T.C.
MARMARA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

NONLİNEER LOJİSTİK REGRESYON VE UYGULAMASI

Esra Zeynep ŞENSOY

YÜKSEK LİSANS TEZİ
MATEMATİK ANABİLİM DALI
UYGULAMALI MATEMATİK PROGRAMI

DANIŞMAN
Prof. Dr. Müjgan TEZ

İKİNCİ DANIŞMAN
Prof. Dr. Aydın ERAR

İSTANBUL 2009

T.C.
MARMARA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

NONLİNEER LOJİSTİK REGRESYON VE UYGULAMASI

Esra Zeynep ŞENSOY

YÜKSEK LİSANS TEZİ
MATEMATİK ANABİLİM DALI
UYGULAMALI MATEMATİK PROGRAMI

DANIŞMAN
Prof. Dr. Müjgan TEZ

İKİNCİ DANIŞMAN
Prof. Dr. Aydın ERAR

İSTANBUL 2009

T.C.
MARMARA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

KABUL ve ONAY BELGESİ

Esra Zeynep ŞENSOY'un **Nonlineer Lojistik Regresyon ve Uygulaması** başlıklı Lisansüstü tez çalışması, M.Ü. Fen Bilimleri Enstitüsü Yönetim Kurulu'nun 29.06.2009 tarih ve 2009/15-09 sayılı kararı ile oluşturulan jüri tarafından Matematik Anabilim Dalı Uygulamalı Matematik Programında YÜKSEK LİSANS Tezi olarak kabul edilmiştir.

Danışman: Prof.Dr. Müjgan TEZ (Marmara Ünv.).....

1. Üye : Yard.Doç.Dr. Birsen E. ERDOĞAN (Marmara Ünv.).....

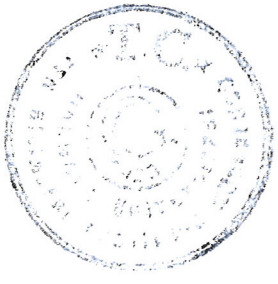
2. Üye : Yard.Doç.Dr. Dursun ÜSTÜNDAĞ (Marmara Ünv.).....

Tezin Savunulduğu Tarih : 22/07/2009

ONAY

M.Ü. Fen Bilimleri Enstitüsü Yönetim Kurulu'nun 28.09.2009 tarih ve 2009/23-32 sayılı kararı ile Esra Zeynep ŞENSOY'un Matematik Anabilim Dalı Uygulamalı Matematik Programında Y.Lisans (MSc.) derecesi alması onanmıştır.

Fen Bilimleri Enstitüsü Müdürü
Prof. Dr.Sevil ÜNAL



TEŞEKKÜR

Çalışmalarımı yakından izleyerek beni yönlendiren ve her türlü yardım ve desteği sağlayan danışman hocalarım sayın Prof. Dr. Müjgan TEZ ve sayın Prof. Dr. Aydın ERAR'a, uygulama çalışmasında istatistiksel program kullanımında her daim yardımcı olan Araş. Gör. Barış AŞIKGİL'e, benden hiçbir zaman desteğini esirgemeyen sevgili aileme ve yüksek lisans arkadaşım Bürya ÖZKARTAL'a en içten dileklerimle teşekkür etmeyi bir borç bilirim.

İÇİNDEKİLER

SAYFA

ÖZET

NONLİNEER LOJİSTİK REGRESYON VE UYGULAMASI

Lojistik Regresyon, bağımsız değişkenlerin bağımlı değişkenler üzerindeki etkisini olasılık olarak ortaya koyar. Risk faktörlerinin olasılık olarak belirlenmesini sağlar. Lojistik Regresyonun amacı, en az değişkeni kullanarak en iyi uyuma sahip olacak şekilde yanıt değişkeni ile bağımsız değişkenler arasındaki ilişkiyi tanımlayabilmek ve amaca yönelik kabul edilebilir model kurmaktır. Lojistik Regresyonda neden sonuç ilişkisinin ortaya konulması amacıyla, bağımlı değişken olumlu-olumsuz, başarılı-başarısız gibi kategorik olarak ikili (binary) kodlanmıştır. Bu yöntem, sayısal verilerle kolay yorumlanabildiği için popülerdir. Bu nedenlerle son zamanlarda epidemiyolojik çalışmalarda, biyolojide, ziraatte, taşımacılıkta, ekonomi gibi alanlarda yaygın bir şekilde kullanılmaktadır.

Lojistik Regresyon modelinde lojistik fonksiyonun doğal logaritmasının alınmasıyla elde edilen lojit fonksiyon her zaman doğrusal olmayabilir. Bazen verinin doğrusallığa uygun olmayışından bazen de model denkleminin istenilen sonucu iyi karşılayamamasından kaynaklanan bu durum model denkleminin karesel ya da kübik olarak biçimlendirilmesiyle çözümlenebilir. Bu durumda splayn foksiyonlarla lojistik regresyon çalışılacaktır.

Bu çalışmada, lojistik regresyon modelinin kurulması ve analizi başlığı altında parametre tahmin yöntemleri, parametrelerin önem testi açıklanmıştır. Modelde bağımsız değişken sayısının birden fazla olması durumunda 'Çoklu Lojistik Regresyon Model incelenmiştir. Lojistik regresyon için katsayıların yorumlanması stratejileri açıklanmıştır.

Lojistik regresyon fonksiyonu doğrusal olmadığında kullanılabilecek yöntemler hakkında bilgiler verilmiştir. Buna göre çalışmamızda yer alan segmentli fonksiyonlardan bahsedilerek splayn fonksiyonlar açıklanmıştır.

Uygulama çalışmamızda, 243 kişilik veri kümesinde hemoglobin kan değeri incelenerek ilgili risk durumu üzerine Doğrusal Olmayan Lojistik Regresyon Modeli

kurulmuştur. SAS version 9.1 kullanılarak model analiz edilmiştir. Ayrıca segmentli lojistik regresyona bir başka örnek olacak Mulla'nın albümin çalışması Matlab programına uyarlanarak incelenmiştir.

Son olarak her iki uygulama çalışmasının sonuçları, farklı istatistiksel programlarda değerlendirilmiştir.

Haziran, 2009 **Esra Zeynep ŞENSOY**

ABSTRACT

NONLINEAR LOGISTIC REGRESSION AND ITS APPLICATION

Logistic Regression produces effects of independent variables on dependent variables as probability. Risk factors could be determined as probability by Logistic Regression. The aim of Logistic Regression Analysis is to establish the best acceptable model with least variable, which gives the relationship between outcome and independent variables. In Logistic Regression, dependent variable is coded binary as positive-negative, successful-unsuccessful for showing dose-response relationship. This method is popular because of easy interpretation by numerical data. Thus it is commonly used in many fields including business and finance, ecology, health policy, agriculture, biology and transportation.

Logit function obtained by computing the natural log of the logistic function in Logistic Regression model can not always be linear. Sometimes data is not available as linear. Sometimes model equation can not satisfy the probable cause. This situation can be analyzed as quadratic or cubic shape. At this stage, logistic regression is studied with spline functions.

At this study, the title under building logistic regression models and analysis, parameter estimation method and test of significance of parameters are explained. When model has more than one variable, 'Multiple Logistic Regression Model' is investigated. For logistic regression, strategies for interpretation coefficients are explained.

The information is given about that the method can be used, when Logistic Regression isn't linear. According to that, spline functions are explained to talk about segmented functions in our study.

In our application work, blood hemoglobin values in the data set of 243 people were examined. The Nonlinear Logistic Regression Model was established based on the risk status. The model is analyzed to use SAS version 9.1.

Additionally, Mulla's albumin study was adapted to Matlab programme as an example of segmented logistic regression.

Finally, the results of both application studies are evaluated by different statistical programmes.

June, 2009 **Esra Zeynep ŞENSOY**

$\mathbf{y}$	: Bağımlı Değişken (Yanıt Değişkeni)
$\mathbf{y_i}$	: i.Deneğin Bağımlı Değişkeninin Değeri
$\boldsymbol{\pi(x_i)}$	: Lojistik Fonksiyon (i. gözleme ait lojistik fonksiyon değeri)
$\boldsymbol{\hat{\pi}(x_i)}$	: Lojistik Fonksiyon Kestirimi (i. gözlem için)
$\boldsymbol{\beta_0}$	: Herhangi Bir Parametre
$\boldsymbol{\beta_1}$	: Eğim Parametresi
$\boldsymbol{\beta}'$	: Parametreler vektörü
$\boldsymbol{\hat{\beta}_0}$	: β_0 Parametresinin Kestirimi
$\boldsymbol{\chi^2}$	: Ki-kare İstatistiği
$\mathbf{100(1-\alpha)\%CI}$	: Güven Aralığı

KISALTMALAR

LR : Lojistik Regresyon

Var : Varyans

HGB : Hemoglobin

EKK : En Küçük Kareler Kestirimi

ML : En Çok Olabilirlik (Maksimum Likelihood)

IRLS : İteratif Olarak Ağırlıklandırılmış En Küçük Kareler

SEMBOLLER

D	: Sapma (Deviance) İstatistiği
E(y/*x*) = π (*x*)	: Koşullu Ortalama (x değeri için y'nin beklenen değeri)
f(*z*)	: Lojistik Fonksiyon (basit form)
g(*x*)	: Lojit Fonksiyon
G	: Hosmer-Lemeshow İstatistiği
h_i	:Leverage Uzaklık Ölçütü
I(β)	: Fisher Bilgi Matrisi
J	: Jacobiyen Matris
J^T	: Jacobiyen Matrisinin Transpozu
L(.)	: Olabilirlik Fonksiyonunun Değeri
OR	: Odds Oranı
$\widehat{OR}$	: Odds Oran Kestirimi
P	: Olasılık Fonksiyonu
Pr	:Anlamlılık Test Ölçütü
ROC	: Receiver Operating Characteristic
SE	: Standart Hata
$\widehat{SE}$	: Standart Hata Kestirimi
SSE	: Artık Kareler Toplamı
SSR	: Regresyon Kareler Toplamı
ST	: Skor Testi
n	: Gözlem Sayısı
V(.)	: Varyans
Var(β)	: β Parametresinin Varyansı
w_i	: i için Ağırlık Fonksiyonu
W	: Wald İstatistiği
ε	: Hata
x	: Bağımsız Değişken
x_i	: i.Deneğin Bağımsız Değişkeninin Değeri

ŞEKİL LİSTESİ

SAYFA NO

TABLO LİSTESİ

SAYFA NO

BÖLÜM I

GİRİŞ

Regresyon analizi bir bağımlı değişken ile bir veya birden fazla bağımsız değişken arasındaki bağıntının açıklanması sürecidir.

Regresyon analizinde bağımsız değişkenler (X), nicel veya nitel olabilir. Bağımlı değişkenin nicel veya nitel olması, kullanılan yöntemin çözümü ve yorumlanmasını etkiler. Bağımlı değişkenin nitel olması durumunda Lojistik Regresyon Analizi uygun bir yöntemdir.

Lojistik Regresyon Analizinin kullanım amacı, İstatistik'teki diğer model yapılandırma teknikleri ile aynıdır: En az değişkeni kullanarak en iyi uyuma sahip olacak şekilde bağımlı ile bağımsız değişkenler arasındaki ilişkiyi tanımlamaktır [3].

I.1 LOJİSTİK REGRESYON MODELİ

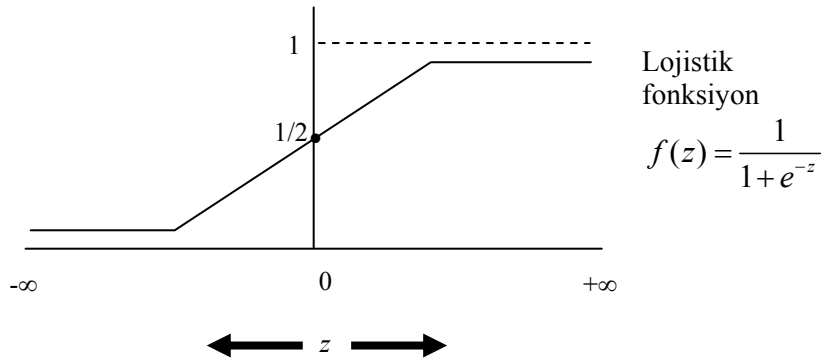
Lojistik Regresyon yöntemi bir veya birden fazla açıklayıcı değişken ile yanıt değişkeni arasındaki ilişkileri inceler. Yanıt değişkeni kesikli olup, iki veya daha fazla olası değere bağlıdır. Bu yöntemde, açıklayıcı değişkenlerin bağımlı değişkenler üzerindeki etkileri olasılık olarak hesaplanarak risk faktörlerinin olasılık olarak belirlenmesi sağlanır [15, 18].

Alışılagelen doğrusal regresyon analizinde bağımlı değişkenin değeri kestirilirken, lojistik regresyon analizinde bağımlı değişkenin alacağı değerlerden birinin gerçekleşme olasılığı kestirilir.

Lojistik modelin matematiksel formu,

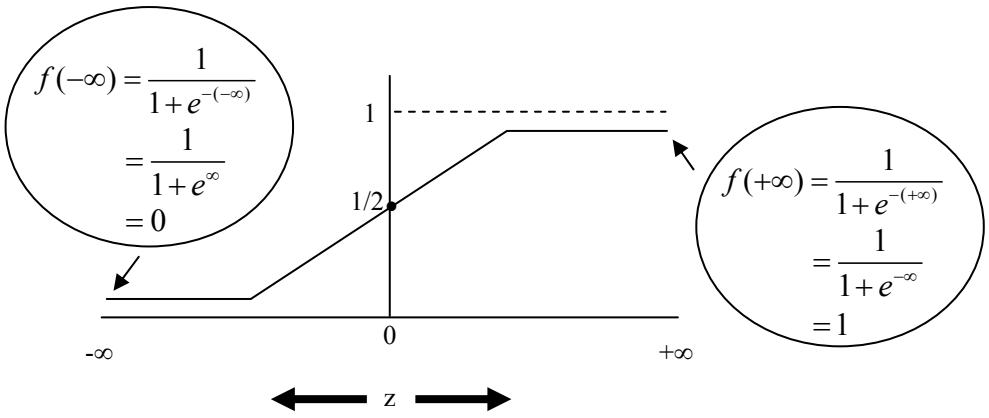
$$f(z)=\frac{1}{1+e^{-z}}$$

şeklinde tanımlanır.



Şekil I.1 Lojistik Fonksiyonun Grafikte Gösterilmesi

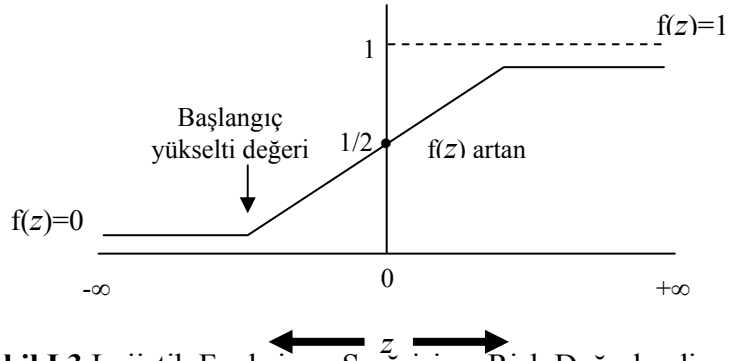
Şekil I.1 ve Şekil I.2'den görüldüğü gibi *z*' nin tanım aralığı (-∞, +∞) dır. *z*' nin değeri ne olursa olsun f(*z*) fonksiyonunun değişim aralığı 0 ile 1 arasında kalır.



Şekil I.2 Lojistik Fonksiyondaki *z* Değerinin Tanım Aralığı

f(*z*) lojistik fonksiyonunun 0 ile 1 arasında bir değişime sahip olması lojistik fonksiyonun tercih edilmesindeki önemli bir nedendir; çünkü model 0 ile 1 arasında yer alan olasılıklar üzerine kurulmuştur. Böylece 0-1 aralığı dışında lojistik model için bir risk kestiricisi ortaya çıkmaz.

Lojistik modelin tercih edilmesinin diğer bir nedeni de lojistik fonksiyonun biçimidir. Şekil I.3'te görüldüğü gibi lojistik fonksiyon S-şeklinde bir sigmoid eğri meydana getirir. S şekilli lojistik fonksiyonda z, çeşitli risk faktörlerinin katılımını gösteren bir indeks olarak kabul edilirse f(z) de z değerindeki riski gösterir. Şekil I.3' den de anlaşılacağı gibi yükselti değerine kadar bireyin riski minimumdur. Sonra risk ortadaki z değerlerinde hızla artmakta ve z yeteri kadar arttığında 1 civarında kalmaktadır.



Şekil I.3 Lojistik Fonksiyon S-eğrisi ve Risk Değerlendirme Şekli

Yükselti değeri, epidemiologlar tarafından örneğin, hastalık koşullarının değişikliğini belirtmek için kullanılmıştır [11].

Lojistik modelde yanıt değişkeninin kesikliliği nedeni ile yanıt değişkeninin açıklayıcı değişken ile ilişkisi X-Y çiziminden açıkça görülemez. Bunun için yanıt değişkeni yerine f(z) olasılık değerine karşı çizim yapılır; f(z) fonksiyonu sürekli olup Şekil I.3'teki gibidir [2].

k sayıda bağımsız (açıklayıcı) değişken varlığında lojistik fonksiyon,

$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$

ve β_0 ve β_i bilinmeyen parametreler olmak üzere,

$$f(z) = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-(\beta_0+\Sigma\beta_i X_i)}}$$

biçiminde yazılır.

Öte yandan lojistik regresyon fonksiyonu x'deki her birim değişme sonucunda $\mathrm{E}(\mathrm{Y}/x)$'de oluşan değişikliği gösterir. Böylece lojistik regresyon modeli $\pi(x)=\mathrm{E}(\mathrm{Y}/x)$ iken,

$$\pi(x)=\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \tag{I.1}$$

$\pi(x)/(1-\pi(x))$

biçiminde verilir [8].

Koşullu ortalama "lojit dönüşüm" adı verilen,

$$g(x)=\ln\left[\frac{\pi(x)}{1-\pi(x)}\right]=\ln\left[\frac{\mathrm{P}(y=1|x)}{\mathrm{P}(y=0|x)}\right]=\beta_0+\beta_1 x \tag{I.2}$$

dönüşümüyle doğrusal biçime getirilebilir. Eşitlikteki $\pi(x)/(1-\pi(x))$ oranı odds olarak adlandırılır.

Bu dönüşümün önemi, $g(x)$'in diğer regresyon modellerinde istenen çoğu özelliği taşımasıdır. Lojit $g(x)$, parametreleri bakımından doğrusaldır. Lojit $g(x)$ süreklidir ve x'in tanım aralığına göre 0 ve 1 arasında değer alabilir.

Lojistik modelin yanıt değişkenini oluşturan lojit dönüşüm $g(x)$'in özellikleri aşağıdaki şekilde sıralanabilir:

i) $\pi(x)$ arttıkça $g(x)$ de artar.

ii) $\pi(x)$ yani lojistik fonksiyon 0 ve 1 arasında iken $g(x)$ tüm gerçel doğru üzerinde değerler alır.

iii) Eğer $\pi(x)<0,5$ ise $g(x)<0$

Eğer $\pi(x)>0,5$ ise $g(x)<0$'dır [2].

Lojistik regresyon modelinde yanıt değişkeninin gözlem değeri $y=E(Y/x)+\varepsilon$ ile gösterildiğinde, ε hata terimi olarak adlandırılır ve gözlemin koşullu ortalamadan sapma miktarını ifade eder. ε'nin klasik modelde sıfır ortalama ve σ^2 sabit varyansı ile normal dağılıma sahip olması önemli bir varsayımdır. Fakat iki sonuçlu bağımlı değişken için, x verildiğinde yanıt değişkeni $y=\pi(x)+\varepsilon$ ile gösterilir ve ε'un bir veya iki olası değeri vardır. Eğer,

y=1 ise $\pi(x)$ olasılığıyla, ε =1-$\pi(x)$,

y=0 ise $\pi(x)$olasılığıyla, ε = -$\pi(x)$ olur. Böylece ε, sıfır ortalama ve $\pi(x)[1-\pi(x)]$varyansına sahip bir dağılım gösterir:

$$E(\varepsilon)=\left[\left(1-\pi(x)\right)\pi(x)\right]+\left[-\pi(x)\left(1-\pi(x)\right)\right]=0$$

$$V(\varepsilon)=E\left(\varepsilon^2\right)-\left[E(\varepsilon)\right]^2=E\left(\varepsilon^2\right)$$

$$=\left[\left(1-\pi(x)\right)^2\pi(x)\right]+\left[\left(-\pi(x)\right)^2\left(1-\pi(x)\right)\right]$$

$$=\left[\left(1-2\pi(x)+\pi(x)^2\right)\pi(x)\right]+\left[\left(\pi(x)\right)^2\left(1-\pi(x)\right)\right]$$

$$=\left[\pi(x)-\pi(x)^2\right]$$

$$=\pi(x)\left[1-\pi(x)\right]$$

Böylece yanıt değişkeninin koşullu dağılımı, $\pi(x)=E(\mathrm{y}/x)$ koşullu dağılımına göre bir binom dağılımıdır.

Özet olarak, yanıt değişkeninin iki düzeyli olması durumunda regresyon analizinde:

1) Regresyon eşitliğindeki koşullu ortalama 0 ve 1 arasında bir değer olmalıdır.
2) Normal dağılımlar değil de binom dağılımları hatanın dağılımını tanımlar ve analiz bunun üzerine kuruludur.
3) Doğrusal regresyonda kullanılan ilkeler, lojistik regresyon analizinde de yol göstericidir [8].

I.2 LOJİSTİK REGRESYONUN LİNEER REGRESYONDAN FARKI ve KULLANIM AMAÇLARI

Lojistik regresyonda, doğrusal regresyon analizinde olduğu gibi bazı değişken değerleri göz önüne alınarak tahmin yapılmaya çalışılır. Fakat bu iki yöntem arasında üç önemli fark vardır:

a) Doğrusal regresyon analizinde tahmin edilecek bağımlı değişken sürekli iken, Lojistik Regresyon analizinde bağımlı değişkenler kesikli bir değer almaktadır.

b) Doğrusal regresyon analizinde bağımlı değişkenin değeri tahmin edilirken, lojistik regresyon analizinde ise bağımlı değişkenin alabileceği değerlerden birinin gerçekleşme olasılığı tahmin edilir.

c) Doğrusal regresyon analizinde bağımsız değişkenin çoklu normal dağılım göstermesi koşulu aranırken lojistik regresyon analizinde böyle bir şart yoktur [3].

Öte yandan Lojistik Regresyon Analizi, diskriminant analizi ve çapraz tablo uygulamalarına alternatif olarak da uygulanmaktadır.

Lojistik regresyon analizinin güncel olmasının nedenleri şöyle özetlenebilmektedir:

a) Yanıt değişkeni kesikli iken açıklayıcı değişkenlerin hem sürekli hem de kesikli olma durumlarında uygulanabilmektedir.

b) Lojistik modelin parametreleri epidemioloji de yapılan ölçümlere benzediği için yorumları kolay olmaktadır. Epidemiolojide katsayıların exponansiyeli hastalık riski olarak yorumlanmaktadır.

c) Lojistik modelin parametre sayısı karşılık gelen doğrusal regresyon modeli ve diskriminant fonksiyonu ile aynı olmaktadır.

d) Lojistik modele dayalı analizler için standart paket programlar vardır

e) Açıklayıcı değişkenlerin olasılık fonksiyonlarının dağılımı üzerinde kısıt olmaması (yarı parametrik) nedeni ile çeşitli testler uygulanabilmektedir.

Epidemioloji ve diğer medikal uygulamaların yanı sıra deneysel verilerin analizinde, askeri konularda, meteorolojide, ziraatte, taşımacılıkta, ekonomi v.b. alanlarda sıkça kullanılan lojistik regresyon analizi farklı varsayımlar durumunda aynı lojistik formülasyona götürdüğü için varsayım bozulmalarına karşı daha güçlü bir yöntemdir [2,7].

Bu çalışmada, Lojistik Regresyon Modeli'nin kurulması, parametrelerin kestirim yöntemleriyle elde edilmesi ve kestirilen bu parametrelerin anlamlılığının incelenmesi açıklanacaktır. Kestirilen katsayılara ilişkin standart hatalar oldukça yüksek çıkıyorsa, önemli olduğu bilinen değişkenler önemsiz gözüküyorsa ve regresyon modelinin veriyi iyi karşılayamaması durumunda Doğrusal Olmayan Lojistik Regresyon Modeli'nin kullanılması önerilir. Bu amaçla Doğrusal Olmayan Lojistik Regresyon ve karşılaşılan model yapıları anlatılacaktır.

BÖLÜM II

LOJİSTİK REGRESYON MODELİNİN KURULMASI VE ANALİZİ

(x_i, y_i), i=1, 2, ..., n, n tane bağımsız gözlem olsun. y_i iki düzeyli yanıt değişkeni ve x_i de i'ninci denek için bağımsız değişken değerini temsil etsin. Yanıt değişkeninin 0 ve 1 kodlanmasının, belirli bir riskin yokluğunu ve varlığını temsil ettiği varsayılsın. $\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$ lojistik regresyon denklemi için bilinmeyen β_0 ve β_1 parametrelerinin kestirilmesi gerekir. Bu bölümde bu parametrelerin kestirimi, modele katkıları, modelin anlamlılığı analiz incelenecektir.

II.1 PARAMETRE KESTİRİM YÖNTEMLERİ

II.1.1 En Çok Olabilirlik Yöntemi

Regresyon analizlerinde, hataların normal dağılması, ortalamanın sıfır olması ve bağımsız değişkenin her bir seviyesi için varyansın sabit kalması önemli varsayımlardır. Fakat iki yanıtlı model olan lojistik regresyon analizinde bu varsayımlar geçerli değildir. Bu durumunda $Y = \pi(x) + \varepsilon$ modelinde yer alan hata terimi, Y=1 ve Y=0 durumunda iki olasılık değerinden birini alır: Y=1 ise $\pi(x)$ olasılığıyla $\varepsilon = 1 - \pi(x)$; Y=0 ise $1-\pi(x)$ olasılığıyla $\varepsilon = -\pi(x)$'dir. Böylece ε'nin ortalamasının sıfır ve varyansının $\pi(x)\left[1-\pi(x)\right]$ olduğu; bağımlı değişkenin(y) binomiyal dağılım gösterdiği görülür [3].

En çok olabilirlik yöntemi, olabilirlik fonksiyonu maksimum olacak biçimde bilinmeyen parametrelerin kestirimidir.

Y, 0 ve 1 değerlerini almış iken,

- $\pi(x)$, x verildiğinde Y'nin 1'e eşit olma koşullu olasılığıdır ve P(Y=1/ x)= $\pi(x)$ ile gösterilir.
- $1-\pi(x)$, x verildiğinde Y'nin 0'a eşit olma koşullu olasılığıdır ve P(Y=0/ x)=$1-\pi(x)$ ile gösterilir.

$\pi(x_i)$, x_i'deki $\pi(x)$ değerini vermek üzere,

y_i=1 ise (x_i , y_i) gözlem çiftinin olabilirlik fonksiyonuna katkısı $\pi(x_i)$,

y_i=0 ise (x_i , y_i) gözlem çiftinin olabilirlik fonksiyonuna katkısı $1-\pi(x_i)$ olasılığı kadardır. $\boldsymbol{\beta}=(\beta_0\ \beta_1)'$ parametre vektörü iken, (x_i , y_i) ikililerinin olabilirlik fonksiyonuna katkısını göstermek için aşağıdaki yol izlenir:

$$\pi(x_i)^{y_i}\left[1-\pi(x_i)\right]^{1-y_i}$$

Gözlemler birbirinden bağımsız olduğundan olabilirlik fonksiyonu,

$$l(\boldsymbol{\beta})=\prod_{i=1}^{n}\pi(x_i)^{y_i}\left[1-\pi(x_i)\right]^{1-y_i} \qquad \text{(II.1)}$$

olarak bulunur. Log-olabilirlik (log likelihood) fonksiyonu ise,

$$L(\boldsymbol{\beta})=ln\left[l(\boldsymbol{\beta})\right]=\sum_{i=1}^{n}\left\{y_i.ln\left[\pi(x_i)\right]+(1-y_i).ln\left[1-\pi(x_i)\right]\right\} \qquad \text{(II.2)}$$

olur. L($\boldsymbol{\beta}$)'yı maksimum yapan $\boldsymbol{\beta}$ değerini bulmak üzere, L($\boldsymbol{\beta}$)'nin β_0 ve β_1'e göre türevleri alınırsa,

$$\sum_{i=1}^{n}\left[y_i-\pi(x_i)\right]=0 \qquad \text{(II.3)}$$

$$\sum_{i=1}^{n}x_i\left[y_i-\pi(x_i)\right]=0 \qquad \text{(II.4)}$$

"olabilirlik eşitlikleri" elde edilir.

Lojistik regresyonda bu eşitlikler β_0 ve β_1'e göre doğrusal olmayan, üstel denklem olduklarından bu denklemlerin çözümü için iteratif yöntemler kullanılır.

Lojistik modelin en çok olabilirlik kestirimlerini bulmak için iterasyona başlarken başlangıç değerlerini vermenin çeşitli yolları vardır. Bunlardan biri

grafiksel gösterimlerden göz ile kestirimde bulunmaktır. Başlangıç değerlerinin doğruluğu iterasyon sayısı ve kestirimlerin etkinliği üzerinde önemli etkiye sahiptir. İyi bir başlangıç değeri ile az sayıda iterasyon sonucu optimal çözüme ulaşılabilmektedir. Ençok olabilirlik yönteminde her adımda yapılacak düzeltme miktarı(δ), tek açıklayıcı değişken durumunda lojit tablolarından elde edilmektedir [2]. Veriler birbirinden çok ayrık olduğunda en çok olabilirlik kestirimlerinde yakınsama elde edilememektedir [14].

(II.3) ve (II.4) eşitliklerinden elde edilen $\boldsymbol{\beta}$ değeri, en çok olabilirlik kestirimidir ve $\hat{\boldsymbol{\beta}}$ ile gösterilir. $\hat{\pi}(x_i)$ de $\pi(x_i)$'nin en çok olabilirlik kestirimidir ve verilen $x = x_i$ değeri için Y'nin 1'e eşit olma koşullu olasılığının kestirimini verir. (II.4) eşitliği $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\pi}(x_i)$ şeklinde yazılabilir. Bu eşitlik, y'nin gözlenen değerlerinin toplamının, y'nin beklenen değerleri toplamına eşit olduğunu gösterir. Bu özellik sonraki bölümlerde modelin uyumu tartışılırken yararlı olacaktır [8].

En çok olabilirlik ve Newton yöntemi yardımıyla parametre kestirimi aşağıdaki gibi bulunur [14]:

$\pi(x_i) = \dfrac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$ lojistik fonksiyon olsun.

(II.2) nolu denklemin parametrelere göre daha açık biçimde yazılmasıyla, L($\boldsymbol{\beta}$),

$$L(\boldsymbol{\beta}) = \sum y_i.\beta_0 + \sum y_i.\beta_1 x_i - \sum \ln(1+e^{\beta_0+\beta_1 x_i})$$

olur ve parametrelere göre türevleri alınıp sıfıra eşitlenirse,

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_0} = \sum y_i - \sum \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} = \sum y_i - \sum \pi(x_i) = 0$$

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_1} = \sum y_i x_i - \sum x_i \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} = \sum y_i x_i - \sum x_i \pi(x_i) = 0$$

elde edilir. Bu iki denklemin Newton yöntemine göre iteratif olarak çözümüyle kestiriciler bulunur:

$$f_1(\beta_0,\beta_1) = \sum y_i - \sum \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} = \sum y_i - \sum \pi(x_i)$$

$$f_2(\beta_0,\beta_1) = \sum y_i x_i - \sum x_i \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} = \sum y_i x_i - \sum x_i \pi(x_i)$$

$$J(\beta)=\begin{bmatrix}\frac{\partial f_1}{\partial \beta_0} & \frac{\partial f_1}{\partial \beta_1}\\ \frac{\partial f_2}{\partial \beta_0} & \frac{\partial f_2}{\partial \beta_1}\end{bmatrix}$$

$$\frac{\partial f_1}{\partial \beta_0}=-\sum\left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)\left(\frac{1}{1+e^{\beta_0+\beta_1 x_i}}\right) \qquad \frac{\partial f_1}{\partial \beta_1}=-\sum x_i\left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)\left(\frac{1}{1+e^{\beta_0+\beta_1 x_i}}\right)$$

$$=-\sum \pi(x_i)\left(1-\pi(x_i)\right) \qquad =-\sum x_i\pi(x_i)\left(1-\pi(x_i)\right)$$

$$\frac{\partial f_2}{\partial \beta_0}=-\sum x_i\left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)\left(\frac{1}{1+e^{\beta_0+\beta_1 x_i}}\right) \qquad \frac{\partial f_2}{\partial \beta_1}=-\sum x_i^{\,2}\left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)\left(\frac{1}{1+e^{\beta_0+\beta_1 x_i}}\right)$$

$$=-\sum x_i\pi(x_i)\left(1-\pi(x_i)\right) \qquad =-\sum x_i^{\,2}\pi(x_i)\left(1-\pi(x_i)\right)$$

$$J=-\begin{bmatrix}\sum \pi_i(1-\pi_i) & \sum x_i\pi_i(1-\pi_i)\\ \sum x_i\pi_i(1-\pi_i) & \sum x_i^{\,2}\pi_i(1-\pi_i)\end{bmatrix}$$

Jakobiyen matrisi oluşturulduktan sonra iteratif çözüm vektörüyle iterasyon denklemi çözülür:

$$\beta^{(k+1)}=\beta^{k}-J^{-1}(\beta^{(k)}).F(\beta^{(k)}) \qquad \text{(II.5)}$$

II.1.2 Yeniden Ağırlıklandırılmış İteratif En Küçük Kareler Yöntemi

Lojistik regresyonda parametre tahmininde kullanılan bir başka yöntem de iteratif olarak ağırlıklandırılmış en küçük kareler yöntemidir. Bu yöntem, hata terimlerine ilişkin varyansların eşit olmadığı zamanlarda, değişen varyanslılık durumu söz konusu olduğunda doğru tahmin sağlar.

Gruplandırılmış verilerde J grubun her birinde n_j denemeden r_j başarı elde edildiğinde, başarı oranı $P_j=\frac{r_j}{n_j}$ şeklinde gösterilsin. Varyansı $\sigma_j^2=\mathrm{Var}(r_j/P_j)=\mathrm{Var}(P_j)=P_j.\left(\frac{1-P_j}{n_j}\right)$ olduğundan her binom dağılımlı gözlem için varyans değişmektedir. Bu durumda lojit(r_j/n_j)'nin açıklayıcı değişkenler üzerinde, $w_j=\frac{n_j}{P_j(1-P_j)}$ ağırlığı ile ağırlıklandırılmış regresyonu uygulanmalıdır. Fakat w_j ağırlık değerleri de P_j'nin bir fonksiyonu olduğu için en küçük kareler yöntemi

iteratif olarak uygulanarak ağırlık değerleri her adımda yeniden elde edilmek suretiyle çözüme ulaşılır [4 , 23].

II.1.3 Minimum Logit Ki-Kare Yöntemi

Ağırlıklı en küçük kareler kestirim yönteminin özel bir biçimidir. Berkson'un (1955) geliştirdiği bu yöntemde, 2xj çapraz tablolarındaki beklenen ve gözlenen lojit değerler arasındaki farktan yararlanılmaktadır. Bu yöntem tekrarlı veriler olması durumunda kullanılmaktadır.

Veriler J grupta tekrar edildiğinde ve her grupta tekrar sayısı çok olduğunda katsayı kestirimleri ağırlıklı en küçük kareler yöntemleri elde edilebilmektedir.

II.1.2 kısmında değinilen yöntemde, P_j başarı olasılığı, lojistik fonksiyon eşitliğinde tanımlandığı gibidir. Bu olasılık üzerinde yapılan lojit dönüşüm, bu yöntemde yanıt değişkenini oluşturmaktadır. Kestirimde kullanılan ağırlık değerleri $n_j\ P_j\ (1\text{-}P_j)$ olarak elde edilmektedir. Yöntem, lojit değeri olarak tanımlanan yanıt değişkeninin açıklayıcı değişkenler üzerindeki ağırlık değeri ile ağırlıklandırılmış regresyonundan EKK kestirimlerini elde etmeye dayanmaktadır. Buradan tek adımda bulunan ağırlıklı EKK kestirimleri minimum lojit Ki-Kare kestirimleri adını almaktadır.

Olasılık değerinin 0 veya 1 olduğu durumda lojit değeri tanımlı olmayacağı için P_j yerine $P_j + 1/2n_j$ değerinin konulduğu ayarlanmış lojit ki-kare yöntemi kullanılmaktadır [2].

II.2 PARAMETRELERİN ÖNEM TESTİ

Lojistik regresyonda katsayıların anlamlılık testleri bağımsız değişkeni içeren ve içermeyen modellerden bulunacak log-olabilirlik fonksiyonu ile yapılmaktadır.

Olabilirlik fonksiyonu kullanarak gözlenen değerlerle kestirilen değerlerin karşılaştırılması,

$$D = -2.ln\left[\frac{\text{Mevcut modelin olabilirliği}}{\text{Doymuş modelin olabilirliği}}\right] \qquad \text{(II.6)}$$

biçiminde verilir. Burada doymuş model veri noktaları kadar parametre içeren modeldir.

Eşitlik (II.6)'te parantez içerisinde verilen ifade olabilirlik oranı olarak bilinir; (-2*ln*) katının alınması, matematiksel olduğu kadar dağılımı bilinen bir değer elde etmektedir. Bu değer hipotez testi amacıyla kullanılmaktadır. Böyle bir test "Olabilirlik Oran Testi " adını alır. Log olabilirlik eşitliği kullanılarak (II.6) eşitliği,

$$D = -2.\sum_{i=1}^{n}\left[y_i ln\left(\frac{\hat{\pi}_i}{y_i}\right) + \left(1-y_i\right) ln\left(\frac{1-\hat{\pi}_i}{1-y_i}\right)\right] \qquad \text{(II.7)}$$

biçiminde elde edilir. Burada $\hat{\pi}_i = \hat{\pi}(x_i)$ 'dir.

(II.7)'deki D istatistiği bazı yazarlar tarafından sapma (deviance) istatistiği olarak adlandırılır ve uyum iyiliğine karar verirken kullanılır [14]. Lojistik regresyon için sapma, doğrusal regresyondaki artık kareler toplamı ile aynı anlamı taşır ve (II.7) eşitliğindeki sapma, doğrusal regresyon için hesaplanırsa $\mathrm{SSE} = \sum_{\mathrm{i=1}}^{\mathrm{n}} (\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i})^2$ 'ye eşit olur.

Bağımsız bir değişkenin anlamlılığına karar vermek için, denklemde bağımsız değişkenin olduğu ve olmadığı durumlardaki D değerleri karşılaştırılır. Modeldeki bağımsız değişkeni kapsamasından dolayı ortaya çıkan D'deki değişim aşağıdaki gibidir:

G = D (değişkensiz model için) – D (değişkenli model için) (II.8)

Bu istatistik lojistik regresyonda, doğrusal regresyonda kullanılan F testinin pay kısmı olan regresyon kareler toplamı ile aynıdır. G'yi hesaplamak için farkı alınacak D değerlerinin her ikisi için de doymuş modelin olabilirlikleri ortak olduğu için G istatistiği şu şekilde ifade edilir:

$$G = -2.ln\left[\frac{\text{Değişkensiz modelin olabilirliği}}{\text{Değişkenli modelin olabilirliği}}\right] \qquad \text{(II.9)}$$

Tek bağımsız değişkenli ve değişkenin modelde olmadığı durumda β_0'ın en çok olabilirlik kestirimi $\ln(n_1/n_0)$'dır. Burada $n_1=\Sigma y_i$ ve $n_0=\Sigma(1-y_i)$ iken, kestirim değeri de n_1/n sabit değeridir. Bu durumda G,

$$\mathrm{G} = -2\ln\left[\frac{\left(\frac{\mathrm{n_1}}{\mathrm{n}}\right)^{\mathrm{n_1}}\left(\frac{\mathrm{n_0}}{\mathrm{n}}\right)^{\mathrm{n_0}}}{\prod_{\mathrm{i=1}}^{\mathrm{n}} \hat{\pi}_\mathrm{i}^{\mathrm{y_i}} (1-\hat{\pi}_\mathrm{i})^{(1-\mathrm{y_i})}}\right] \qquad \text{(II.10)}$$

ya da

$$G = 2\left\{\sum_{i=1}^{n}\left[y_i ln(\hat{\pi}_i) + (1-y_i) ln(1-\hat{\pi}_i)\right] - \left[n_1 ln(n_1) + n_0 ln(n_0) - n ln(n)\right]\right\} \qquad \text{(II.11)}$$

olur. β_1=0 hipotezi altında, G istatistiği 1 serbestlik derecesiyle ki-kare (χ^2) dağılımına sahiptir.

Bağımsız değişkenin modele katkısının önemini test etmek için Wald ve Score testleri de mevcuttur. Bu testler için gerekli varsayımlar, olabilirlik oran testindeki varsayımlarla aynıdır.

Wald testi, eğim parametresinin en çok olabilirlik kestiriminin, onun standart hatasına oranı ile yapılır. Elde edilen oran, β_1=0 hipotezi altında standart normal dağılımı gösterir:

$$W = \frac{\hat{\beta}_1}{\widehat{SE}(\hat{\beta}_1)} \sim N(0,1)$$

Bazı yazarlar Wald testi ile bağımsız değişkenin katsayısı gerçekte önemli olduğu halde test sonucunda katsayının önemsiz çıkabileceğini belirtmişlerdir [5]. Hem G olabilirlik oran testi, hem de W wald testinde β_1 için en çok olabilirlik kestirimlerini hesaplamak gerekir.

Bir başka test de Score testidir. Hesaplama işlemlerini azaltması bu testin en büyük avantajı olmakla beraber, birçok paket programda bulunmaması ise kullanılmasını kısıtlamaktadır [6]. Score testi, log olabilirliğinden türetilen dağılım teorisine dayanmaktadır. Aslında matris hesaplamaları gerektiren çok değişkenli bir testtir.

Tek değişkenli durumda, bu test, (II.3) eşitliğindeki türevlerin verilmesiyle (II.4) eşitliğindeki türevlerin koşullu dağılımına dayanır. Burada β_0 ve β_1 katsayılarının hesaplanmasında (II.4) eşitliği kullanır: $\beta_0=\ln(n_1/n_0)$ ve $\beta_1 = 0$ dır. Bu parametre değerlerine bağlı olarak $\hat{\pi} = n_1 / n = \bar{y}$ olur. Böylece (II.4) eşitliğinin sol tarafı $\sum x_i(y_i - \bar{y})$'ye dönüşür. Kestirilmiş varyans $\bar{y}(1-\bar{y})\sum(x_i - \bar{x})^2$ dir. Score testi (ST) için test istatistiği,

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

olarak verilir [8].

II.3 ÇOKLU LOJİSTİK REGRESYON MODELİ

Bu bölümde, lojistik regresyon modelini birden çok bağımsız değişken olması durumu için kısaca ele alacağız.

X'=(x_1, x_2, ..., x_k) vektörü ile gösterilen k tane bağımsız değişken kümesi olsun. Yanıt değişkeni için P(Y=1/x) = $\pi(x)$ yazılsın. Çoklu lojistik regresyon modelinin lojiti aşağıdaki denklemle ifade edilir:

$$g(x) = \beta_0+\beta_1 x_1+ \beta_2 x_2+...+\beta_k x_k \quad \text{(II.12)}$$

Buradaki lojistik regresyon modeli,

$$\pi(x) = \frac{e^{g(x)}}{1+e^{g(x)}} \quad \text{(II.13)}$$

şeklindedir.

Bağımsız değişkenlerden bazıları nominal ise bunları sürekli değişkenlermiş (aralık ölçekli) gibi kabul ederek modele dahil etmek uygun değildir. Değişkenin farklı düzeylerini göstermek için göstermelik (dummy, kukla) ya da tasarım (desing) değişkenleri kullanılır [8].

Genelde, nominal bir değişken d kategoriye sahipse, d-1 tane göstermelik (D_{ju} , u=1,…,d-1) tanımlanır. Böylece nominal ölçekli değişken içeren modelin lojiti aşağıdaki gibi olur:

$$g(x) = \beta_0 + \beta_1 x_1 + ... + \sum_{i=1}^{d_j-1} \beta_{ju} D_{ju} + \beta_k x_k \quad \text{(II.14)}$$

β'=(β_0,β_1,...,β_k) iken çok değişkenli durumda, parametre kestirimleri için en çok olabilirlik yöntemi kullanılır. Olabilirlik fonksiyonu yaklaşık olarak (II.1) nolu denklemde verildiği gibi olmalıdır. Fakat tek değişiklik olarak $\pi(x)$ denklem (II.13) de tanımlandığı gibidir. Log olabilirlik fonksiyonunun k+1 tane katsayıya göre türevi alınarak k+1 tane olabilirlik eşitlikleri, j=1,2,...,k olmak üzere,

$$\sum_{i=1}^{n}\left[y_i - \pi(x_i)\right] = 0 \quad \text{(II.15)}$$

ve

$$\sum_{i=1}^{n} x_i \left[y_i - \pi(x_i)\right] = 0 \quad \text{(II.16)}$$

biçiminde elde edilir [11].

Kestirilen katsayıların varyans ve kovaryansları, log olabilirlik fonksiyonunun ikinci dereceden kısmî türevler matrisinden elde edilir. Bu kısmi türevler, j,u=0,1,2,...,k, $\pi_i = \pi(x_i)$ olmak üzere aşağıdaki biçimdedir:

$$\frac{\partial L(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 \pi_i (1-\pi_i) \qquad \text{(II.17)}$$

ve

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_u} = -\sum_{i=1}^{n} x_{ij} x_{iu} \pi_i (1-\pi_i) \qquad \text{(II.18)}$$

Eşitlik (II.17) ve (II.18)'de verilen terimlerin negatiflerini kapsayan (k+1)x(k+1) boyutlu matris **I(β)** ile gösterilir ve "bilgi (information) matrisi" olarak adlandırılır. Kestirilen katsayıların varyans ve kovaryansları bilgi matrisin tersinden elde edilir. **Var(β)=I^{-1}(β).** Çok özel durumların dışında bu matrisin elemanlarını açık bir şekilde yazmak mümkün değildir. Matrisin j. köşegen elemanı Var(β_j)'dir; köşegen dışındaki matris elemanları ise Cov(β_j.β_u) ile gösterilir. Varyans ve kovaryansların kestirimlerini Var(**β**)'yı $\hat{\boldsymbol{\beta}}$ kullanarak buluruz. Kestirilmiş katsayıların kestirilmiş standart hataları , (j=0,1,2,...,k) için,

$$\widehat{\mathrm{SE}}(\hat{\beta}_j) = \left[\widehat{\mathrm{Var}}(\hat{\beta}_j)\right]^{1/2}$$

ile verilir [8].

Kestirilen modelin incelenmesinde ve uyumun değerlendirilmesinde faydalı olacak bilgi matrisi $\hat{\mathbf{I}}(\hat{\boldsymbol{\beta}})=\mathbf{X'VX}$ şeklindedir. Burada **X**, n_x(k+1) boyutundaki bir matris olup her bir birey ve denek için verileri kapsar:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}$$

V ise n x n boyutlu ve köşegen elemanları $\hat{\pi}_i(1-\hat{\pi}_i)$, olan (i=1,2,...,n) bir matristir [8 , 11] :

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

Öte yandan anlamlılık testleri için, tek bağımsız değişkenli durumda verilen olabilirlik oran testinin çok değişkenli duruma genelleştirilmiş biçimi kullanılır. Olabilirlik oran testi (II.5) de verildiği gibi G istatistiğine bağlıdır. Aralarındaki tek fark $\hat{\boldsymbol{\beta}}$ vektöründe k+1 parametreyi kapsayan modelde $\hat{\pi}$ değeridir. Modelde birlikte değişenler için k tane "eğim" katsayısının sıfıra eşit olması hipotezi (null hipotezi) altında, G istatistiği p serbestlik derecesiyle (χ^2) ki-kare dağılımı gösterir. H_0 hipotezinin reddedilmesi doğrusal regresyondaki yoruma benzer yoruma sahiptir: En az bir veya tüm k katsayılarının sıfırdan farklı olması yorumudur.

Genel olarak, test edilecek hipotez,

H_0:$\beta_q=\beta_{q+1}=...=\beta_{k-1}=0$

H_1: En az bir β_p değeri sıfırdan farklıdır.

iken tüm parametreler üzerinde bulunacak olabilirlik fonksiyonu L(F); hipoteze göre kısaltılmış model için ise olabilirlik fonksiyonu L(R) olsun.

Test istatistiği,

$$\chi^2 = 2ln\left[\frac{L(R)}{L(F)}\right] = -2\left[lnL(R) - lnL(F)\right]$$

olur. Örnek hacmi yeterince büyük olduğunda H_0 hipotezi doğru ise χ^2 istatistiği yaklaşık olarak $\chi^2_{(1-\alpha;k-q)}$ şeklinde dağılım gösterir. Serbestlik derecesi, V=(n-q)-(n-k)=k-q şeklindedir. Böylece $\chi^2 \leq \chi^2_{(1-\alpha;\ k-q)}$ olduğunda H_0 kabul edilirken $\chi^2 > \chi^2_{(1-\alpha;\ k-q)}$ olduğunda H_1 kabul edilir [3].

Katsayıların hepsinin veya bir kısmının sıfırdan farklı olduğu sonucuna varmadan önce, değişkenler tek tek Wald test istatistiği ile test edilebilir. Wald testi için kullanılan eşitlik, $W_j = \hat{\beta}_j / \widehat{SE}(\hat{\beta}_j)$'dir. "Bir katsayının $(\hat{\beta}_j)$ sıfıra eşit olması" hipotezi altında Wald istatistiği standart normal dağılım gösterir.

Öte yandan, Score testinin çok değişkenli durumlardaki ifadesi, L($\boldsymbol{\beta}$)'nin $\boldsymbol{\beta}$'ya göre k tane türevinin koşullu dağılımı dikkate alınarak hesaplanır. Bu testin hesaplanması Wald testiyle aynı zorluklara sahiptir ve Score testinin de olabilirlik oran testine karşı bir üstünlüğü yoktur [8].

II.4. LOJİSTİK REGRESYON MODELİNDE KATSAYILARIN YORUMLANMASI

Önceki bölümlerde model kurma ve modeldeki değişkenlerin önem testi üzerinde durulmuştu. Burada ise katsayıların yorumlanması incelenecektir. Lojistik regresyonda model uygunluğunun değerlendirilmesi ile ilgili yöntemler doğası gereği daha karışıktır.

II.4.1 İki Sonuçlu (Dichotomous) Bağımsız Değişken

x bağımsız değişkeni iki düzeyli olsun ve 0 , 1 biçiminde kodlandığını varsayalım. x=1 ve x=0 için lojitteki fark $g(1)$-$g(0)$ =[β_0+β_1]-[β_0]=β_1 şeklindedir. Bu durumda konuyu açıklamak için "Odds Ratio" larına gerek duyulur.

Lojistik regresyonda x bağımsız değişkeni 0 ve 1 ile kodlandığında mümkün olan olasılıklar $\pi(x)$ ve $1-\pi(x)$ değerleridir. Bu durum Tablo II.1'de gösterilmiştir. x=1 iken riskin bireyler arasındaki olma olasılığı $\pi(1)/\left[1-\pi(1)\right]$ olarak tanımlanır. Benzer şekilde x=0 iken sonucun bireyler arasında olma olasılığı da $\pi(0)/\left[1-\pi(0)\right]$olarak tanımlanır. Odds Oranı (Odds Ratio) OR ile gösterilir ve aşağıdaki denklem ile verilir:

$$OR = \frac{\pi(1)/\left[1-\pi(1)\right]}{\pi(0)/\left[1-\pi(0)\right]} \qquad \text{(II.19)}$$

Tablo II.1 Bağımsız Değişkenin İkili Olması Durumunda Lojistik Regresyon Modelinin Değerleri

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

Tablo II.1'deki değerler (II.19) denkleminde yerine konulursa odds oranı aşağıdaki gibi bulunur:

$$OR = \frac{\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right) \Big/ \left(\frac{1}{1+e^{\beta_0+\beta_1}}\right)}{\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right) \Big/ \left(\frac{1}{1+e^{\beta_0}}\right)} \tag{II.20}$$

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

$$= e^{(\beta_0+\beta_1)-\beta_0}$$

$$= e^{\beta_1}$$

Dolayısıyla iki düzeyli bağımsız değişkenin lojistik regresyonu için odds oranı $OR = e^{\beta_1}$ dir. Odds değerlerinin logaritması lojit olarak adlandırılır. Bu değerlere ilişkin lojit değerler,

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

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

olmak üzere log-odds (odds) logaritması,

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

$$= ln\left(e^{\beta_1}\right)$$

$$= \beta_1$$

şeklinde ifade edilir; lojit fark (logit difference) olarak da adlandırılır. Yani *x* değişkeninin iki farklı değeri için bulunan log-odds oranlarının farkı, bu iki koşulun karşılaştırılmasında kullanılacak log-odds oranını verir. Yukarıdaki örnekte log-odds oranı $ln(OR) = ln\left(e^{\beta_1}\right) = \beta_1$ olup, regresyon modelindeki eğim katsayısını vermektedir. Bir başka yorum ise, $g(1)$-$g(0)$ olmasıyla $\beta_1 = g(x+1)-g(x)$ şeklinde yazılarak bağımsız değişkendeki bir birimlik değişmeye karşılık lojitteki değişmeyi göstermesidir [11].

Odds Oranı geniş kullanıma sahip bir ölçüm birimidir. Epidemiolojik araştırmalarda güçlü bir analitik metot olduğu kanıtlanan lojistik regresyonun seçilmesinin temel nedeni katsayılarının yorumlanabilir olmasıdır. Lojistik regresyonda katsayıların yerine odds oranlarının yorumlanması oldukça büyük önem

taşır. x=1 olan bireyler arasında sonuç değişkeninin gerçekleşme olasılığı, x=0 olan bireylere göre ne kadar çok veya az olduğu açısından özellikle epidemiolojide yaygın olarak kullanılan bir ilişki ölçüsüdür. Örneğin y yanıt değişkeni bireyin akciğer kanseri olup olmadığını, x değişkeni de bireyin sigara kullanım durumunu göstersin. Eğer $\widehat{\mathrm{OR}} = 2$ ise, akciğer kanserinin sigara içen kişilerde içmeyenlere göre iki kat daha sık görüldüğünü belirtir.

Odds oranı yardımı ile yorumlama birçok durumda göreli risk (relative risk) adı verilen bir niceliğe dayanır. Bu parametre $\pi(1)/\pi(0)$ oranına eşittir. Eşitlik (II.18)den $[1-\pi(0)]/[1-\pi(1)] \approx 1$ ise odds oranı göreli riske yaklaşır. Bu yaklaşım ancak $\pi(x)$, x=0 ve x=1 için yeterince küçük ise geçerli olur [8].

Odds oranı, OR, lojistik regresyonda önemli bir parametredir. Teorik olarak örneklem genişliği yeterince büyük olduğu zaman $\widehat{\mathrm{OR}}$'nin dağılımı normaldir. Ancak örneklem büyüklüğü çoğu çalışmada sağlanamamaktadır. Bundan dolayı çıkarsamalar genellikle $ln(\widehat{OR}) = \hat{\beta}_1$'in örneklem dağılımına dayanır. Odds oranının 100(1-α)% güven aralığı kestirimi, ilk olarak β_1 katsayısı üzerinden elde edilir:

$$\exp\left[\hat{\beta}_1 \pm Z_{1-\alpha/2}\widehat{SE}(\hat{\beta}_1)\right]$$

Güven aralığının sınırları 1'den büyük çıktığı zaman dağılım sağa yatık olur ve aralıkta 1 değerinin olmaması durumunda odds oranının istatistiksel olarak önemli olduğuna karar verilir. Bu durum sadece bağımsız değişkenin 0 ve 1 değerleriyle kodlanmasında geçerlidir [8, 11].

İlişki ölçüsü olarak odds oranının öneminden dolayı, nokta ve aralık kestirimleri genellikle lojistik regresyon analizi sonucunda tablolarda bulunur.

Bağımsız değişkenin diğer kodlama türlerinde ise lojit farkın değeri yeni kodlar üzerinden yeniden hesaplanır; sonra da e'nin üssü şeklinde yazılır. İki farklı düzeydeki x=a'ya karşı x=b için herhangi bir bağımsız değişkenin odds oranının logaritması, bu iki değer için hesaplanan lojitlerin farkına eşittir:

$$\begin{aligned} ln\left[\widehat{OR}(a.b)\right] &= \hat{g}(x=a) - \hat{g}(x=b) \\ &= (\hat{\beta}_0 + \hat{\beta}_1 a) - (\hat{\beta}_0 + \hat{\beta}_1 b) \\ &= \hat{\beta}_1 (a-b) \end{aligned} \qquad \text{(II.21)}$$

Kestirilen odds oranı, aşağıdaki gibidir:

$$\widehat{OR}(a.b) = \exp\left[\hat{\beta}_1(a-b)\right]. \tag{II.22}$$

Bu ifade sadece (a-b)=1 olduğu zaman $\exp(\hat{\beta}_1)$'e eşit olur. Eşitlik (II.21) ve (II.22)'deki $\widehat{\mathrm{OR}}(\mathrm{a.b})$ Odds notasyonunu gösterelim:

$$\widehat{OR}(a.b) = \frac{\hat{\pi}(x=a)/\left[1-\hat{\pi}(x=a)\right]}{\hat{\pi}(x=b)/\left[1-\hat{\pi}(x=b)\right]} \tag{II.23}$$

a ve b, marjinal yöntem olarak da bilinen, eğer a=1, b=-1 (örneğin; kadın:-1, erkek:+1) biçiminde kodlanmış ise, (II.21) eşitliği,

$$\begin{aligned} ln\left[\widehat{OR}(kadın, erkek)\right] &= \hat{g}(D=1) - \hat{g}(D=-1) \\ &= \left[\beta_0 + \beta_1(D=1)\right] - \left[\beta_0 + \beta_1(D=-1)\right] \\ &= 2\hat{\beta}_1 \end{aligned}$$

ve kestirilen odds oranı $\widehat{OR} = \exp(2\hat{\beta}_1)$ olur.

Kodlama metodu, aynı zamanda odds oranının güven aralığının alt ve üst sınırlarının hesaplanmasını da etkiler. Genel olarak denklem (II.23) de verilen odds oranı için güven aralığının alt ve üst sınırları,

$$\exp\left[\hat{\beta}_1(a-b) \pm Z_{1-\alpha/2}\left|a-b\right|\widehat{SE}(\hat{\beta}_1)\right] \tag{II.24}$$

şeklindedir; $\left|\mathrm{a}-\mathrm{b}\right|$, $(\mathrm{a}-\mathrm{b})$ değerinin mutlak değeridir. Genelde tüm iki düzeyli değişkenlerin 0 veya 1 diye kodlanması ve bu değişkenlerin aralık ölçekli değişken olarak varsayılması önerilir [8, 11].

II.4.2 Çok Sonuçlu (Polychotomous) Bağımsız Değişken

Nominal bir bağımsız değişkenin ikiden fazla düzey içerdiği durumda, bu değişkenin kategorilerini gösteren göstermelik/tasarım değişkenlerinin oluşturulması gerekir.

Göstermelik değişkenlerin seçimi, bir düzey referans seçilerek yapılabilir. Referans grup olarak seçilen düzey için 0 ve göstermelik değişkenler 1 değerini alır. İki düzeyin birbirine göre karşılaştırılması, örneğin,

$$ln\left[\widehat{OR}(siyah,beyaz)\right]=\left[\hat{\beta}_0+\hat{\beta}_1(D_1=1)+\hat{\beta}_2(D_2=0)+\hat{\beta}_3(D_3=0)\right]$$
$$-\left[\hat{\beta}_0+\hat{\beta}_1(D_1=0)+\hat{\beta}_2(D_2=0)+\hat{\beta}_3(D_3=0)\right]$$
$$=\hat{\beta}_1$$

biçiminde uygulanır [8].

Odds oranları için güven aralıkları, iki düzeyli bağımsız değişkenlerde kullanılan yaklaşımdan aynı şekilde elde edilir. İlk olarak Log-odds (lojistik regresyon katsayısı) için güven limitleri bulunur ve daha sonra bu limitleri ve üssü alınarak odds oranı için güven limitleri elde edilir. Genel olarak, katsayı için %100(1-α) güven limitleri $\hat{\beta}_j \pm Z_{1-\alpha/2}\widehat{SE}(\hat{\beta}_j)$ dir. Odds oranı için güven aralığı bu sınırların e-üssü alınarak,

$$\exp\left[\hat{\beta}_j \pm Z_{1-\alpha/2}\widehat{SE}(\hat{\beta}_j)\right] \qquad \text{(II.25)}$$

elde edilir.

Tasarım değişkenlerinin kodlanmasının bir başka yolu ise ortalamadan sapma metodudur. Bu kodlama yöntemi genel ortalamadan grup ortalamasının sapmasının etkisini açıklar. Lojistik regresyonda, "grup ortalaması" grubun lojitidir ve "genel ortalama" ise tüm grupların ortalama lojitidir. Bu yönteme göre, tasarım değişkeninin tüm değerleri -1 alınıp, geri kalan diğer değişkenler için 0 ve 1 kodlaması kullanılır.

Ortalamadan sapma kodlaması metoduyla kestirilen katsayıların yorumlanması, referans grup metodu kadar kolay ve açık değildir. Kestirilen katsayıların üssü, belirli bir grup için odds değerinin, oddsların geometrik ortalamasına oranını verir. Fakat buradan çıkan sonuç gerçeği yansıtmaz. Çünkü pay ve paydada iki farklı kategori için oddslar mevcuttur. Kestirilen katsayının üssü ortalama odds değerine göre ifade edilmektedir.

II.4.3 Sürekli Bağımsız Değişken

Lojistik regresyon modeli bir sürekli bağımsız değişkeni içerdiği zaman, kestirilen katsayıların yorumlanması değişkenin modele nasıl girdiğine ve değişkenin birimine bağlıdır. Doğrusallık varsayımı altında, lojit, $g(x)=\beta_0+\beta_1 x$ 'dir. β_1 eğim katsayısı, x'deki "1" birimlik artışın log-odds değerinde meydana getirdiği değişimi verir. Herhangi bir x değeri için $\beta_1=g(x+1)-g(x)$ dir. "1" değeri çoğu zaman biyolojik olarak pek ilginç değildir. Örneğin, yaş değişkenindeki "1" birim artış veya sistolik

kan basıncındaki 1 mm Hg artış önemli sayılmayacak kadar küçük bir değişimdir. Yaştaki 10 yıllık veya 10 mm Hg değişimi daha yararlı kabul edilebilir. Diğer yandan x'in tanım aralığı 0'dan 1'e doğru ise, "1" birimlik değişim çok büyük olacaktır; 0,01 birimlik artış da daha gerçekçi olacaktır. Bu nedenden dolayı sürekli ölçekli kovaryantlar hakkında faydalı bir yorum sağlamak için kovaryanttaki "*c* " değişimi için nokta ve aralık kestirimine bakmak gerekir.

x'deki *c* birimlik bir değişim için log-odds oranı $g(x+c)-g(x)=c\beta_1$ lojit farkından elde edilir ve karşılık gelen odds oranı $OR(c)=OR(x+c,x)=\exp(c\beta_1)$ olarak bulunur. β_1'in en çok olabilirlik kestirimi $\hat{\beta}_1$, güven aralığı kestiriminde aşağıdaki gibi kullanılır: OR(*c*)' nin %100 (1-α) güven aralığı kestirimi,

$$\exp\left[c\hat{\beta}_1 \pm Z_{1-\alpha/2}c\,\widehat{SE}(\hat{\beta}_1)\right]$$

şeklindedir.

Güven aralığının başlangıç ve bitiş noktalarının *c*' nin seçimine bağlı olmasından dolayı *c*' nin belirli bir değeri tüm tablo ve hesaplamalarda açıkça belirtilmelidir. *c*' nin daha rastgele, keyfi seçimi bazı durumlarda zorluklar çıkarabilir.

Lojitin kovaryantta doğrusal olmadığına inanılıyorsa, göstermelik değişkenlerin kullanılması ve gruplandırılması düşünülmelidir. Alternatif olarak, yüksek dereceli terimler (örn: x^2, x^3, ...) veya kovaryantta doğrusal olmayan ölçekleme (örn. log(*x*)) düşünülmelidir. Dolayısıyla, sürekli kovaryantları modellemenin önemli noktası lojitteki ölçeklendirmedir [8].

Özet olarak, sürekli bir değişkenin kestirilmiş katsayısının yorumlanması, nominal ölçekli değişkenlerinkine benzer. En önemli fark ise sürekli değişken için anlamlı bir değişimin tanımlanmasıdır.

II.4.4 Etkileşim Varlığında Odds Oranlarının Kestirimi

Zaman zaman iki ya da daha çok değişken söz konusu olduğunda etkileşim terimleri de modele girebilir. Risk faktörü ve başka bir değişken arasında etkileşim olduğunda risk faktörünün odds oranı kestirimi, onunla etkileşime giren değişkenin değerine bağlıdır. Bu durumda önceki bölümlerde verilen odds oranlarını kestirmek için kullanılan formülde bir değişiklik yaparak etkileşim içinde olan değişkenler arasındaki lojit farkı da hesaba katılır [8].

Bir modelde iki bağımsız değişken ve bunların etkileşimi yer alsın. F risk faktörü, *x* kovaryant ve onların etkileşimi de F.X şeklinde ifade edilsin. F=*f* ve X=*x* için bu modelin lojiti,

$$g(f.x) = \beta_0 + \beta_1 f + \beta_2 x + \beta_3 fx \qquad \text{(II.26)}$$

şeklinde olur. X=*x* sabit tutulduğunda F=*f* 'e karşı F=f_1 düzeyleri için log-odds oranı,

$$g(f_1.x) = \beta_0 + \beta_1 f_1 + \beta_2 x + \beta_3 f_1 x$$

ve

$$g(f_0.x) = \beta_0 + \beta_1 f_0 + \beta_2 x + \beta_3 f_0 x$$

ifadelerinin farkının lojitinin alınmasıyla,

$$ln\left[OR\left(F=f_1, F=f_0, X=x\right)\right] = g(f_1,x) - g(f_0,x) \qquad \text{(II.27)}$$

$$= \left(\beta_0 + \beta_1 f_1 + \beta_2 x + \beta_3 f_1 x\right) - \left(\beta_0 + \beta_1 f_0 + \beta_2 x + \beta_3 f_0 x\right)$$

$$= \beta_1(f_1 - f_0) + x\beta_3(f_1 - f_0)$$

şeklinde elde edilir. Buradaki parametreler yerine kestirilmiş değerlerini kullanarak kestirilmiş log-odds değeri bulunmuş olur. Bu eşitliğin varyans kestirimi,

$$\widehat{Var}\left\{ln\left[\widehat{OR}(F=f_1, F=f_0, X=x\right]\right\} = \left(f_1 - f_0\right)^2 .\widehat{Var}(\hat{\beta}_1) + \left[x(f_1 - f_0)\right]^2 .\widehat{Var}(\hat{\beta}_3)$$

$$+2x\left(f_1 - f_0\right)^2 .\widehat{Cov}(\hat{\beta}_1, \hat{\beta}_3) \qquad \text{(II.28)}$$

biçiminde verilir [8].

Lojistik regresyon programlarının birçoğu modeldeki kestirilen parametrelerin varyans ve kovaryansının kestirimini verir. Kestirimler elde edildikten sonra, eşitlik (II.28) de bu değerler yerine konularak kestirilen log-odds oranının varyansının bir kestirimi bulunur. Log-odds oranı için %100(1-α) güven aralığı kestiriminin sınır (başlangıç ve bitiş) noktaları,

$$\exp\left(\left[\hat{\beta}_1(f_1 - f_0)\right] + x\hat{\beta}_3(f_1 - f_0)\right) \pm Z_{1-\alpha/2}\,\widehat{SE}\left\{ln\left[\widehat{OR}(F=f_1, F=f_0, X=x\right]\right\} \qquad \text{(II.29)}$$

olarak elde edilir.

F iki düzeyli bir risk faktörü ise bu ifade, varyansı ve log-odds kestiricileri daha basit şekil alır. Eğer f_1=1, f_0=0 alırsak log-odds oranının kestirimi,

$$ln\left[\widehat{OR}(F=1, F=0, X=x)\right] = \hat{\beta}_1 + \hat{\beta}_3 x$$

olur. Varyans kestirimi

$$\widehat{Var}\left\{ln\left[\widehat{OR}(F=f_1,F=f_0,X=x\right]\right\}=\widehat{Var}(\hat{\beta}_1)+x^2\widehat{Var}(\hat{\beta}_3)+2x\widehat{Cov}(\hat{\beta}_1,\hat{\beta}_3)$$

şeklindedir. Odds oranı için kestirilen güven aralığının sınır noktaları

$$\exp\left(\left(\hat{\beta}_1+\hat{\beta}_3 x\right)\pm Z_{1-\alpha/2}\widehat{SE}\left\{ln\left[\widehat{OR}\left(F=1,F=0,X=x\right)\right]\right\}\right)$$

şeklinde elde edilir [8].

II.5 MODEL UYUMUNUN BELİRLENMESİ

Modele gerekli tüm değişkenler alındıktan sonra kestirilen lojistik regresyon modelinin sonuç değişkenini tanımlamakta ne kadar etkili olduğu, uyum iyiliği (goodness-of-fit) testiyle anlaşılır.

Bağımlı değişkeninin gözlenen değerleri **y** vektörü şeklinde gösterilsin ve $\mathbf{y}'=(y_1,y_2,...,y_n)$ olsun. Model tarafından kestirilen değerler $\hat{\mathbf{y}}$ ile gösterilsin ve $\hat{\mathbf{y}}'=(\hat{y}_1,\hat{y}_2,...,\hat{y}_n)$ olsun. Eğer **y** ve $\hat{\mathbf{y}}$ arasındaki uzaklık, yani artık(residual), özet ölçüleri küçükse ve her bir $(y_i,\hat{y}_i)$, i=1,2,3...,n ikilisinin bu özet ölçülere katkısı sistematik değil ve modelin hata yapısına göre küçükse modelin uyumlu olduğuna karar verilir. Böylece, uygun bir modele tamamen karar vermek için hem **y** ve $\hat{\mathbf{y}}$ arasındaki uzaklığın özet ölçüsünün hem de bu ölçülerin her bir parçasının tek tek incelenmesi gereklidir.

Uyum iyiliği için şu aşamalar uygulanır: (a) tümel uyum ölçülerinin hesaplanması ve değerlendirilmesi, (b) genellikle grafiksel olarak özet istatistik parçalarının ayrı ayrı incelenmesi ve (c) **y** ve $\hat{\mathbf{y}}$ parçaları arasındaki uzaklık veya farkın diğer ölçümlerinin incelenmesi.

Öte yandan, uyum iyiliği kovaryantın hepsiyle değil, modeldeki kovaryantlarla belirlenir. Örneğin $\mathbf{x}'=(x_1,x_2,...,x_k)$ bağımsız değişken vektörü ve J gözlenen *x* 'in farklı değerlerinin sayısını göstersin. m_j denek sayısını göstersin, j=1,2,..,,J.

- Eğer analize dahil edilen denekler arasında aynı kovaryant değerine sahip denekler varsa, kovaryant sayısı (J) < toplam denek sayısı (n) olur, (m_j değeri büyük olma eğiliminde).

- Model geliştirme aşamasında, bağımsız değişkenlerin değerleri her bir denek için farklı ölçüm değerine sahipse J=n dir [8].

Özellikle modelde sürekli değişkenler yer alıyorsa J≈n olması beklenilir. Uyum iyiliği testinde kullanılan farklı istatistikler vardır. Bunlar aşağıda açıklanmıştır ve bu istatistiklerde J=n olduğu varsayılmıştır.

II.5.1 Ki-Kare İstatistiği (χ^2_{B0})

Modelde sadece sabit terim varken sözkonusu hatayı gösterir. Diğer bir anlatımla (χ^2_{B0}) istatistiği, modelde sadece sabit terim olduğunda -2LogL istatistiğini vermektedir. Yani ilk ki-kare istatistiği modeldeki tüm β katsayılarının sıfır olduğu hipotezini kabul eden -2LogL istatistiğidir.

II.5.2 Pearson Ki-Kare İstatistiği ve Sapma Ölçütü

Lojistik regresyonda gözlenen ve kestirilen değerler arasındaki farkın (doğrusal regresyondaki $y-\hat{y}$ artık fonksiyonu gibi) birçok ölçüsü vardır. Lojistik regresyonda kestirilen değerler her bir kovaryantın deseni için hesaplanır ve o kovaryant deseni için kestirilen olasılığa bağlıdır; kestirilen değer $\hat{y}_j$ ile gösterilir:

$$\hat{y}_j = m_j \hat{\pi}_j = m_j \frac{e^{\hat{g}(x_j)}}{1+e^{\hat{g}(x_j)}}$$

Buradaki $\hat{g}(x_j)$ kestirilen lojittir. m_j kovaryant değerleri birbirinden farklı olan denek sayısıdır.

Gözlenen ve kestirilen değerler arasındaki farkın iki ayrı ölçüsünü inceleyelim. Bunlar Pearson artığı ve deviance artığıdır. Kovaryant deseni için Pearson artığı

$$r(y_j, \hat{\pi}_j) = \frac{\left(y_j - m_j \hat{\pi}_j\right)}{\sqrt{m_j \hat{\pi}_j \left(1-\hat{\pi}_j\right)}} \qquad \text{(II.30)}$$

dir. Bu artıklar üzerinden hesaplanan istatistik Pearson ki-kare istatistiği,

$$\chi^2 = \sum_{j=1}^{J} r(y_j, \hat{\pi}_j) \qquad \text{(II.31)}$$

ile verilir. Sapma artığı aşağıdaki gibi tanımlanır:

$$d\left(y_j, \hat{\pi}_j\right) = \pm \left\{ 2 \left[y_j . \ln\left(\frac{y_j}{m_j \hat{\pi}_j}\right) + \left(m_j - y_j\right) . \ln\left(\frac{(m_j - y_j)}{m_j (1-\pi_j)}\right) \right] \right\}^{1/2} \qquad \text{(II.32)}$$

Burada denklemin önündeki (±) işareti $(y_j - m_j\hat{\pi}_j)$ ifadesinin işareti ile aynıdır. m_j=1 ve y_j=0 kovaryant deseni için deviance artığı aşağıdaki şekle dönüşür:

$$d\left(y_j, \hat{\pi}_j\right) = -\sqrt{2m_j \left|ln(1-\hat{\pi}_j)\right|}$$

Başka bir örnek olarak m_j=1 ve y_j=1 için sapma artığı,

$$d\left(y_j, \hat{\pi}_j\right) = \sqrt{2\left|ln(\hat{\pi}_j)\right|}$$

olur. Sapma artıklarına dayanan özet istatistiği,

$$D = \sum_{j=1}^{J} d(y_j, \pi_j)^2 \qquad \text{(II.33)}$$

olarak yazılır [8].

Kurulan modelin tüm yönlerden doğru olduğu varsayımı altında χ^2 ve D istatistiklerinin dağılımının J-(k+1) serbestlik derecesiyle ki-kare dağılımına sahip olduğu varsayılır. Aynı zamanda D istatistiği (k+1) parametreli kestirim modeline karşılık J parametreli doymuş modelin olabilirlik oran testidir. Benzer durum H_0 hipotezi geçerli iken χ^2 istatistiğinin gösterdiği dağılım için de geçerlidir:

$$D = -2.ln\left[\frac{\text{Uyarlanmış modelin olabilirliği}}{\text{Doymuş modelin olabilirliği}}\right]$$

Diğer bir deyişle olabilirlik oran testi,

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

ile verilir. j=n alınmasıyla bu istatistik D, sapma olarak da adlandırılır [12, 14].

Sapma, genelde analize bağımsız değişken ilave edildiğinde modelin hatasını gösterir. Bu nedenle -2LogL istatistiği bağımlı değişkendeki açıklanmayan varyansın anlamlılığını gösterir. Bu istatistiğin anlamlı olmaması lojistik regresyon analizinde istenen durumu göstermektedir. Log olabilirlik değeri 0 ve 1 arasında değerler almaktadır. Bu oran bağımlı değişkenin bağımsız değişkenler tarafından tahmin edilme olasılığını gösterir. -2LogL istatistiği yaklaşık olarak ki-kare dağılımına uyduğundan lojistik regresyon analizindeki bu istatistik regresyon analizindeki artık kareler toplamına benzer. Olabilirlik oranı 1 ise -2LogL sıfıra eşit olur. Model, doymuş modelle uyumlu tam olarak temsil edilmesi durumunda bu sonuç gerçekleşir. Özetle -2LogL istatistiği ne kadar küçük değere sahipse daha iyi bir model oluşur [8, 14].

II.5.3 Hosmer-Lemeshow (G) İstatistiği

Hosmer ve Lemeshow, kestirilen olasılık değerlerinin gruplandırılmasını önermişlerdir. J=n ise veri matrisindeki n tane sütun. n tane kestirilen olasılık değerine karşılık gelir. Burada gruplama yapmadaki amaç daha düzgün bir beklenen değer oluşturulmasıyla mevcut dağılımı ki-kare dağılımına yaklaştırarak anlamlı, yorumlanabilir model elde etmektir. Gruplama iki farklı tipte yapılabilir:

(a) kestirilen olasılıkların yüzdesi dikkate alınarak

(b) kestirilen olasılıkların sabit değerleri dikkate alınarak.

İlk yöntemde gözlenen ve kestirilmiş beklenen frekansları karşılaştıran $\hat{C}$ istatistiğini inceleyeceğiz. Burada 10'lu risk grubu kullanılır. Yani tüm gözlemler 10 gruba ayrılır. Bu yöntem yeterli sayıda gözlem olduğunda geçerlidir. Olumsuz yönü ise gerçek değerlerin göz ardı edilmesidir. $\hat{C} \sim \chi^2_{(g-2)}$,g grup sayısıdır.

İkinci yöntemde sabit kesim noktaları üzerine bir gruplandırma yapılır. Örneğin kestirilen olasılıklara göre sabit gruplar oluşturularak deneklerin ilgili grupta yer alması sağlanır.

Gruplama yoluyla verileri azaltma işleminde, gruplardaki veri sayısı azaldığından dolayı uyumdan sapmalar görülebilir. Hosmer-Lemeshow testi yorumlama ve geniş veri kümesini rahat çözmek amaçlı en yaygın tercih edilen bir uyum testidir. Bu istatistik SPSS'de "Hosmer-Lemeshow G" olarak bilinmesinin yanı sıra Model ki-kare istatistiği olarak adlandırılır. Bağımsız değişkenlerden hiçbirinin bağımlı üstünlük oranıyla anlamlı doğrusal bir ilişki göstermediğini ileri süren sıfır hipotezini test etmektedir. Diğer bir anlatımla bu istatistik, sabit terimin dışındaki tüm lojit katsayılarının sıfıra eşit olup olmadığını sınamaktadır.

Model ki-kare istatistiği bir olabilirlik oran testidir ve bu yüzden modelde bağımsız değişkenin olmadığı -2LogL istatistiği ile modelde bağımsız değişkenlerin olduğu -2LogL istatistiği arasındaki fark olarak hesaplanmaktadır. İncelenen modelin parametre sayısı ile yalnız sabit terimli modelin parametreleri arasındaki farka eşit bir serbestlik derecesi ile ki-kare dağılımına uymaktadır. Lojistik regresyon analizinde model ki-kare değerinin anlamlı olması arzu edilen durumu göstermektedir. Model ki-kare testi regresyon analizindeki F testine benzer [8, 10].

II.5.4 Blok Ki-Kare İstatistiği

Bir blok değişkeninin modele dahil edilmesiyle model ki-kare istatistiğine meydana gelen değişmeyi gösterir. Bu istatistik adımsal lojistik regresyon analizinde "Step ki-kare" adıyla hesaplanır. Her adımda modele tek bir değişken ilave edilip çıkartılıyorsa, doğal olarak blok ve step ki-kare istatistikleri eşit olmaktadır. Kategorik bağımsız değişken modele dahil edildiğinde blok ki-kare istatistiği ile test edilmektedir. Bu durumda kategorik değişken ile ilgili tüm göstermelik değişkenler blok olarak modele dahil edilmektedir [8, 10].

II.5.5 Sınıflandırma Tablosu

Kurulan bir lojistik modelin sonuçlarını özetlemek için sezgiye dayanan en iyi yol sınıflandırma tablosu oluşturmaktır. Bu tablo, sonuç değeri y'nin düzeyleri ile kestirilen lojistik olasılıklar tarafından türetilen ikili bir değişkenin çapraz sınıflandırılmasıyla elde edilir.

Türetilen bu ikili bağımsız değişkeni elde etmek için c kesim noktası belirlenir ve kestirilen her bir olasılık değeri c ile karşılaştırılır. Eğer kestirilen olasılık c değerini geçerse, türetilen ikili değişken 1'e eşit olur, diğer durumlarda 0'a eşittir. c' nin en yaygın kullanılan değeri 0.5'dir.

Bu yaklaşımda kestirilen olasılıklar grup üyeliğini tahmin etmek için kullanılır:

Y=0 grubunda $P(Y=1)=Q_1$ ve $X\sim N(0,1)$

Y=1 grubunda $X\sim N(\mu,1)$ olduğu varsayılırsa lojistik regresyon için eğim katsayısı $\beta_1=\mu$ olur ve sabit değer,

$$\beta_0 = ln\left[\frac{Q_1}{(1-Q_1)}\right] - \frac{\mu^2}{2}$$

ile verilir. Lojistik regresyon modelinin Hatalı Sınıflandırma Olasılığı (PMC),

$$PMC = Q_1\Phi\left\{\frac{1}{\beta_1} ln\left[\frac{(1-Q_1)}{Q_1}\right] - \frac{\beta_1}{2}\right\} + (1-Q_1)\Phi\left\{\frac{1}{\beta_1} ln\left[\frac{Q_1}{(1-Q_1)}\right] - \frac{\beta_1}{2}\right\}$$

dir. Φ, N(0,1) dağılımının kümülatif dağılım fonksiyonudur. Böylece beklenen hata oranı, eğimin büyüklüğünün bir fonksiyonu olduğundan, model uyumuyla ilişkisiz olduğundan doğru veya yanlış sınıflandırma uyum iyiliği için bir kriter olarak dikkate alınmaz. Sınıflandırma tablosu bağımlı değişken gruplarındaki denek,

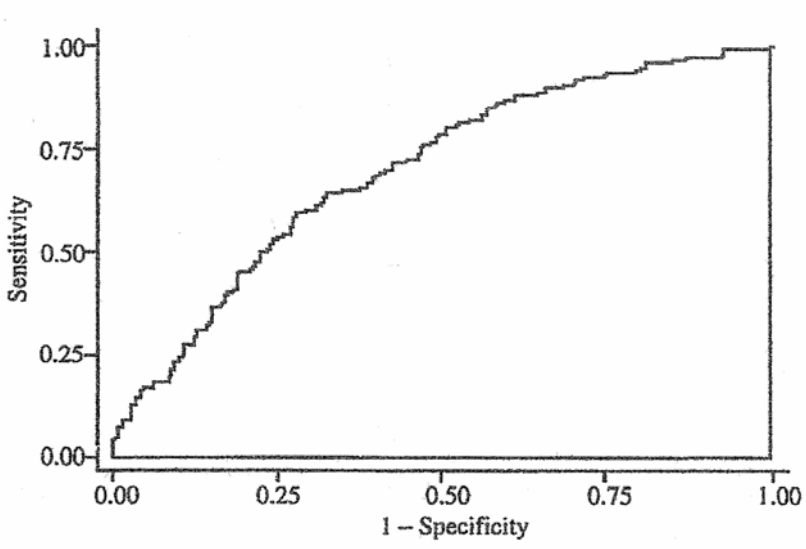
gözlem sayısına duyarlıdır. Grup büyüklüğü arttıkça doğru sınıflandırma olasılığının güvenirliği artar [8].

II.5.6 ROC Eğrisi Altındaki Alan

Doğruluk yüzdeleri, bir test sonucunun doğru sınıflandırılmasındaki tek kesim noktasına dayanır. ROC (Receiver Operating Characteristic) eğrisi altındaki bölge, sınıflandırmada doğruluk tanımından daha fazlasını verir. ROC eğrisi altında kalan alan 0'dan 1'e kadar değişir. Gerçekleşmeyen duruma karşı gerçekleşen değerlerin farkını görmek bu ayrımı yapabilmek için modelin bir yeterlilik ölçüsü olarak kullanılır.

Sınıflandırma tablosunda duyarlılık ve özgüllük $\left(\underset{(\hat{y}=1 \wedge y=0)}{\text{sensitivity}}, \underset{(\hat{y}=0 \wedge y=0)}{\text{specificity}} \right)$ sadece ek bilgi verir, model uyumuna karar vermede doğrudan etkili değildir.

Hedefimiz, sınıflandırmanın amacı için optimal bir kesim noktasını seçmek ise (sensitivity ve specificity) doğruluk yüzdelerinin her ikisini de maksimum yapan bir kesim noktası seçmek olacaktır.



Şekil II.1 ROC Eğrisi Altında Kalan Alanın Grafiksel Gösterimi

Bu Şekil II.1'de ROC eğrisi altındaki alan 0.6989'dur. İstatistiksel paket programlarla ilgili değerler (kesim noktası gibi) girildikten sonra ROC eğri altındaki alana ulaşılır. Genel bir kural olarak: (ROC eğrisi altında kalan alanın doğru sınıflandırma ölçüsüne göre bir olasılık olduğu ve (0,1) arası değer aldığı unutulmamalıdır).

ROC = 0.5 alana sahipse ayrım yapılamaz, yorumlanamaz.

$0.7 \leq ROC < 0.8$ ise kabul edilebilir ayrım olduğu düşünülür.

$0.8 \leq ROC < 0.9$ ise mükemmel ayrımdır.

$ROC \geq 0.9$ ise çok iyi ayrımdır.*

* Pratikte ROC eğrisi altındaki alanın 0.9'dan büyük gözlenmesi olağan dışıdır. Gerçekte önceki konularda belirtildiği gibi tam bir ayrım olduğunda lojistik regresyon modelinin katsayılarını kestirmek imkansızdır, fakat tam ayrıma, sınıflandırmaya yakın bir ihtimal ROC eğrisi altında kalan alan > %90 olduğunda kabul edilebilir [8].

II.6 MODEL DEĞERLENDİRME

İstatistikte geliştirilen bir modelin geçerliliğinin değerlendirilmesi büyük önem taşır. Lojistik regresyon modelinin uygunluğunun değerlendirilmesinde genelde gerçek olasılıklar ile kestirilen olasılıklar arasındaki standart farka bakılır [10]. Gözlemin, yeni modelin hangi açıdan etkili olduğu, değişik hataların dağılımları, ilişki ölçümü gibi göstergeler incelenir. Lojistik regresyonda hesaplanabilen hata türleri:

a) Standart Olmayan Hatalar

Standart olmayan hatalar (e_i), fiili (gerçek) olasılıklar ile kestirilen olasılıklar arasındaki farka eşittir. Modelin hataları lojit ölçekte hesaplanmış ise bu hatalara lojit hatalar adı verilir. Lojit hatalar aşağıdaki formülle hesaplanır:

$$\text{Lojit Hata} = \frac{e_i}{P_i(1-P_i)} \qquad \text{(i. birimin standart olmayan hatası)}$$

b) Standart Hatalar

Standart hatalar, standart olmayan hataların (e_i) kendi standart sapmalarına bölünmesiyle elde edilir. Standart hatalar, aşağıdaki formülle hesaplanır:

$$Z_i = \frac{e_i}{\sqrt{P_i(1-P_i)}} \qquad \text{(i. birimin standart hatası)}$$

Her bir birimin standart hatası χ^2 uygunluk istatistiğinin bir bileşeni olarak görülebilir. Büyük örnekler için standart hatalar 0 ortalama ve 1 standart sapma ile normal dağılıma uyar.

c) Sapma Değerleri

Modelin serbestlik derecesiyle de ilgili olan sapma değeri deviance artığından elde edilir. y=0 riskli karşılaşılmayan durum için sapma değeri,

$$\text{Sapma} = \sqrt{-2.\ln(P_i)}$$

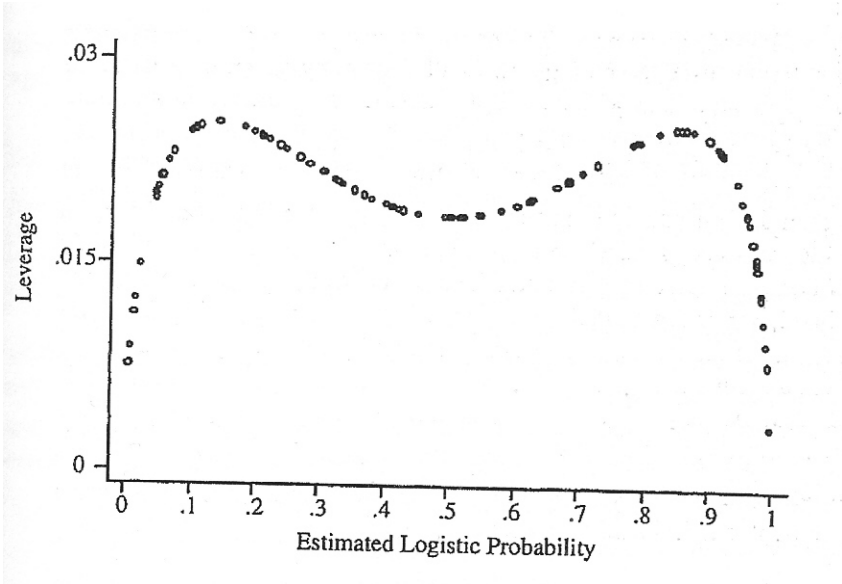
y=1 riskiyle karşılaşıldığında,

$$\text{Sapma} = \sqrt{-2.\ln(1-P_i)}$$

formülünden hesaplanır. Büyük sapma değerleri modelin ilgili verileri iyi temsil etmediğini gösterir. Büyük örnekler için sapma değerleri yaklaşık olarak normal dağılıma uymaktadır.

d) Uzaklık (Leverage) Değerleri

Uzaklık değerleri kestirilen değerler üzerinde büyük etkisi olan birimlerin belirlenmesi amacıyla kullanılır. Uzaklık değerleri 0 (tamamen etkisiz) ve 1 (tamamen etkili) aralığı içinde değerler alır. Uzaklık değerlerinin ortalaması P/n oranına eşittir. P, sabit terim dahil, modelde tahmin edilen parametre sayısını ve n örnek hacmini göstermektedir. Böylece uzaklık değerleri ortalama uzaklık değeri ile karşılaştırılmaktadır. Ayrıca herhangi birbirimin kestirilen olasılığı %10'dan küçük veya %90'dan büyük olması durumunda, söz konusu birim etkili bir birim olsa bile uzaklık değeri küçük hesaplanabilmektedir [10].



Şekil II.2 Leverage Değerine (h) Karşı Kestirilen Olasılık Değerinin ($\hat{\pi}$) Çizimi

Eğer kestirilen olasılık 0.1 ve 0.9 arasında ise, leverage uzaklık olarak düşünülebilecek bir değer verir. Kestirilen olasılık 0.1 ve 0.9 aralığı dışında olduğu zaman, leverage değeri uzaklığı tam anlamıyla ölçemeyebilir.

e) Cook Uzaklığı (Cook's Distance)

Cook uzaklığı değeri herhangi bir birimin model üzerindeki etkisini gösterir. Cook uzaklığı belirli bir birimin modelden çıkartılması durumunda lojistik regresyon katsayılarının ne kadar değişeceğini gösterir. Aşağıdaki formülle hesaplanır:

$$CU_i = Z_i^2 \left(\frac{h_i}{1 - h_i} \right)$$

Formülde Zi standartlaştırılmış hataları ve h_i ise uzaklık (leverage) değerini gösterir. Formülden kolayca görülebileceği gibi Cook uzaklığı hem standart hataya hem de uzaklık değerine bağlıdır.

f) Grafik Yöntemler

Yukarıda bahsedilen istatistiklerden uygun olanlar kullanılarak normal olasılık ve diğer grafikler elde edilebilmektedir. Örneğin Şekil II.7 grafik yöntem olarak incelenip, yorumlanabilir. Uzaklık değerleri ortalama uzaklık değeri ile karşılaştırıldığından, kestirilen olasılık 0.1 ve 0.9 arasında olduğunda ortalama

uzaklık değerinden uzaklaşan gözlemin etkililiğine bakılır. İlgili gözlem noktasının artık değeri çok fazla büyük çıkıyorsa etkili bir gözlemdir ve bu durumda, bu gözlem model yapısını etkileyeceğinden gözlem noktasının veriden çıkarılamayacağına karar verilir.

Sonuç olarak bu modellerin değerlendirilmesi, yorumlanmasıyla modellerin uyum iyiliği araştırılmış olur [10].

BÖLÜM III

DOĞRUSAL OLMAYAN LOJİSTİK REGRESYON MODELİ

Bağımlı değişkenin değerinde gözlemlenen değişim bağımsız değişkenin aldığı değere bağlı olarak sabit çıkmıyorsa, bağımlı değişken ile bağımsız değişken arasında doğrusal olmayan bir ilişki vardır denir. Burada doğrusal olmayan regresyon modelleri olarak "segmented" veya diğer bir deyişle "splayn" fonksiyonlar incelenecektir.

Bağımlı değişken ile bağımsız değişkenler arasındaki doğrusal olmayan ilişkiyi bulmak için lojistik regresyon modelinin doğrusal, karesel, kübik veya daha yüksek dereceli bir model olup olmadığı sınanmalıdır. Eğer türetilen bağımsız değişken sayısı çok fazla ise, kestirilen katsayılara ilişkin standart hatalar oldukça yüksek çıkma eğilimi gösterir ve modele dahil edilen doğrusal terimler istatistiksel anlamda önemli olduğu halde önemsiz çıkar. Bu durumda lojistik regresyon modelinin doğrusallık varsayımının bozulduğu görülür. Böylece doğrusal olmayan lojistik regresyon analizi kullanılarak verilerin modellenmesi tercih edilir.

Doğrusal olmayan lojistik regresyon modeli bilinmeyen parametre vektörünün bileşenlerinin en az birine göre doğrusal olmayan bir fonksiyona sahiptir. Doğrusal olmayan lojistik regresyon analizinin teorisi ve yöntemleri, doğrusal lojistik regresyon modelinin teorisi ve yöntemleriyle benzerdir.

Doğrusal olmayan regresyon modelleri oldukça yararlı olmasına karşın, verilerin normal olmaması halinde, kullanılan basit model uygun değildir. Bazen dönüşüm yapılarak bu sorun çözülebilir. Bununla birlikte, uygulamada veriler göz önüne alınarak en uygun fonksiyonel şekli tahmin etmek zordur. Parametrik modelin uygun olduğu düşünülebilen durumlarda bile, bağımsız değişkenlere göre bağımlı değişkenin fonksiyonel bağımlılığını ortaya çıkarmada, doğrusal olmayan regresyon yararlı bir tekniktir [22].

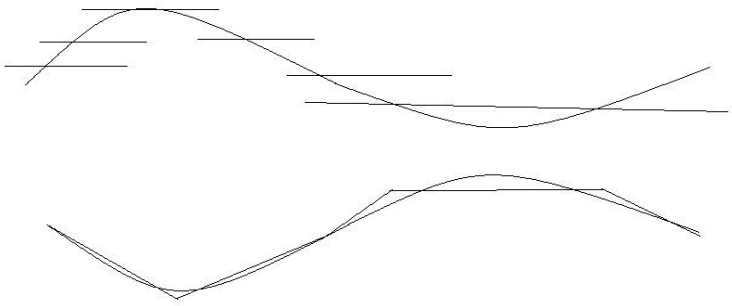
III.1 SPLAYNLAR VE PARÇALI POLİNOM UYDURMA

"Splayn", bir dizi veri noktalarına polinomial bir eğri uydurma veya bu noktalar arasından pürüzsüz olarak geçen ve birçok parçadan oluşan esnek bir eğridir. Splayn fonksiyonlar yeni yeni geliştirilen matematiksel araçlardır. Bu fonksiyonların temel düşüncesi, tanımlanan aralığı bağımsız değişkenlerin gözlem değerleri yardımıyla alt aralıklara bölerek, her bir alt aralıkta farklı bir polinomial fonksiyon ile bağımlı ve bağımsız değişkenler arasındaki ilişkiyi modellendirerek istenilen mertebeden türevi olan sürekli fonksiyon elde etmektir [1].

Bazen düşük dereceli polinomların veriye uydurulmasında zayıf kaldığı, polinom derecesinin yükseltilmesinin de durumu düzeltmediği görülür. Bunun nedeni artık kareler toplamındaki başarısızlık veya açıklanmamış yapıları gösteren artık çizimleridir. Bu problem, fonksiyon, *x* dizisinin farklı parçalarında farklı davrandığında ortaya çıkar. Bazen *x* ve/veya y'deki dönüşümler bu problemi ortadan kaldırır. İdeal yaklaşım, verilerdeki *x* dizisini bölmek ve her bir bölüme uygun bir eğri oluşturmaktır. Splayn fonksiyonlar parçalı polinom uydurmayı gerçekleştirmek için kullanışlı bir yoldur [16].

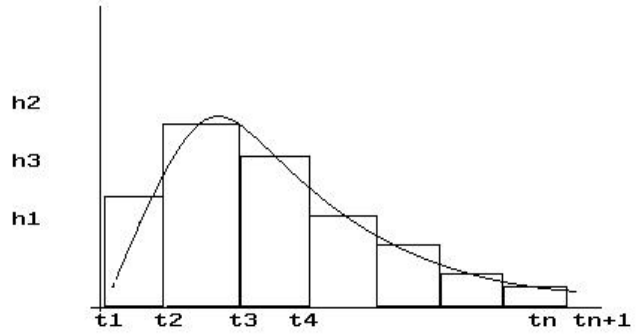
Parçalı polinomların sıradan polinomlara göre bükülgenliği zayıftır, iyi düzgünleştirilmeleri zordur. Splaynlar n-1 sürekli türevlerle kırılma noktalarıyla n.dereceden parçalı polinomlardan oluşur. Splaynların kırılma noktaları, polinom parçalarının birbirine bağlandığı ortak noktalar olan "düğümler(knots)" olarak adlandırılır. n=2 veya 3 veya daha fazla ise splayn düzgün, esnek şekildedir.

Splaynların iki klasik örneği vardır: 0.dereceden olan adım fonksiyonları ve 1.dereceden splaynlar olan kırık çizgiler. Bu örnekler düzgünleştirilmenin olmadığı durumlar şeklinde aşağıda gösterilmiştir.



Şekil III.1 Adım Fonksiyonu ve Kırık Çizgi. 0. ve 1.Dereceden Splayn Yaklaşımıyla İki Düzgün Eğri

Bir histograma ait parabolik splayn yaklaşımı Şekil III.2' de oldukça düzgünleştirilmiş bir biçimde gösterilmiştir.



Şekil III.2 Bir Parabolik Splayn. 0.Dereceden Bir Splayn Histogramının Her Bir Parçasındaki Alana Eşit Olan Parabolik Splayn

Histogram h_i yüksekliğine, t_i düğüm noktasına sahiptir ve parabolik splaynın her parçasının altındaki bölge histogramdaki basamaklara karşılık gelir. Parabolik splaynın ayrıntılı incelenmesi eğrilerin 2.türevlerindeki sıçramaları gösterir.

Genelde parçalı polinomlardan çok splaynların tercihi daha fazladır; çünkü düzgünleştirme durumları bazı katsayıları yok sayar. Bu, düğümlerdeki değerler ve düğümlere ihtiyaç duyulan kırık çizgilerin oldukça kolay görülmesini sağlar. Parçalı polinom $pp(x)$, (k+1) kırılma noktalı, n.dereceden k tane $p_i(x)$ polinom kümesine sahip olan denklem,

$$pp(x) = p_i(x) \qquad x \in [t_i, t_{i+1}], \qquad i = 1, 2, \ldots, k$$

biçiminde gösterebilir. $p_i(x)$ fonksiyonu $a_{i0}, a_{i1}, \ldots, a_{in}$ katsayılarına sahiptir. Kırılma noktaları kümesi $T = (t_1, t_2, \ldots, t_{k+1})$ dir. Parçalı polinomun açık şekilde temsili aşağıdaki gibidir:

$$pp(x) = a_{i0} + a_{i1}(x - t_i) + a_{i2}(x - t_i)^2 + \ldots + a_{in}(x - t_i)^n$$

pp(x)'i hesaplamak için doğru i.aralığı bulmamızı sağlayan eşitlik,

$$a_{0i} + \left[\frac{a_{0i+1} - a_{0i}}{t_{i+1} - t_i}\right](x - t_i)$$

ile verilir. Dolayısıyla değerlendirme boyunca karşılaştığımız $(x - t_i)$'nin katsayısı olan a_{1i} yi hesaplamak zorunluluğu vardır.

Benzer bir gösterim yüksek dereceli splaynlar için mevcuttur. İlk [t_1, t_2] aralığında splayn, herhangi bir polinom olabilir; dolayısıyla a_{i0}, a_{i1}, …, a_{in} katsayılarına ihtiyaç vardır. Sonraki aralıkta $p_2(x)$'in ilk (n-1) türevini t_2 deki $p_1(x)$'e eşitlemek için, t_2 değerinde hesaplama yapılır. Bu polinomlar j=0,…,n için a_{1j} 'den hesaplanabilir. t_2 değeri için $p_2(x)$'in n.türevi, $p_1(x)$'den $p_2(x)$ polinomuna geçiş için bir sıçrama noktasıdır. Çünkü bu geçiş noktası bir polinomun diğeriyle birleştiği ortak bir yerdir. Bu şekilde 3. ve 4. ve sonraki düğümler hesaplanabilir. Dolayısıyla (i=2, 3, …, k) b_i sıçramasının $p_1(x)$ açılımında yer alan katsayıların bulunmasıyla $pp(x)$ gösterimi elde edilmiş olur [20].

"Kesilmiş üslü ifade"(truncated power basis) kullanarak basit bir matematiksel biçim oluşturulmasıyla fonksiyon gösterimi geliştirilir. t, x fonksiyonunun bir parametresi olacak şekilde fonksiyon,

$$(x-\mathrm{t})_+^{\mathrm{i}} = \begin{cases} (x-\mathrm{t})^{\mathrm{i}} & , \ x \geq \mathrm{t} \\ 0 & , \ x \leq \mathrm{t} \end{cases}$$

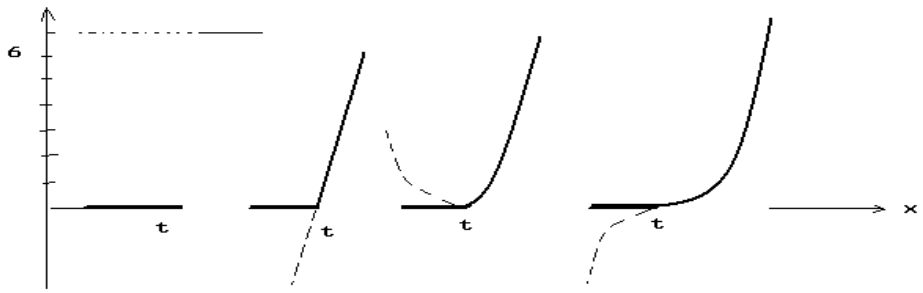
olarak yazılabilir.

Şekil III.3 'de i=0, 1, 2, 3, 4 için fonksiyon çizimleri yer almaktadır. $(x-\mathrm{t})_+^{\mathrm{i}}$ nin türevleri ve integralleri kolayca hesaplanır. Özel olarak x^j fonksiyonuna benzer bir örnek aşağıda verilmiştir:

$$\frac{d^j (x-t)_+^i}{dx^j} = i(i-1)\ldots(i-j+1)(x-t)_+^{i-j} \qquad , \mathrm{j} \leq \mathrm{i} \quad \text{için.}$$

Böylece, $(x-\mathrm{t})_+^{\mathrm{i}}$ nin n.türevi adım fonksiyonudur. t' nin sol tarafı 0, sağ tarafı n! olur. [t_2, t_3] ikinci aralığında $p_1(x)+b_2(x-t_2)_+^n$, $p_1(x)$'den $p_2(x)$'e geçişi sağlayan bir fonksiyon adım fonksiyonudur. Bu fonksiyon, t_2 de n.türevdeki n!.b_2 nin bir sıçramasıyla $p_1(x)$'den oluşturulur. Bu şekilde splaynın bir başka gösterimi de elde edilmiş olur:

$$pp(x) = \sum_{j=0}^{n} a_j (x-t_i)^j + \sum_{i=2}^{k} b_i (x-t_i)_+^n$$



Şekil III.3 Kesilmiş Üslü İfade. t-2≤x≤t+2 için $6(x-t)^0_+$, $6\,(x-t)_+$, $3\,(x-t)^2_+$ ve $(x-t)^3_+$ fonksiyon çizimleri.

Şekil III.3'deki fonksiyonlar düzgünleştirilmiştir ve kübik olanın kırılma noktası gözle görülmeyecek şekildedir [20].

Özetle, splaynlar k.dereceden parçalı polinomlardır. Parçaların ortak noktaları olan düğümler önemli bir rol üstlenir. Çoğunlukla fonksiyon değerlerinin ve ilk (k-1) türevlerinin düğümlere uygunluk gösterdiği görülür; çünkü splayn, (k-1) sürekli türevlere sahip bir fonksiyondur.

İlk ve ikinci türevlerde sürekli kübik splayn (k=3), genellikle uygulamalı problemler için uygundur. h düğüm noktalı bir kübik splayn, $t_1 < t_2 < \ldots < t_h$, aşağıdaki gibi yazılır:

$$E(y) = s(x) = \sum_{j=0}^{3} \beta_{0j} x^j + \sum_{i=1}^{h} \beta_i \left(x - t_i\right)_+^3$$

Düğüm noktalarını açık bir biçimde gösteren splayn modelde süreklilik şartı aranmayabilir. Çünkü değişim noktalarına kadar zaten sürekli olmakla beraber ayrı parçalar şeklinde incelendiğinden bu koşulun aranmasına gerek yoktur. Yukarıda tanımlanan splayn fonksiyon buna örnek olarak verilebilir. Bu şekilde daha az süreklilik koşulu, daha uygun modelin elde edilmesini sağlar. Splayn fonksiyonların kesikli polinomlardan oluşması yani verinin bazı bölgelerde süreklilik şartının sağlamaması durumunda kullanılması ve bu noktalarda değişim noktalarının kullanılmasıyla oluşturulan farklı dereceli polinomların düğüm noktalarıyla birbirine bağlanması ilgili verinin kolayca çözülmesine yardımcı olur.

Düğüm noktalarının bilindiğini varsayalım. Eğer düğüm noktaları kestirilmiş parametreler ise problem bir doğrusal olmayan regresyon problemidir. Düğüm

noktalarının yeri bilindiğinde yukarıda tanımlanan eşitliğin oluşturulması, doğrusal en küçük karelerin açık bir uygulamasıyla tamamlanabilir [16].

Her bir parçadaki polinom derecesine, düğümlerin yeri ve sayısına karar vermek kolay değildir. Wold (1974), bölüm başına en az 4-5 veri noktasıyla mümkün birkaç düğümün var olabileceğini belirtmiştir. Buradaki uygulama önemlidir; çünkü splayn fonksiyonların büyük ölçüdeki önemli esnekliği, "overfit" (aşırı uyumu) veriyi çok kolay kılar. Ayrıca Wold, her bölümün bir büküm noktası ve bir ekstrem noktasından daha fazla olmayacağını belirtmiştir. Yani her bölümde sadece bir değişim söz konusudur. Ekstrem noktaları bölüm merkezinde yer almalı ve büküm noktaları da düğümlere yakın olmalıdır [16].

Pastor ve Guallar'ın makalelerindeki spline yönteminin ise anlaşılması daha kolaydır ve Matlab programında çözümlenmesi oldukça yalındır. Makalelerinde iki parçalı lojistik regresyon modellerini tanımlamışlardır. Epidemiyolojik çalışmalar hastalık riski ve maruz kalan değişken arasındaki ilişkiyi açıklamak için ideal olduğundan makalelerinde alkol alımı ile miyokard enfarktüsünü incelemişlerdir. Burada doz-tepki ilişkisinin, ilgili değişken x, bilinmeyen bir eşik düzeyine ulaştığı zaman beklenmedik şekilde değiştiği görülmüştür ve bu değişim noktası λ ile gösterilmiştir. Alışılmış doz-tepki analiz yöntemleri değişim noktalarının yerini veya güven aralığını tahmin etmede çıkarsama sağlamadığı için kategorik olarak ilgili x değişkenini alt aralıklara ayırmışlardır. Sınırlardaki değişim noktalarının kararsızlığından kaçınmak için splayn regresyon kullanarak her bir kategoride doğrusal veya karesel model bulmayı amaçlamışlardır. Splayn regresyon kategoriler içinde risk değişimine izin verdiğinden doz-tepki analizi dediğimiz lojistik regresyon ilişkisini kestirmede bu yöntem kategorik analizden daha esnektir. Düğüm yeri araştırmacı tarafından keyfi olarak ayarlanmak zorundadır.

Pastor ve Guallar'ın yaptıkları çalışmada kullandıkları yöntem, bu çalışmadaki yönteme paralel bir çalışmadır.

Buradaki lojistik regresyon modelindeki amaç, doz-tepki ilişkisinin belirgin bir biçimdeki değişimlerinde, maruz kalan x değişkeninde bir potansiyel değişim noktası kestirmektir. Lojit ve x değişkeni arasında değişim noktasının her iki yanında iki farklı polinomial fonksiyonun oluşturulması istenmektedir. İki segmentli lojistik regresyon lojit terimle açıklanabilir:

$$g(x, z_1, ..., z_p) = \log\left(\frac{\pi(x, z_1, ..., z_p)}{1-\pi(x, z_1, ..., z_p)}\right)$$

$$= f(x, \lambda) + \alpha_1 z_1 + ... + \alpha_p z_p \qquad \text{(III.1)}$$

Burada $\pi(x, z_1, ..., z_p) = P(y = 1 \mid x, z_1, ..., z_p)$ verilen kovaryant kümesi için hastalık olasılığını belirtir. λ ise bilinmeyen değişim noktalarıdır. İki segmentli polinomial fonksiyon $f(x, \lambda) = \begin{cases} f_1(x) \ , & \text{x} < \lambda \\ f_2(x) \ , & \textit{diğer durumlarda} \end{cases}$ ise farklı çok terimliler için $f_1(x)$ ve $f_2(x)$ vardır. $\alpha_1, ..., \alpha_p$ kovaryantların regresyon katsayılarıdır.

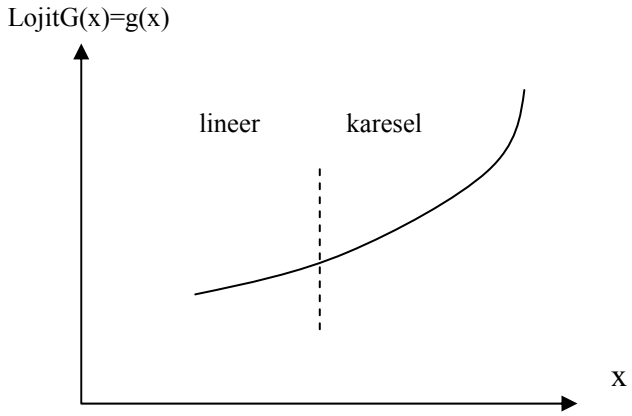
Pratik uygulamalar için doğrusal, karesel veya kübik gibi polinomlar sınıfı belirtilmelidir. Değişim noktasında düzgün geçişin istenilen özelliği en küçük karesel polinomlardan birini seçmektir. Diğer taraftan kareselden daha büyük dereceli polinomların katsayılarının epidemiyolojide yorumlanması zor olacaktır. İki segmentli lojistik regresyon için temel alternatifler doğrusal-karesel, karesel-doğrusal veya karesel-karesel gibi modellerdir.

Kovaryantların bulunduğu durumda doğrusal-karesel biçimdeki iki segmentli lojistik regresyon aşağıdaki gibi ifade edilebilir:

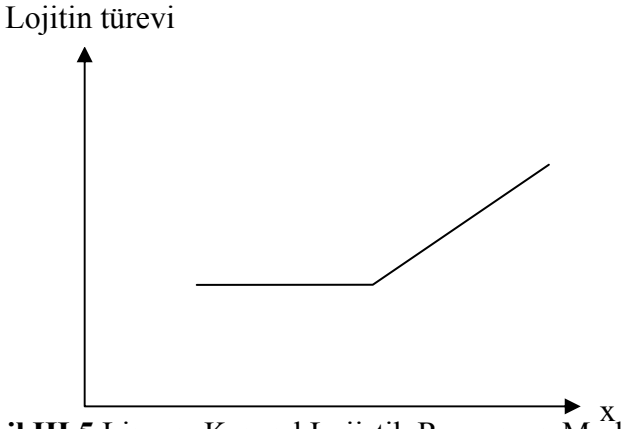
$$g(x, z_1, ..., z_p) = \beta_0 + \beta_1 x + \beta_2 (x - \lambda)^2 I_{(x \geq \lambda)} + \alpha_1 z_1 + ... + \alpha_p z_p \qquad \text{(III.2)}$$

$$I_{(x \geq \lambda)} = \begin{cases} 1 \ , & \text{x} \geq \lambda \\ 0 \ , & \textit{diğer durumlarda} \end{cases}$$

λ bilinmeyen değişim noktasını temsil eder. β_0, β_1, β_2 etkilenen x değişkeninin bilinmeyen regresyon katsayılarıdır. Doğrusal-karesel lojitte iki segmentli modelde, etki, λ değişim noktasının alt parçasında kalan lojitte doğrusaldır. Bu parçada lojit $\beta_1 + 2\beta_2(x - \lambda)$ şeklindedir. λ değişim noktasının üst tarafında, etki kareseldir ve $\beta_0 + \beta_1 x + \beta_2 (x - \lambda)^2$ şeklinde verilir. Bu modelde lojitin türevi sürekli olduğundan iki parçaya karşı risk eğiliminin düzgün bir değişimi vardır. Eğim, λ değişim noktasının aşağısında ise β_1 'dir. Değişim noktasının yukarısına doğru bir değişiklik başlar; burada eğim $\beta_1 + 2\beta_2(x - \lambda)$ şeklinde doğrusal olacak şekilde sürekli bir değişim gözlenir.



Şekil III.4 Lineer- Karesel Lojistik Regresyon Modeli



Şekil III.5 Lineer- Karesel Lojistik Regresyon Modelin Türevi

Şekil III.4 'deki doğrusal-karesel lojistik regresyon modelini uydurmak için $\alpha_1,...,\alpha_p$ kovaryantlarını ; β_0, β_1, β_2 maruz kalan değişken katsayılarını ve λ değişim noktasını kestirmek gerekir. Bu bilinmeyen parametrelerin en çok olabilirlik kestiricileri $\hat{\lambda}$, $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ ve $\hat{\alpha}_1,...,\hat{\alpha}_p$ şeklindedir. Bunlar log-likelihood fonksiyonunu maksimize edecek şekilde kestirilir:

$$l = \log(L)$$

$$= \sum_{i=1}^{n} [y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i)]$$

$$= \sum_{i=1}^{n} [y_i g_i - \log(1 + e^{g_i})]$$

y_i i.kişi için hastalık durumunu belirtir. $\pi_i = \pi\left(x_i, z_{i1}, ..., z_{ip}\right)$ ve $g_i = g\left(x_i, z_{i1}, ..., z_{ip}\right)$ hastalık olasılığını ve i.kişi için bu olasılığın lojitini verir. (III.2) nolu eşitlik doğrusal olmadığından ve devamında sabit olmadığından standart iteratif olarak yeniden ağırlıklandırılmış en küçük kareler yöntemi bu modelin uydurulmasında kullanılamaz. Bunun yerine tüm model parametrelerini eşzamanlı olarak kestirmek için IRLS yönteminin başka bir versiyonu önerilmiştir [19].

Standart regresyondaki çıkarsamalar, en çok olabilirlik oran istatistiğinin asimptotik χ^2 dağılımından direkt olarak elde edilen olabilirlik temelli yöntemler ya da ML (maksimum likelihood) kestiricilerinin asimptotik normal dağılımı üzerine kurulu, Wald tipi yaklaşımlar kullanılmasına yönlendirilir. İki segmentli regresyonda, bu asimptotik özellikler belli bir değişim noktasının varlığına bağlıdır. Değişim noktasının varlığı için kavramsal bir test, doz-yanıtın doğrusal homojen bir örneği sayılan $H_0 : \beta_2 = 0$ hipotezinin test edilmesidir. Eğer $\beta_2 = 0$ ise ikinci segmentin olmadığı sonucu ortaya çıkar; ama bu istenilen bir durum değildir: Çünkü segmente sahip bir veri üzerinde çalışılır. O halde λ iyi tanımlanmamıştır; parametre uzayının bir bozulması mevcuttur. Bu durumda ML kestiricileri asimptotik olarak normal değildir; Wald ve olabilirlik oran istatistiklerinin dağılımları sırasıyla standart normal ve χ^2 dağılımlarına yakınsamaz. Burada hipotez testleri veya güven aralıkları gibi tanımlı tüm çıkarsamalar, λ'nın varlığına dayandırılır. λ için %100(1- α) güven aralığında Wald modelinde $\hat{\lambda} \pm Z_{1-\alpha/2} s(\hat{\lambda})$ kestirilir. $s(\hat{\lambda})$, modifiye edilmiş IRLS algoritmasının son iterasyonunda Fisher bilgi matrisinin tersinden elde edilen standart hatanın bir kestiricisidir. λ için daha düzgün güven aralığı olabilirlik oran istatistiğinin kullanılmasıyla elde edilebilir:

$$\wedge_{\lambda} = 2l(\hat{\lambda}, \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) - 2l(\lambda, \hat{\beta}_0(\lambda), \hat{\beta}_1(\lambda), \hat{\beta}_2(\lambda))$$

$\hat{\beta}_j(\lambda)$, λ verildiğinde (j=0, 1, 2) β_j'nin ML kestirimidir. $\chi^2_{(l;\ 1-\alpha)}$, 1 serbestlik dereceli %100(1- α) güven düzeyli ki-kare değeri olmak üzere, λ olabilirliğe dayalı güven aralığı $\wedge_{\lambda} \leq \chi^2_{(l;\ 1-\alpha)}$ biçiminde gösterilir. Bu aralık genellikle Wald yaklaşımıyla oluşturulan terimlerde daha geçerlidir.

Spline fonksiyondaki gruplama farklı şekillerde gösterilebilir. Liu ve arkadaşlarının, doğum ile yumurtalık kanserindeki geçici azalmaya ilişkin çalışmalarında λ değişim noktası fark biçiminde gösterilmiş olup bu şekilde her bir değişim noktasından sonraki bölüm için değişken sayısının artmasına neden olmuşlardır. Makalelerinde kullandıkları spline model,

$\beta_0 + \beta_p$ (parity) $+ \beta_a$ (age) $+ \beta_{aafd}$ (age at first delivery) $+ \beta_1 x_1 + \beta_2 x^2_1 + \beta_3 x^3_1 + \beta_4 x^2_2 + \beta_5 x^2_3$

dir. Burada yaş ve ilk doğumdaki yaş birer kategorik zaman değişkenleridir. x_1 ise ilk doğumdan itibaren geçen zamanın sürekli değişkeni olmak üzere λ değişim noktaları 3 ve 38 olduğundan diğer değişkenler şöyledir [13]:

$$\begin{cases} x_2 = x_1 - 3 & , \ x_1 > 3 \\ x_2 = 0 & , \ \textit{diğer durumlarda} \end{cases}$$

$$\begin{cases} x_3 = x_1 - 38 & , \ x_3 > 38 \\ x_3 = 0 & , \ \textit{diğer durumlarda} \end{cases}$$

Pastor ve Guallar, tüm parametre uzayında kalan parametrelerin sürekli olacak şekilde ilk kısmi türevleriyle iki segmentli bir polinomial fonksiyon olan $g(x, z_1, ..., z_p) = f(x, \lambda) + \alpha_1 z_1 + ... + \alpha_p z_p$ lojistik modeline uyan bir algoritma geliştirdiklerini belirtmektedirler. $f(x, \lambda)$'nın özel parametrizasyonlarına örnek olarak,

$f(x, \lambda) = \beta_0 + \beta_1 x + \beta_2 (x - \lambda)^2 I_{(x \geq \lambda)}$ doğrusal-karesel,

$f(x, \lambda) = \beta_0 + \beta_1 x + \beta_2 (x - \lambda)^2 I_{(x \leq \lambda)}$ karesel- doğrusal,

$f(x, \lambda) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 (x - \lambda)^2 I_{(x \geq \lambda)}$ karesel-karesel,

modeller verilebilir. Buradaki başlıca amaç iki segmentli lojistik regresyon modelinin log-olabilirlik fonksiyonu,

$$l = \sum_{i=1}^{n} \left[y_i g_i - \log(1 + e^{g_i}) \right] \qquad \text{(III.3)}$$

ifadesinin $Q = (\lambda, \beta_0, \beta_1, ..., \beta_q, \alpha_1, ..., \alpha_p)$ bilinmeyen parametreler vektörüyle maksimize edilmesiyle en çok olabilirlik kestiricilerini $\hat{Q} = (\hat{\lambda}, \hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_q, \hat{\alpha}_1, ..., \hat{\alpha}_p)$ elde etmektir. Skor vektörü $u(Q) = \partial l / \partial Q$, aşağıdaki biçimde hesaplanabilir:

$$\frac{\partial l}{\partial Q}=\sum_{i=1}^{n}(y_i-\pi_i)\frac{\partial g}{\partial Q_r} \quad , \; r=1,...,p+q+2 \tag{III.4}$$

g_i parametre uzayı olan Q'ya göre birinci kısmi türevlere sahip sürekli bir ifadedir. $E(y_i)=\pi_i$ ve $\text{var}(y_i)=\pi_i(1-\pi_i)$ olduğundan, $u(Q)$ 0 olasılığına sahiptir ve kovaryans matrisi $I(Q)$ Fisher bilgi matrisine eşittir:

$$I_{r,s}(Q)=\left\{E(u(Q)u^T(Q))\right\}_{r,s}=\sum_{i=1}^{n}\pi_i(1-\pi_i)\frac{\partial g_i}{\partial Q_r}\frac{\partial g_i}{\partial Q_s} \tag{III.5}$$

Buradaki $I_{r,s}(Q)$ Fisher matrisinin (r,s)ninci elemanlarını gösterir ($r,s=1,...,p+q+2$). En çok olabilirlik kestiricisi $\hat{Q}$, $u(Q)=0$ olabilirlik denklemlerinin $p+q+2$ tanesinin çözümüyle elde edilir. Bu denklemler Q parametrelerinde doğrusal olmadıklarından iteratif yöntemlerle $\hat{Q}$ elde edilmelidir. İki segmentli lojistik regresyon modeli Fisher skorlama yöntemi kullanılarak uyarlanabilir. Fisher bilgi matrisi $I(Q^{(k)})$ ve skor vektörü $u(Q^{(k)})$ olmak üzere Q'nın mevcut kestirimi $Q^{(k)}$ olsun. $Q^{(k+1)}$ bir sonraki yeni parametre kestirimi olmak üzere $I(Q^{(k)})(Q^{(k+1)}-Q^{(k)})=u(Q^{(k)})$ denkleminden türetilir. Bu da (III.4) ve (III.5) denklemlerinin değerlendirilmesiyle aşağıdaki denklem sisteminin sonuçlarıdır ($r=1,...,p+q+2$):

$$\sum_{i=1}^{n}\pi_i^{(k)}(1-\pi_i^{(k)})\frac{\partial g_i}{\partial Q_r}\Big|_{Q^{(k)}}\left[\sum_{s=1}^{p+q+2}\frac{\partial g_i}{\partial Q_s}\Big|_{Q^{(k)}}(Q_s^{(k+1)}-Q_s^{(k)})-\frac{y_i-\pi_i^{(k)}}{\pi_i^{(k)}(1-\pi_i^{(k)})}\right]=0 \tag{III.6}$$

g_i'nin $Q^{(k)}$ 'ya göre birinci dereceli Taylor serisi kullanılarak, $\left[\partial g_i/\partial Q|_{Q^{(k)}}\right]^T(Q^{(k+1)}-Q^{(k)})=g_i^{(k+1)}-g_i^{(k)}$ ifadesi (III.6)'deki eşitlikten,

$$\sum_{i=1}^{n}\pi_i^{(k)}(1-\pi_i^{(k)})\frac{\partial g_i}{\partial Q_r}\Big|_{Q^{(k)}}\left[g_i^{(k+1)}-\left(g_i^{(k)}+\frac{y_i-\pi_i^{(k)}}{\pi_i^{(k)}(1-\pi_i^{(k)})}\right)\right]=0 \tag{III.7}$$

şeklinde yazılabilir. Burada regresyon modelinin ağılıkları $w_i^{(k)}=\pi_i^{(k)}(1-\pi_i^{(k)})$ ve yanıt değişkenleri $v_i^{(k)}=g_i^{(k)}+(y_i-\pi_i^{(k)})/w_i^{(k)}$ olmak üzere (III.7) nolu denklem ağırlıklandırılmış en küçük kareler denklemidir. Böylece $Q^{(k+1)}$ ağırlıklandırılmış

modelden hesaplanabilir ve bu her bir iterasyondan yeniden hesaplandırılması yoluyla olur [19].

Gauss-Newton algoritmasının Hartley modifikasyonu, ağırlıklandırılmış iki segmentli regresyon modellerinin çözümünde kullanılabilir. Temel iteratif sürecin k'nıncı derecesiyle iç içe olan metodun l 'inci iterasyondaki Q'nın bir kestirimi $Q^{(k,l)}$ olsun. Her bir gözlem için $Q^{(k,l)}$ çevresinde $g_i(Q)$'nın birinci dereceden Taylor serisi $g_i(Q) \cong g_i(Q^{(k,l)}) + \left[\partial g_i / \partial Q|_{Q^{(k,l)}}\right]^T (Q - Q^{(k,l)})$ denklemini verir. $r_i(k,l) = v_i^{(k)} - g_i(Q^{(k,l)})$ artığı, $r(k,l)$ artık vektör ve $G(k,l)$ $\partial g_i / \partial Q|_{Q^{(k,l)}}$'den (i,r) elemanlı bir matris olmak üzere $r(k,l) = G(k,l)(Q - Q^{(k,l)}) + \varepsilon$ tüm gözlemlerden oluşan bir matris biçiminde veya aşağıdaki gibi açıklanabilir:

$$r_i(k,l) = \left[\frac{\partial g_i}{\partial Q}|_{Q^{(k,l)}}\right]^T (Q - Q^{(k,l)}) + \varepsilon \qquad \text{i=1,…,n.} \qquad \text{(III.8)}$$

$\mathbf{W}(k) = diag\left\{\pi_i^{(k)}(1-\pi_i^{(k)})\right\}$ ağırlıklarının diagonal matrisiyle doğrusallaştırılan regresyon modelinin çözümü -ordinary- ağırlıklandırılmış en küçük kareler yöntemi kullanılarak elde edilir. $\mathbf{d}(k,l) = \left[\mathbf{G}^{\mathbf{T}}(k,l)\mathbf{W}(k)\mathbf{G}(k,l)\right]^{-1}\mathbf{G}^{\mathbf{T}}(k,l)\mathbf{W}(k)r(k,l)$ dir. Dolayısıyla çözüm, segmentli polinomial regresyon modelleri için Gauss-Newton algoritmasının Hartley modifikasyonu gereğince, bir sonraki kestirici $Q^{(k,l+1)}$, iki segmentli regresyon modelinin ağırlıklandırılmış hata toplamını minimize eden $Q^{(k,l)} + \eta\mathbf{d}(k,l)$ ifadesinde 0 ve 1 arasındaki bir skaler olan $\eta(k,l)$'nın yer aldığı $Q^{(k,l)} + \eta(k,l)\mathbf{d}(k,l)$ tarafından belirlenir.

Bu iteratif aşamaların her biri, $Q^{(k,l+1)}$ ve $Q^{(k,l)}$ arasındaki bağıl değişim önemsiz olduğu zaman sona erer. Son kestirici, $Q^{(k+1)}$ sonraki iteratif sürecin başlangıç değeri olarak kullanılır. $w_i^{(k+1)}$ ağırlığı ve $v_i^{(k+1)}$ yanıt değişkeni $Q^{(k+1)}$'da yeniden hesaplanmalıdır ve bu iteratif işlemleri kullarak iki segmentli regresyon modelinin tekrar ağırlıklandırılmasına uygun olmalıdır. Bu süreç yakınsama olana kadar tekrarlanır.

Burada sözü edilen algoritmanın doğru bir yakınsamasını elde etmek için $Q^{(0)}$ başlangıç kestiriminin seçimi çok önemlidir; çünkü (III.3) log-olabilirlik fonksiyonu yerel maksimuma yol açabilir. Başlangıç noktalarını elde etmek için iyi

bir altenatif, grid-search yaklaşımının kullanılmasıdır. Bu yöntem x maruz kalan değişkenin aralığında yer alan eşit aralıklı noktaları önceden belirler. Her bir nokta için iki segmentli lojitli bir splayn lojistik regreyon bulunabilir ve log-olabilirliği değerlendirilir. Log-olabilirlikteki nokta, $Q^{(0)}$ başlangıç değeri olarak splayn lojistik regresyondan diğer parametre kestirimleriyle birlikte kullanılan en yüksek değere ulaşır [19].

Kullanılan splayn lojistik model iki segmentli olmasa da parametrelerin bulunmasına kadar olan işlemler Gauss-Newton algoritmasıyla aynıdır. (III.7) nolu denklemde açık bir şekilde gösterilmiş olmasa da bu yöntemle ağırlıklandırılmış olarak çözümlenmektedir. Sonuç olarak iyi bir başlangıç değeriyle Gauss-Newton algoritmasının kullanımı önemli rol oynamaktadır.

Splayn regresyon doz-yanıt ilişkisini kestirmek için kullanılan geleneksel tüm metotlardan daha avantajlıdır. Kompleks dağılımlara uyarlanabilir ve grafiklerde değişime daha elverişlidir [4].

BÖLÜM IV

IV.1 UYGULAMA-1

Bu uygulamada, 243 kişilik hemoglobin verisi üzerinden splayn yöntem aracılığıyla doğrusal olmayan lojistik regresyon analizi incelendi. Bu çalışmanın amacı, hemoglobin eksikliğinin etkili olduğu hastaların yaşama şansını tahmin edecek bir doğrusal olmayan lojistik regresyon modeli kurmaktır. Öncelikle bağımsız değişken olarak modelde yer alan hemoglobin hakkında bilgi vermek doğru olacaktır.

Alyuvarlara karbondioksit ve oksijen taşıyabilme yeteneği kazandıran hemoglobin molekülleridir. Hemoglobin aynı zamanda kana kırmızı rengini de veren maddedir.

Hemoglobin yapısında bulunan zincirlerden yalnızca birinin bile aminoasit yapısında ortaya çıkan bir değişiklik, anormal hemoglobin üretimine neden olur. Bu yapısal değişiklik çoğunlukla belirti vermezken, bazı durumlarda hemoglobinin oksijen taşımasını önemli ölçüde etkileyerek hastanın yaşamını tehdit edebilir.

Anemi (Kansızlık), kan hemoglobin düzeyinde veya kırmızı kan hücreleri sayısında azalma ve sonucunda ortaya çıkan bulgulardır. Bir başka ifadeyle, anemi hemoglobin miktarının yaş ve cinsiyete göre dünya sağlık örgütü tarafından kabul edilen kriterlerin altında kalmasıdır. Bu kriterler erişkin erkeklerde 13 g/dL, kadınlarda 12 g/dL nin altı kabul edilir. 6 ay ile 6 yaş arası çocuklarda 11 g/dL nin, 6-14 yaşlarda 12 g/dL nin altı anemidir [24].

Ortaya çıkan şikayetler ve saptanan bulgular doku ve hücrelere yetersiz oksijen taşınmasına bağlı olarak gelişmektedir. Anemili hastalarda yorgunluk, hafif

çarpıntı ve nefes darlığı gelişebilir. İleri düzeyde bir anemide ise, bütün bu bulgular, istirahat halinde görülmesinin yanı sıra; kulak çınlaması, baş dönmesi, baş ağrısı, uyuma güçlüğü, iştahsızlık, kilo kaybı, adet kanamalarının düzensizliği veya fazlalığı, adet görmeme ve iktidarsızlık gibi bulgular ortaya çıkabilir. Aneminin sık görülen bulgularından çarpıntı, anemi yüzünden dokularda oluşan oksijen açlığını gidermek amacıyla, kalbin atım hızını ve her atımda pompaladığı kan miktarını artırması nedeniyle ortaya çıkar. Buna rağmen dokularda yeterli oksijen sağlanamıyorsa, solunum sayısının artması ve nefes darlığı ortaya çıkar. Uzayan anemilerde ve yaşlı kişilerde veya kalp hastalığı olanlarda kalp yetmezliğine ait bulgular gelişebilir. Anemideki en belirgin bulgulardan birisi de solukluktur. Aneminin şiddetine bağlı olarak, ağız ve göz kapağı içindeki deride ilk olarak fark edilebilen solukluk, aneminin ilerlemesi ile avuç içinde, tırnak yataklarında ve deride de belirginleşir. Anemiye yol açan nedene bağlı olarak çok çeşitli bulgular gelişebilir.

En sık rastlanan anemi türleri demir eksikliğine bağlı anemi, folik asit eksikliğine bağlı anemi, Vitamin B-12 eksikliği anemisidir.

Anemi gibi kan hücrelerinin sayısındaki deformasyona bağlı olarak gelişen ve tehlike içeren bir durum da lösemidir. Bunun tipik bir örneği de hemoglobin miktarındaki azalmayla görülen Akut Myeloid Lösemi (AML)'dir. Olgunlaşmamış bu hücreler kemik iliğinde çok yüksek sayılara ulaşırlar ve normal kan hücrelerinin üretimini azaltırlar. Sonuçta anemi (kansızlık - kırmızı kan hücresi üretiminde azalma) ve sık enfeksiyona yakalanma (beyaz kan hücresi üretiminde azalma) durumu ortaya çıkabilir. Ergenlik çağında ve 20 li yaşlarda saptanan lösemilerin %50 sini, yetişkinlerdeki lösemilerin de %20 sini AML oluşturur [24].

Genel olarak lösemiler tüm kanserlerin %2 sini oluştururlar. Erkeklerde lösemi daha sık gözlenmektedir. Ayrıca beyaz ırkta da daha sıktır. Yetişkinlerde lösemi tanısı konma sıklığı çocuklardan 10 kat daha fazladır ve risk yaşla birlikte artar. Çocuklar arasında ise 4 yaş altında daha sık gözlenir. Bu çalışmada yetişkin kadınları içeren 243 kişilik bir veri kümesinde hemoglobin değerlerine ait bulgular incelendi.

Çalışmada bağımsız değişken hemoglobin değerlerinden oluşmaktadır; bağımlı değişken ise hemoglobin eksikliğinin neden olduğu anemi ve lösemi risklerinin mevcut olduğu y=1 ve risk barındırmayan y=0 şeklinde iki sonuçlu değişkendir.

Burada riske neden olan diğer kan sayımı bulguları dikkate alınmadı, tek değişkenli model üzerinde çalışma gerçekleştirildi.

SAS programında Mulla'nın önerdiği algoritmaya göre hemoglobin verisi çözümlendi. Mulla yaptığı çalışmada bağımsız değişken olan x değişkenini değişim aralıklarına göre gruplandırarak kategorik şekle getirmiştir. Bu çalışmada, hemoglobinin belirgin etki noktaları ise 8.6 g/dL ve 12 g/dL'dir. Bu değişim aralıklarına göre kategorilendirme sonucu her kategori için yeni isimlendirmeler aşağıdaki biçimdedir:

```
if hemoglobin1<8.6 then hgbcat=1;
if hemoglobin1>=8.6 and hemoglobin1<=11.9 then hgbcat=2;
if hemoglobin1>=12.0 then hgbcat=3;

if hemoglobin1>8.6 then hgb2=hemoglobin1-8.6;
                                    else hgb2=0;

if hemoglobin1>12.0 then hgb3=hemoglobin1-12.0;
                                    else hgb3=0;
```

Burada üç tane bölge vardır. İlk bölge 8.6 g/dL değerinin altında kalan bölgedir ve ikinci bölge 8.6 g/dL≤ hemoglobin < 12 g/dL ve son olarak üçüncü bölge ise 12 g/dL ≤ hemoglobin şeklindedir. Bu algoritmada ilk bölgedeki bağımsız değerlere 'hemoglobin1' denilirse hgb1 adlandırmasıyla, ikinci bölgedeki bağımsız değişken kategorilendirme gereği hgb2=hgb1-8.6 , üçüncü bölgedeki değişken ise hgb3=hgb1-12.0 olarak yazılabilir. Değişken isimlendirmesi bir önceki bölümde bahsedilen Liu ve arkadaşlarının kullandıkları tanımlamaya benzer şekildedir. Doğrusal olmayan bir model yapısı var olduğundan doğrusal olmama durumunu modelde gösterebilmek adına çeşitli algoritma atamaları oluşturularak PROC LOGISTIC, Version 9.1 SAS / STAT paket programı kullanıldı.

Hemoglobin çalışması iki farklı model yapısı üzerinden incelendi. Modellerde değişim noktasının aynı olmasıyla model yapılandırmasının öneminin belirtilmesi amaçlandı.

Değişim noktalarının ayırdığı bölgeler kategorilendirme yapılarak aşağıdaki modelleme yapıldı. Model_1 çalışması için,

```
hgbz1=hemoglobin1;
hgbz2=(hemoglobin1**2)-(hgb3**2);
hgbz3=(hgb2**2)-(hgb3**2);
```

model_2 çalışması için de,

```
hgbz1=hemoglobin1;
 hgbz2=(hemoglobin1**2)-(hgb3**2);
 hgbz3=(hgb2**2)+(hgb3);
```

yapılandırılması oluşturulmuştur. Buna göre 243 gözlemin her iki model için yanıt profili aşağıdaki gibidir:

Tablo IV.1 HGB Modelleri için Yanıt Profili.

Ordered Value	Outcome	Total Frequency
1	1	46
2	0	197

Her iki model için model uygunluk istatistiği birbirine benzer sonuçlar vermiştir. Hangi modelin daha iyi olacağını belirlemede henüz yeterli bir bilgi olmamakla beraber sonuçları aşağıdaki tabloda yer almaktadır:

Tablo IV. 2 HGB Model_1 için Model Uygunluk İstatistiği.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	237.811	87.672
SC	241.304	101.644
-2 Log L	235.811	79.672

Tablo IV. 3 HGB Model_2 için Model Uygunluk İstatistiği.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	237.811	88.530
SC	241.304	102.502
-2 Log L	235.811	80.530

Modeller için sıfır hipotez testi aşağıda verilmiştir.

Tablo IV. 4 HGB Model_1 için Sıfır Hipotez Testi.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	156.1386	3	<.0001
Score	147.3557	3	<.0001
Wald	24.7096	3	<.0001

Tablo IV. 5 HGB Model_2 için Sıfır Hipotez Testi.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	155.2805	3	<.0001
Score	141.9816	3	<.0001
Wald	28.3060	3	<.0001

Her iki model için dikkat edilmesi gereken belirleyici aşamalardan biri en çok olabilirlik kestirimi analizidir. Model_1'de parametrelerden birinin kesim noktası ve standart hatasının oldukça yüksek çıkması düşündürücüdür. Modeller arasında şimdilik standart hata kestirimine bakarak model_2'nin daha iyi sonuç verdiği söylenilebilir. Kestirilmiş katsayının standart hatasına bölünmesi sonucu elde edilen standart normal dağılım olarak tanımlanan Wald istatistiğinin karesinin alınmasıyla Wald ki-kare testi elde edilir ve bu test modele girmesi muhtemel değişkenlerin ne kadar gerekli olduğunu sınar. Bu durumda sıfır hipotezi ile, ilgili değişkenin modelde sıfır olup reddedilmesi gerekip gerekmediğine karar verilir. Burada *p* değerleri Wald ki-kare istatistiğinden büyük çıktığı için Sıfır hipotezi reddedilir. Bir başka deyişle ilgili parametrenin sıfır olarak kabul edilmesiyle oluşturulan hipotez koşulu sağlanmaz ve bu da demektir ki parametrenin modelde bulunması, yani sıfır olarak alınmaması gereklidir. Modelde yer alması önemlidir.

Tablo IV. 6 HGB Model_1 için En Çok Olabilirlik Kestirimi.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr >ChiSq
Intercept	1	541.9	1023.3	0.2804	0.5964
hgbz1	1	-125.7	238.6	0.2776	0.5983
hgbz2	1	7.3134	13.9093	0.2765	0.5990
hgbz3	1	-8.2594	14.1303	0.3417	0.5589

Tablo IV. 7 HGB Model_2 için En Çok Olabilirlik Kestirimi.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr >ChiSq
Intercept	1	90.4324	205.7	0.1933	0.6602
hgbz1	1	-20.1474	48.6131	0.1718	0.6785
hgbz2	1	1.1439	2.8675	0.1592	0.6899
hgbz3	1	-1.9040	3.1569	0.3638	0.5464

Karşılaştıracağımız bir diğer özellik ise Odds oran kestirimidir. Hastalık riskini diğer değişkenlere göre oransal ifade eden Odds kestirimi modelin yorumlanması ve kabul edilebilirliğinde önemli rol oynar. Model_1'deki nokta kestirimleri güven aralıklarının dışında kalmaktadır. Bu nedenle model_2'nin tercih edilmesi daha uygun olacaktır.

Tablo IV. 8 HGB Model_1 için Odds Ratio Kestirimi.

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
hgbz1	<0.001	<0.001	>999.999
hgbz2	>999.999	<0.001	>999.999
hgbz3	<0.001	<0.001	>999.999

Tablo IV. 9 HGB Model_2 için Odds Ratio Kestirimi.

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
hgbz1	<0.001	<0.001	>999.999
hgbz2	3.139	0.011	866.219
hgbz3	0.149	<0.001	72.491

Buradaki odds oran kestiriminde daha iyi model olarak kabul edilen model_2'de hgbz1 değerinin kestirimi güven aralığı sınırında çıkmaktadır.

HGB model_2 çalışmasının sonuçları model istatistiği, en çok olabilirlik kestirimi ve odds oran kestirim sonuçlarına dayanarak model_1'e göre çok daha kabul edilebilir bir çalışmadır. Ayrıca ROC (receiver operating characteristic) eğrisinin altında kalan alan olarak tanımlanan c (concordance rate) ile sembolize edilen uyumluluk oranının 1'e yakın olması modelin kabul edilebilirliğini istatistiksel olarak desteklemektedir. Hemoglobin çalışmasında model_2'nin c değeri 0.955 olarak sonuç vermiştir. Model_2 için Hemoglobin SAS Sonucu, her bir bağımsız değişkenle risk durumları, Ek-1'de verilmiştir. Model_2'nin SAS program sonucu aşağıda gösterilmektedir:

The LOGISTIC Procedure

Model Information

Data Set	WORK.STUDY	
Response Variable	outcome	outcome
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	243
Number of Observations Used	243

Response Profile

Ordered Value	outcome	Total Frequency
1	1	46
2	0	197

Probability modeled is outcome=1.

Model Convergence Status

Quasi-complete separation of data points detected.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariate
AIC	237.811	88.530
SC	241.304	102.502
-2 Log L	235.811	80.530

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	155.2805	3	<.0001
Score	141.9816	3	<.0001
Wald	28.3060	3	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Standard Estimate	Wald Error	Chi-Square	Pr > ChiSq
Intercept	1	90.4324	205.7	0.1933	0.6602
hgbz1	1	-20.1474	48.6131	0.1718	0.6785
hgbz2	1	1.1439	2.8675	0.1592	0.6899
hgbz3	1	-1.9040	3.1569	0.3638	0.5464

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
hgbz1	<0.001	<0.001	>999.999
hgbz2	3.139	0.011	866.219
hgbz3	0.149	<0.001	72.491

Association of Predicted Probabilities and Observed Responses

Percent Concordant	93.8	Somers' D	0.910
Percent Discordant	2.8	Gamma	0.943
Percent Tied	3.4	Tau-a	0.281
Pairs	9062	c	0.955

IV.1 UYGULAMA-2

Uygulama-1'de SAS programında splayn model yapılandırılması incelendi. Uygulama-2'de ise Matlab programında bir başka uygulama olan serum albümin incelenecektir. Kan plazmasının veya serumunun en önemli bölümünü oluşturan albumin, dokuların temel maddelerinden olup önemli bir proteindir. Albümin, suda çözünürlükleri düşük olan yağ asitlerinin kandaki başlıca taşıyıcısıdır. Bunun yanı sıra, oksijen serbest radikallerine bağlanarak bunları kontrol altına alır, ayrıca bilirubin (hem molekülünün yıkımı sırasında ortaya çıkar) gibi suda çözünmeyen bazı zehirli metabolizma ürünlerine bağlanarak onları zararsız kılar. Kandaki albümin seviyesinin düşüklüğü ciddi risk teşkil eder. Dolayısıyla bu çalışmada, x değişkeni bağımsız değişken olmak üzere 117 kişiden oluşan serum albümin değeridir. y değişkeni de albümin eksikliğinden kaynaklanan ölüm riskidir. Değişim noktaları da tıbbi açıdan önemli sayılan 2.5g/100mL ve 3.5g/100mL değerleridir. Bu değerlerin altında kalan noktalar riski temsil eder. Norm aralık birinci dereceden 3.5g/100mL'dir. Bu değerden daha düşük seviyeler, karaciğerde sentezlenen protein türevi olan albüminin artık sentezlenmemesi sebebiyle vücutta ödem oluşturur. İleri derecede albümin eksikliğinin ölüme neden olduğu bilinmektedir. Belirleyici değişim noktası 3.5g/100mL'den sonra ciddi tehlike arz eden seviye 2.5g/100mL olarak görülmüştür [17].

Matlab programında Pastor ve Guallar'ın model yapısına benzer bir model serum albümin çalışması için oluşturuldu. Bu modelde değişim noktaları lamda1 ve lamda2 olarak ifade edilmiştir. Buna göre model bu değişim noktalarından öncesi ve sonrası olarak es1 ve es2 olarak belirtilmiştir. Daha açık bir ifadeyle,

```
if x(i)>=2.5, es1(i)=1;
        else es1(i)=0;
      end

if x(i)>=3.5, es2(i)=1;
        else es2(i)=0;
      end
```

$$es1_i = \begin{cases} 1 & , x_i \geq 2.5 \\ 0 & , \text{ diğer durumlarda} \end{cases}$$

$$es2_i = \begin{cases} 1 & , x_i \geq 3.5 \\ 0 & , \text{ diğer durumlarda} \end{cases}$$

şeklinde kategorilendirilebilir. Model fonksiyon ise

$$\pi(x_i) = epi(i) = \frac{\exp\left(b_0 + b_1(b_2 - x_i)^3 + b_3(x_i - \lambda_1)^2 es1_i + b_4(x_i - \lambda_2) es2_i\right)}{1 + \exp\left(b_0 + b_1(b_2 - x_i)^3 + b_3(x_i - \lambda_1)^2 es1_i + b_4(x_i - \lambda_2) es2_i\right)}$$

biçiminde alınmıştır. Parametreleri kestirmede doğal logaritması alınmış denklem ile çalışmak daha kolay olduğundan Bölüm II'de de belirtildiği gibi olabilirlik denkleminin logaritması alınır. Logaritması alınmış denkleme ML yöntemi uygulanır. Buradan elde edilen denklem sistemi Gauss-Newton yöntemi aracılığıyla çözümlenir; her bir parametreye göre türev alınarak Newton denklemindeki F matrisi ve jakobiyen matrisi oluşturulur. Eğer jakobiyen matrisinin tersi singüler çıkıyorsa burada bir çoklu iç ilişkiden söz edilebilir. Bağımsız değişkenler arasında çoklu iç ilişkinin oluşmasının üç temel nedeni vardır. Deney tasarımındaki planlamadan ya da zayıf gözlemsel verilerden kaynaklanır, ikincisi ise bağımsız değişkenlerin kuvvetleri ya da çarpımları gibi matematiksel işlemler sonucu oluşturulan yeni değişkenlerin bulunduğu model yapısından kaynaklanır. Bir diğeri de bağımsız değişkenler üzerindeki kısıtların neden olduğu çoklu iç ilişkidir. Bu durumda SVD (Singuler Value Decompozition) ayrışımı uygulanır. Matlab programında yazılmış olan model sistemimiz aşağıda gösterilmektedir:

```
clc; clear
b0=-0.3; b1=0.3; b2=0.6; b3=0.5; b4=4; lamda1=2.5; lamda2=3.5; eps=0.01;
Jacob=zeros(7); ilkteta=zeros(7,1);
ilkteta(1)=b0; ilkteta(2)=b1; ilkteta(3)=b2; ilkteta(4)=b3; ilkteta(5)=b4;
ilkteta(6)=lamda1; ilkteta(7)=lamda2; enbsayi=1;

load veri_zuber.m
[n,k1]=size(veri_zuber)
y=veri_zuber(:,1); x=veri_zuber(:,2);


for iter=1:enbsayi

for i=1:n

   if x(i)>=2.5, es1(i)=1;
            else es1(i)=0;
   end

   if x(i)>=3.5, es2(i)=1;
            else es2(i)=0;
   end
end



   for i=1:n
       eus(i)=exp(b0+ b1*(b2-x(i))^3+ b3*(x(i)-lamda1)^2*es1(i)+ b4*(x(i)-lamda2)*es2(i));
       eust(i)=1+eus(i);
       epi(i)=eus(i)/eust(i);
       cpi(i)=epi(i)*(1-epi(i));   end


  toppi11=0; toppi12=0; toppi13=0; toppi14=0; toppi15=0; toppi16=0; toppi17=0;
```

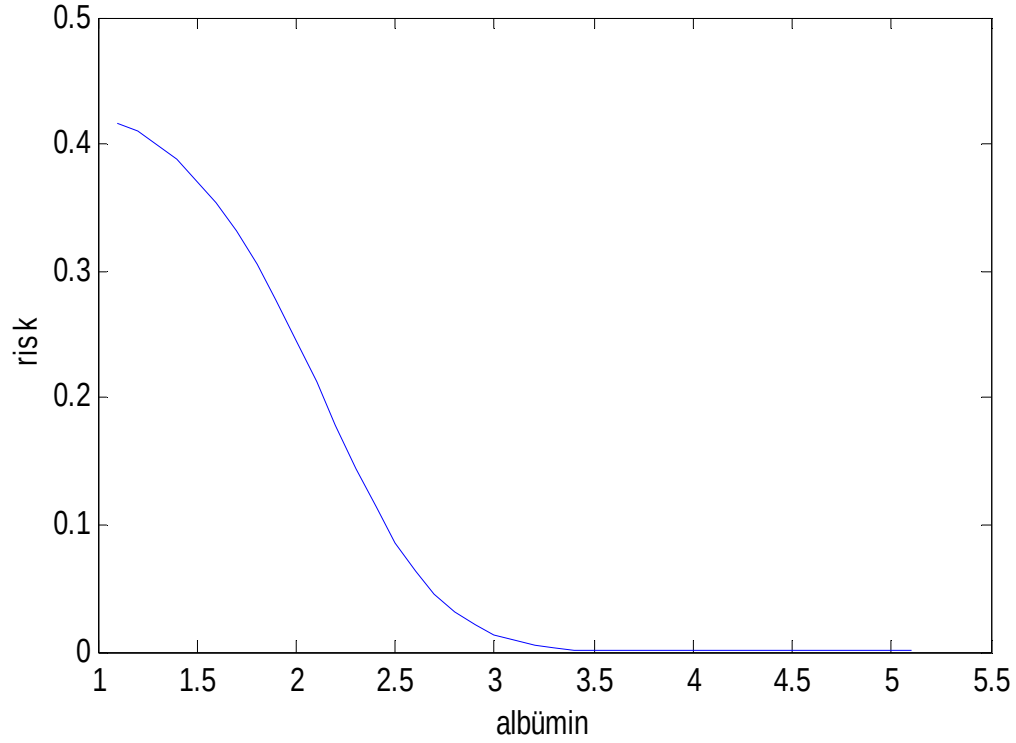
```
toppi21=0; toppi22=0; toppi23=0; toppi24=0; toppi25=0; toppi26=0; toppi27=0;
toppi31=0; toppi32=0; toppi33=0; toppi34=0; toppi35=0; toppi36=0; toppi37=0;
toppi41=0; toppi42=0; toppi43=0; toppi44=0; toppi45=0; toppi46=0; toppi47=0;
toppi51=0; toppi52=0; toppi53=0; toppi54=0; toppi55=0; toppi56=0; toppi57=0;
toppi61=0; toppi62=0; toppi63=0; toppi64=0; toppi65=0; toppi66=0; toppi67=0;
toppi71=0; toppi72=0; toppi73=0; toppi74=0; toppi75=0; toppi76=0; toppi77=0;

for i=1:n
toppi11=toppi11+ (-cpi(i));
toppi12=toppi12+ (-(b2-x(i))^3*cpi(i));
toppi13=toppi13+ (-3*b1*(b2-x(i))^2*cpi(i));
toppi14=toppi14+ (-(x(i)-lamda1)^2*es1(i)*cpi(i));
toppi15=toppi15+ (-(x(i)-lamda2)*es2(i)*cpi(i));
toppi16=toppi16+ (2*b3*(x(i)-lamda1)*es1(i)*cpi(i));
toppi17=toppi17+ (b4*es2(i)*cpi(i));

toppi21=toppi21+ (-(b2-x(i))^3*cpi(i));
toppi22=toppi22+ (-(b2-x(i))^6*cpi(i));
toppi23=toppi23+ (3*y(i)*(b2-x(i))^2- 3*(b2-x(i))^2*epi(i)- 3*b1*(b2-x(i))^5*cpi(i));
toppi24=toppi24+ (-(b2-x(i))^3*(x(i)-lamda1)^2*es1(i)*cpi(i));
toppi25=toppi25+ (-(b2-x(i))^3*(x(i)-lamda2)*es2(i)*cpi(i));
toppi26=toppi26+ (2*b3*(b2-x(i))^3*(x(i)-lamda1)*es1(i)*cpi(i));
toppi27=toppi27+ (b4*(b2-x(i))^3*es2(i)*cpi(i));
```

```
  toppi31=toppi31+ (-3*b1*(b2-x(i))^2*cpi(i));
  toppi32=toppi32+ (3*y(i)*(b2-x(i))^2- 3*(b2-x(i))^2*epi(i)- 3*b1*(b2-x(i))^5*cpi(i));
  toppi33=toppi33+ (6*y(i)*b1*(b2-x(i))- 6*b1*(b2-x(i))*epi(i)- 9*b1^2*(b2-x(i))^4*cpi(i));
  toppi34=toppi34+ (-3*b1*(b2-x(i))^2*(x(i)-lamda1)^2*es1(i)*cpi(i));
  toppi35=toppi35+ (-3*b1*(b2-x(i))^2*(x(i)-lamda2)*es2(i)*cpi(i));
  toppi36=toppi36+ (6*b1*b3*(b2-x(i))^2*(x(i)-lamda1)*es1(i)*cpi(i));
  toppi37=toppi37+ (3*b1*b4*(b2-x(i))^2*es2(i)*cpi(i));

  toppi41=toppi41+ (-(x(i)-lamda1)^2*es1(i)*cpi(i));
  toppi42=toppi42+ (-(b2-x(i))^3*(x(i)-lamda1)^2*es1(i)*cpi(i));
  toppi43=toppi43+ (-3*b1*(b2-x(i))^2*(x(i)-lamda1)^2*es1(i)*cpi(i));
  toppi44=toppi44+ (-(x(i)-lamda1)^4*es1(i)^2*cpi(i));
  toppi45=toppi45+ (-(x(i)-lamda1)^2*(x(i)-lamda2)*es1(i)*es2(i)*cpi(i));
  toppi46=toppi46+ (-2*y(i)*(x(i)-lamda1)*es1(i)+ 2*(x(i)-lamda1)*es1(i)*epi(i)+ 2*b3*(x(i)-
lamda1)^3*es1(i)^2*cpi(i));
  toppi47=toppi47+ (b4*(x(i)-lamda1)^2*es1(i)*es2(i)*cpi(i));

  toppi51=toppi51+ (-(x(i)-lamda2)*es2(i)*cpi(i));
  toppi52=toppi52+ (-(b2-x(i))^3*(x(i)-lamda2)*es2(i)*cpi(i));
  toppi53=toppi53+ (-3*b1*(b2-x(i))^2*(x(i)-lamda2)*es2(i)*cpi(i));
  toppi54=toppi54+ (-(x(i)-lamda1)^2*(x(i)-lamda2)*es1(i)*es2(i)*cpi(i));
  toppi55=toppi55+ (-(x(i)-lamda2)^2*es2(i)^2*cpi(i));
  toppi56=toppi56+ (2*b3*(x(i)-lamda1)*(x(i)-lamda2)*es1(i)*es2(i)*cpi(i));
```

```
  toppi57=toppi57+ (-y(i)*es2(i)+ es2(i)*epi(i)+ b4*(x(i)-lamda2)*es2(i)^2*cpi(i));

  toppi61=toppi61+ (2*b3*(x(i)-lamda1)*es1(i)*cpi(i));
  toppi62=toppi62+ (2*b3*(b2-x(i))^3*(x(i)-lamda1)*es1(i)*cpi(i));
  toppi63=toppi63+ (6*b1*b3*(b2-x(i))^2*(x(i)-lamda1)*es1(i)*cpi(i));
  toppi64=toppi64+ (-2*y(i)*(x(i)-lamda1)*es1(i)+ 2*(x(i)-lamda1)*es1(i)*epi(i)+ 2*b3*(x(i)-
lamda1)^3*es1(i)^2*cpi(i));
  toppi65=toppi65+ (-2*b3*(x(i)-lamda1)*(x(i)-lamda2)*es1(i)*es2(i)*cpi(i));
  toppi66=toppi66+ (2*y(i)*b3*es1(i)- 2*b3*es1(i)*epi(i)- 4*b3^2*(x(i)-
lamda1)^2*es1(i)^2*cpi(i));
  toppi67=toppi67+ (-2*b3*b4*(x(i)-lamda1)*es1(i)*es2(i)*cpi(i));

  toppi71=toppi71+ (b4*es2(i)*cpi(i));
  toppi72=toppi72+ (b4*(b2-x(i))^3*es2(i)*cpi(i));
  toppi73=toppi73+ (3*b1*b4*(b2-x(i))^2*es2(i)*cpi(i));
  toppi74=toppi74+ (b4*(x(i)-lamda1)^2*es1(i)*es2(i)*cpi(i));
  toppi75=toppi75+ (-y(i)*es2(i)+ es2(i)*epi(i)+ b4*(x(i)-lamda2)*es2(i)^2*cpi(i));
  toppi76=toppi76+ (-2*b3*b4*(x(i)-lamda1)*es1(i)*es2(i)*cpi(i));
  toppi77=toppi77+ (-b4^2*es2(i)^2*cpi(i));
  end
```

```
Jacob=[toppi11 toppi12 toppi13 toppi14 toppi15 toppi16 toppi17;
      toppi21 toppi22 toppi23 toppi24 toppi25 toppi26 toppi27;
      toppi31 toppi32 toppi33 toppi34 toppi35 toppi36 toppi37;
      toppi41 toppi42 toppi43 toppi44 toppi45 toppi46 toppi47;
      toppi51 toppi52 toppi53 toppi54 toppi55 toppi56 toppi57;
      toppi61 toppi62 toppi63 toppi64 toppi65 toppi66 toppi67;
      toppi71 toppi72 toppi73 toppi74 toppi75 toppi76 toppi77]

top1=0;
for i=1:n
    top1=top1+(y(i)-epi(i));end

top2=0;
for i=1:n
    top2=top2+(y(i)*(b2-x(i))^3- (b2-x(i))^3*epi(i));end

top3=0;
for i=1:n
    top3=top3+(3*y(i)*b1*(b2-x(i))^2- 3*b1*(b2-x(i))^2*epi(i));end

top4=0;
for i=1:n
    top4=top4+(y(i)*(x(i)-lamda1)^2*es1(i)- (x(i)-lamda1)^2*es1(i)*epi(i));end
```

```
  top5=0;
  for i=1:n
      top5=top5+(y(i)*(x(i)-lamda2)*es2(i)- (x(i)-lamda2)*es2(i)*epi(i));end

  top6=0;
  for i=1:n
      top6=top6+(-2*y(i)*b3*(x(i)-lamda1)*es1(i)+ 2*b3*(x(i)-lamda1)*es1(i)*epi(i));end

  top7=0;
  for i=1:n
      top7=top7+(-y(i)*b4*es2(i)+ b4*es2(i)*epi(i));end

F(1)=top1; F(2)=top2; F(3)=top3; F(4)=top4; F(5)=top5; F(6)=top6; F(7)=top7;
F
TersJacob=inv(Jacob)
teta=ilkteta+TersJacob*F'
fark=abs(teta-ilkteta);
maxfark=max(fark);
if maxfark<eps, break
end
ilkteta=teta;
b0=teta(1); b1=teta(2); b2=teta(3); b3=teta(4); b4=teta(5); lamda1=teta(6); lamda2=teta(7);
end
```



Şekil IV.1 Serum Albüminin Risk Grafiği

Şekil IV.1'den de görüleceği gibi serum albümin değerinin 2.5g/100mL'den düşük olduğu seviyelerde ölüm oranı riski %40'lara kadar ulaşmaktadır. Orta derecede risk olarak kabul edilen 2.5g/100mL ve 3.5g/100mL aralığında ise %14'lere kadar risk görülmekte olup bu aralıktan sonrasında gittikçe azalan risk durumu söz konusudur. Hatta serum albümin için epidemiyolojik olarak ideal aralık 3.5g/100mL'den sonrası olarak kabul görmektedir. Serum albümin için model yapısının Matlab programında çözümlenmesiyle ilgili sonuçlar aşağıda gösterilmektedir:

```
n =
   117

k1 =
     2
```

```
Jacob =

  Columns 1 through 3

   -6.1219   26.6493  -13.5509
   26.6493 -199.5820  301.2090
  -13.5509  301.2090  -93.7363
   -0.1253    1.9502   -0.6917
   -0.0012    0.0432   -0.0119
    0.3030   -3.8839    1.4757
    0.0524   -1.4190    0.4240

 Columns 4 through 6

   -0.1253   -0.0012    0.3030
    1.9502    0.0432   -3.8839
   -0.6917   -0.0119    1.4757
   -0.0507   -0.0023  -13.3077
   -0.0023   -0.0004    0.0017
  -13.3077   -0.0017    9.2965
    0.0641   -2.9819   -0.0574

  Column 7

    0.0524
   -1.4190
    0.4240
    0.0641
   -2.9819
   -0.0574
   -0.2095



F =

  Columns 1 through 3

   14.0470 -183.3585   63.9666

  Columns 4 through 6

    5.9325    0.3988   -6.6888

  Column 7

  -11.9476
```

```
TersJacob =

  Columns 1 through 3

   -0.2205   -0.0049    0.0163
   -0.0049    0.0012    0.0045
    0.0163    0.0045    0.0016
   -0.0014    0.0000   -0.0005
    0.0007   -0.0000   -0.0017
    0.0005   -0.0000    0.0004
   -0.0000    0.0000    0.0001

  Columns 4 through 6

   -0.0014    0.0007    0.0005
    0.0000   -0.0000   -0.0000
   -0.0005   -0.0017    0.0004
   -0.0524    0.0002   -0.0749
    0.0002    0.0234   -0.0015
   -0.0749   -0.0015    0.0003
    0.0000   -0.3354    0.0001

  Column 7

   -0.0000
    0.0000
    0.0001
    0.0001
   -0.3354
    0.0001
    0.0000



teta =

   -1.4760
    0.3035
    0.0903
    0.6323
    7.9319
    2.0905
    3.3680
```

Dikkat edilirse iterasyon sayısı fazla ilerlememektedir. Doğrusal olmayan sistem çözümlerinde Gauss-Newton yönteminde genelde görülen bir durumdur. Greg Ridgeway makalesinde lojistik regresyonun doğrusal olmaması durumunda parametre kestiriminde iterasyonun fazla gitmeyeceğini belirtmiştir. Bu nedenle doğrusal olmayan yapılarda bu durumun görülmesi muhtemeldir [21].

Serum albümin için kurulan modeldeki parametreler b_0=-1.4760 , b_1=0.3035 , b_2=0.0903 , b_3=0.6323 , b_4=7.9319 , λ_1=2.0905 , λ_2=3.3680 olarak elde edilmiştir. Başlangıç değerlerine oldukça yakın olarak bulunan λ değerleri doğru bir modele yaklaştığımızın da bir kanıtı olarak gösterilebilir. Doğrusal olmayan bu lojistik regresyon modeli, en az iterasyonla en iyi modeli yakalayabilmiştir. Şekil IV.1'den de görüleceği gibi, tıbbi olarak serum albümin için geçerli değerler kurulan modelden de geçerliliğini korumaktadır.

BÖLÜM V

SONUÇLAR VE TARTIŞMA

Uygulama-1 çalışmasında model_2 gerek en çok olabilirlik kestirimi gerekse de odds oran kestirimi bakımından daha iyi sonuç vermiştir. Yalnız burada söz konusu olan bir şey vardır ki o da parametre kestirimlerinin aralıklarıdır. Model_2'nin parametre kestirimleri ve güven aralıklar hgbz1 hariç daha idealdir.

PROC NLIN gibi SAS paket programı kullanıldığında karşılaşılan problem ele alınan parametreler için her zaman alt ve üst sınırların elde edilmemesi şeklinde olup yakınsama ile ilgilidir. Hataların normal dağılıma sahip olduğu doğrusal olmayan regresyon analizlerinde Gauss-Newton, Marquaart ve Newton yöntemleri çok rahat elde edilmekte ancak doğrusal olmayan lojistik regresyon analizinde hataların binom dağılımına sahip olmasından dolayı sorunla karşılaşılmaktadır. Pek çok durumda istikrarlı çözümler sağlamasa bile, doğrusal olmayan lojistik regresyon analizi kullanılarak yapılan modellemeler parametrelerin daha performanslı kestirimlerini sağlar [9].

Hemoglobin çalışmasında kabul edilebilirliği sınayıcı bir başka istatistik ise c (concordance ratio) değeridir. Bu değer 0 ile 1 arasında bir değer alır ve ROC eğrisinin altında kalan alana eşdeğer olduğundan yaptığımız uygulamada 0.955 ile oldukça iyi sonuç vermiştir.

Uygulama-2 Matlab programında yapılan serum albümin çalışması, splayn yöntemin lojistik regresyonda kullanılmasına dair iyi bir örnek olup parametre kestirimleri de anlamlıdır. Mevcut şekilden de desteklenerek epidemiyolojik olarak kabul edilebilir bir model kurulmuştur. Yapılan iterasyonla önemli değişim noktası olarak gösterilen 2.5g/100mL ve 3.5g/100mL değerlerine yaklaşık bir kestirim elde edilmiştir. Bu uygulama göstermiştir ki, serum albümin değerinin 2.5g/100mL'den düşük çıkması ölüm riskini %40 seviyelerine kadar arttırmaktadır. İdeal seviye olan

3.5g/100mL'ye yaklaştıkça risk azalmaktadır. Sonuç itibariyle doğrusal olmayan lojistik regresyonda model kestirimleri ve yorumların anlamlılığı daha iyi sonuç vermektedir ve programın geliştirilmesi ile daha ayrıntılı ve geniş çaplı bir kullanımın ortaya çıkacağı düşünülmektedir.

KAYNAKLAR

[1] Aydın, D.: "*Semiparametrik Regresyon Modellemede Splayn Düzeltme Yaklaşımı İle Tahmin ve Çıkarsamalar",* Doktora Tezi, Anadolu Üniversitesi Fen Bilimleri Enstitüsü İstatistik, Eskişehir, Türkiye, **(2005)**.

[2] Başarır, G.: "*Çok Değişkenli Verilerde Ayırsama Sorunu ve Lojistik Regresyon Analizi",* Doktora Tezi, Hacettepe Üniversitesi Fen Bilimleri Enstitüsü Uygulamalı İstatistik, Ankara, Türkiye, **(1990)**.

[3] Bircan, H.: "Lojistik Regresyon Analizi ve Tıp Verileri Üzerine Bir Uygulama", *Kocaeli Üniversitesi Sosyal Bilimler Dergisi,* **(2004)**.

[4] Boucher, K. M., Slattery, M. L., Berry, T. D., et al.: "*Statistical Methods to Analyze Dose-Response and Trend Analysis in Epidemiologic Studies*", Journal of Clinical Epidemiology, Vol 51, No 12, **(1998)**.

[5] Dobson, Annette J.: *An Introduction to Generalized Linear Models,* Chapmen Hall, London, **(1990)**.

[6] Erdoğan Eygi, B.: "*Bankaların Mali Performanslarının Lojistik Regresyon ile Analizi ve İleriye Yönelik Tahmin",* Doktora Tezi, Marmara Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul, Türkiye, **(2002)**.

[7] Gardside, P. S.; Glueck, C. J.: *The Important role of Modifiable Dietary and Behaviour Characteristic in the Causation and Prevention of Coronary Heart Disease Hospitalization and Mortality,* Journal of American College of Nutrition, **(1995)**.

[8] Hosmer, David W.; Lemeshow, S.: *Applied Logistic Regression,* Second Edition, John Wiley, New York, **(2000)**.

[9] İyit, N.: "*Lineer Olmayan Lojistik Regresyon Analizinde Model Kurma Stratejileri ve Bir Uygulaması",* Yüksek Lisans Tezi, Selçuk Üniversitesi Fen Bilimleri Enstitüsü İstatistik, Konya, Türkiye, **(2003)**.

[10] Kalaycı, Ş.: *SPSS Uygulamalı Çok Değişkenli İstatistik Teknikleri,* Asil Yayın, Ankara, Türkiye, **(2006)**.

[11] Kleinbaum, David G.: *Logistic Regression, A Self-Learning Text,* Springer-Verlag, New York, **(1994)**.

[12] Lesaffre, E.; Albert, A.: *A Multiple Group Logistic Regression Diagnostics,* Applied Statistics, **(1989)**.

[13] Liu, Q.; Lambe, M.; Baik, I.; et al.: "*A Prospective Study of the Transient Decrease in Ovarian Cancer Risk Following Childbirth",* Cancer Epidemiology Biomarkers & Prevention , Vol. 15, 2508-2513, **(2006)**.

[14] McCullagh, P.; Nelder, J. A.: *Generalized Lineer Models,* Chapmen Hall, London, **(1989)**.

[15] Menard, S.: *Applied Logistic Regression Analysis,* Sage Publication, **(1995)**.

[16] Montgomery, Doughlas C.; Peck, Elizabeth A.; Vining, G. Geoffrey: *Introduction to Linear Regression Analysis,* Third Edition, Wiley & Sons, New York.

[17] Mulla, Z.: "*Spline Regression in Clinical Research*", The West Indian Medical Journal, Vol 56 (1), **(2007)**.

[18] Özdamar, K.: *Paket Programlar ile İstatistiksel Veri Analizi,* Kaan Kitabevi, Eskişehir, Türkiye, (**2004)**.

[19] Pastor, R.; Guallar, E.: "*Use of Two-segmented Logistic Regression to Estimate Change-points in Epidemiologic Studies",* American Journal of Epidemiology , Vol. 148, No. 7, 631-642, **(1998)**.

[20] Rice, John R.: *Numerical Methods, Software, and Analysis,* McGraw Hill Higher Education, **(1983)**.

[21] Ridgeway, G.: "*Discussion. Additive Logistic Regression: A Statistical View of Boosting*", The Annals of Statistics, Vol 28, No 2, **(2000)**.

[22] Schimek, Michael G.: *Smoothing and Regression (Approaches, Computation and Application),* John Wiley & Sons, New York, **(1999)**.

[23] Tatlıdil, H.: *Uygulamalı Çok Değişkenli İstatistiksel Analiz,* H.Ü. Fen Fakültesi İstatistik Bölümü, Ankara, Türkiye, **(1992)**.

[24] http://www.hekimce.com/index.php?klm=hemoglobin (04/2009)

EK-1 Hemoglobin SAS Sonucu

x1	risk	lclper	uclper
5.5	100.000	0.0000	100.000
5.6	100.000	0.0000	100.000
5.7	100.000	0.0000	100.000
5.8	99.999	0.0000	100.000
5.9	99.999	0.0000	100.000
6.0	99.998	0.0000	100.000
6.1	99.996	0.0000	100.000
6.2	99.992	0.0000	100.000
6.3	99.986	0.0000	100.000
6.4	99.976	0.0000	100.000
6.5	99.959	0.0001	100.000
6.6	99.932	0.0005	100.000
6.7	99.888	0.0024	100.000
6.8	99.821	0.0102	100.000
6.9	99.721	0.0395	100.000
7.0	99.574	0.1391	100.000
7.1	99.365	0.4442	100.000
7.2	99.075	1.2812	100.000
7.3	98.684	3.3074	99.999
7.4	98.173	7.5307	99.997
7.5	97.525	14.8545	99.989
7.6	96.729	25.0592	99.962
7.7	95.780	36.2770	99.890
7.8	94.688	46.0559	99.732
7.9	93.474	52.8502	99.457
8.0	92.172	56.3793	99.076
8.1	90.829	57.2257	98.655
8.2	89.500	56.4495	98.247
8.3	88.243	55.2720	97.854
8.4	87.117	54.7698	97.420
8.5	86.174	55.6887	96.866
8.6	85.461	58.3762	96.100
8.7	84.768	61.4092	95.113
8.8	83.843	63.3881	93.959
8.9	82.656	64.3487	92.638
9.0	81.170	64.3220	91.155
9.1	79.340	63.3364	89.514
9.2	77.114	61.4307	87.697
9.3	74.437	58.6701	85.659
9.4	71.252	55.1530	83.320
9.5	67.509	51.0079	80.569
9.6	63.175	46.3830	77.284
9.7	58.248	41.4334	73.342
9.8	52.773	36.3095	68.654
9.9	46.851	31.1487	63.202
10.0	40.649	26.0729	57.081
10.1	34.388	21.1943	50.528
10.2	28.316	16.6284	43.894
10.3	22.674	12.5050	37.563
10.4	17.654	8.9585	31.838
10.5	13.373	6.0903	26.874
10.6	9.869	3.9260	22.683
10.7	7.105	2.4036	19.193

10.8	4.999	1.4016	16.303
10.9	3.443	0.7809	13.907
11.0	2.324	0.4168	11.917
11.1	1.540	0.2136	10.259
11.2	1.003	0.1052	8.872
11.3	0.642	0.0499	7.708
11.4	0.404	0.0228	6.727
11.5	0.250	0.0100	5.898
11.6	0.153	0.0043	5.195
11.7	0.092	0.0017	4.597
11.8	0.054	0.0007	4.086
11.9	0.032	0.0003	3.650
12.0	0.018	0.0001	3.275
12.1	0.008	0.0000	3.405
12.2	0.004	0.0000	3.877
12.3	0.002	0.0000	4.838
12.4	0.001	0.0000	6.598
12.5	0.000	0.0000	9.755
12.6	0.000	0.0000	15.396
12.7	0.000	0.0000	25.187
12.8	0.000	0.0000	40.604
12.9	0.000	0.0000	60.351
13.0	0.000	0.0000	78.781
13.1	0.000	0.0000	90.838
13.2	0.000	0.0000	96.662
13.3	0.000	0.0000	98.930
13.4	0.000	0.0000	99.691
13.5	0.000	0.0000	99.919
13.6	0.000	0.0000	99.980
13.7	0.000	0.0000	99.996
13.8	0.000	0.0000	99.999
13.9	0.000	0.0000	100.000
14.0	0.000	0.0000	100.000
14.1	0.000	0.0000	100.000

EK-2 Albümin Matlab Sonucu

x	epi
1.1000	0.1433
1.2000	0.1312
1.4000	0.1036
1.6000	0.0744
1.7000	0.0605
1.7000	0.0605
1.8000	0.0478
1.8000	0.0478
1.9000	0.0364
1.9000	0.0364
1.9000	0.0364
2.0000	0.0269
2.0000	0.0269
2.0000	0.0269
2.1000	0.0191
2.1000	0.0191
2.2000	0.0131
2.2000	0.0131
2.3000	0.0086
2.3000	0.0086
2.3000	0.0086
2.3000	0.0086
2.3000	0.0086
2.4000	0.0054
2.4000	0.0054
2.4000	0.0054
2.4000	0.0054
2.5000	0.0036
2.5000	0.0036
2.5000	0.0036
2.5000	0.0036

2.5000 0.0036
2.5000 0.0036
2.6000 0.0022
2.6000 0.0022
2.6000 0.0022
2.6000 0.0022
2.7000 0.0013
2.7000 0.0013
2.7000 0.0013
2.7000 0.0013
2.8000 0.0007
2.8000 0.0007
2.8000 0.0007
2.8000 0.0007
2.8000 0.0007
2.8000 0.0007
2.8000 0.0007
2.8000 0.0007
2.8000 0.0007
2.8000 0.0007
2.8000 0.0007
2.8000 0.0007
2.9000 0.0004
2.9000 0.0004
2.9000 0.0004
2.9000 0.0004
2.9000 0.0004
2.9000 0.0004
3.0000 0.0002
3.0000 0.0002
3.0000 0.0002
3.1000 0.0001
3.1000 0.0001
3.1000 0.0001

3.1000 0.0001
3.1000 0.0001
3.1000 0.0001
3.2000 0.0001
3.2000 0.0001
3.3000 0.0000
3.3000 0.0000
3.3000 0.0000
3.3000 0.0000
3.3000 0.0000
3.3000 0.0000
3.4000 0.0000
3.4000 0.0000
3.4000 0.0000
3.4000 0.0000
3.4000 0.0000
3.5000 0.0000
3.5000 0.0000
3.5000 0.0000
3.5000 0.0000
3.5000 0.0000
3.5000 0.0000
3.5000 0.0000
3.5000 0.0000
3.5000 0.0000
3.6000 0.0000
3.6000 0.0000
3.6000 0.0000
3.6000 0.0000
3.6000 0.0000
3.7000 0.0000
3.7000 0.0000
3.8000 0.0000
3.8000 0.0000

3.8000 0.0000
3.9000 0.0000
3.9000 0.0000
4.0000 0.0000
4.1000 0.0000
4.1000 0.0000
4.1000 0.0000
4.2000 0.0000
4.2000 0.0000
4.3000 0.0000
4.4000 0.0000
4.4000 0.0000
4.4000 0.0000
4.4000 0.0000
4.7000 0.0000
4.7000 0.0000
4.8000 0.0000
5.1000 0.0000

ÖZGEÇMİŞ

1984 yılı Üsküdar doğumlu Esra Zeynep ŞENSOY, 1990 yılında İstanbul'da başladığı eğitim hayatını ortaöğretim eğitimi bitimine kadar bu şehirde başarılı bir şekilde sürdürmüş olup, 2002 yılında yabancı dil ağırlıklı lisede okul ikinciliği elde etmiştir. Aynı sene Denizli'de Pamukkale Üniversitesi Fen-Edebiyat Fakültesi Matematik Bölümü'nde okumaya hak kazanarak bu üniversitede görmüş olduğu bir yıllık eğitim hayatından sonra bölüm birincisi olmuş ve Bursa Uludağ Üniversitesi Fen-Edebiyat Fakültesi Matematik Bölümü'ne yatay geçiş yapmıştır. Uludağ Üniversitesi'nde tamamladığı lisans eğitiminden sonra, Marmara Üniversitesi Fen Bilimleri Enstitüsü Uygulamalı Matematik alanında tezli yüksek lisans programına kayıt yaptırmıştır.