştır. Buna göre, metastazın olması, kilo kaybı ve radyoterapi tedavi türünün 0.05 yanılma düzeyinde istatistiksel olarak önemli bulunduğu görülmüştür. Ayrıca, metastazın var olması ve kilo kaybının olması ile ilgili tahmin edilen katsayılarının pozitif olduğu ve hastalığın nüksetme ihtimalini artırdığı, radyoterapi uygulamasının ise negatif işaretli bir katsayıya sahip olduğu ve hastalığın nüksetme ihtimalini azalttığı sonucuna ulaşılmıştır.

İkinci aşamada veri kümesindeki tüm hastalar için hastalığın nüksetme olasılıkları hesaplanmıştır. Bu aşama için parametrik olmayan regresyon tahmin edicilerinden NW çekirdek kestiricisi kullanılmıştır. Gaussian ve Epanechnikov çekirdek fonksiyonları için ayrı ayrı tahminler elde edilmiştir.

Gaussian çekirdek fonksiyonu kullanılarak elde edilen tahminlerin Epanechnikov çekirdeği kullanılarak elde edilenlere oranla daha düşük HKO ve daha yüksek DSO'ya sahip olduğu görülmüştür. Örnek olması bakımından veri kümesindeki ilk hasta bilgisi kullanılarak iki farklı çekirdek fonksiyonuna göre hastalığın nüksetme olasılığı tahmin edilmiş ve yorumlanmıştır. Her iki çekirdek fonksiyonu için elde edilen tahminler grafiklenmiş, sonuçların çekirdek fonksiyonlarına göre önemli derecede farklılık gösterdiği görülmüş ve modelin uyum kalitesini gösteren ölçütler dikkate alındığında, Gaussian çekirdeği ile elde edilen sonuçların yorumlanmasının daha doğru olduğu belirlenmiştir.

Sonuç olarak bu çalışmada, parametrik model varsayımlarının doğruluğu kontrol edilmeden elde edilen sonuçlar yanıltıcı olabileceğinden dolayı, doğrusal indeks varsayımı dışında başka ek varsayımlara gereksinim duymayan yarı parametrik yöntemlerin istatistiksel çalışmalarda uygulanabilirliği gösterilmiştir.

Kaynaklar

[1] A. Agresti, (1990), *An Introduction to Categorical Data Analysis*, 1st. ed., John Wiley&Sons, New York.
[2] E. Akkaya, (2008), Mide Kanseri Verileri İçin Cox Regresyon Çözümlemesi, Rapor, Hacettepe Üniversitesi Fen Fakültesi İstatistik Bölümü, İleri İstatistik Projeleri, Ankara.
[3] J.H. Aldrich, F. D. Nelson, (1984), *Linear Probability, Logit and Probit Models*, Sage Publications, London.
[4] D. R. Cox, N. Wermuth, (1992), A Comment on the Coefficient of Determination for Binary Responses, *The American Statistician*, 46 (1), 1-4.
[5] W. Hardle, M. Müller, S. Sperlich, A. Werwatz, (2004), *Nonparametric and Semiparametric Models*, Springer-Verlag, New York.
[6] J. L.Horowitz, (1998), *Semiparametric Methods in Econometrics*, Springer-Verlag, New York.
[7] J. L. Horowitz, (1993), Semiparametric Estimation of a Work-Trip Mode Choice Model, *Journal of Econometrics*, 58, 49-70.
[8] W. Klein, R. H. Spady, (1993), An Efficient Semiparametric Estimator for Binary Response Models, *Econometrica*, 61, 387-421.
[9] E. A. Nadaraya, 1964, On Estimating Regression, *Theory of Probability and its Applications*, 10, 186-190.
[10] D. A. Powers, Y. Xie, (2000), *Statistical Methods for Categorical Data Analysis*, Academic Press.
[11] I. Proença, A. Werwatz, (1994), *Comparing Parametric and Semiparametric Binary Response Models*, Sonderforschungsbereich, Humboldt Universitaet, Berlin.
[12] I. Proença, S. Silva, 2001, Parametric and Semiparametric Specification Tests for Binary Choice Models: A Comparative Simulation Study, *Econometrics*, Econ WPA, No: 0508008.
[13] G. S. Watson, (1964), Smooth Regression Analysis*, Sankhya*, Series A, 26, 359-72.