12596

ÇOK DEĞİŞKENLİ VERİLERDE AYRIMSAMA SORUNU VE LOJİSTİK REGRESYON ANALİZİ

Gülay Başarır

Hacettepe Üniversitesi
Fen Bilimleri Enstitüsü Yönetmeliğinin
İstatistik Anabilim Dalı İçin Öngördüğü
DOKTORA TEZİ
olarak hazırlanmıştır.

Kasım-1990

Fen Bilimleri Enstitüsü Müdürlüğü'ne,

İşbu çalışma, jürimiz tarafından, İSTATİSTİK Anabilim Dalında DOKTORA TEZİ olarak kabul edilmiştir.

Başkan : Prof. Dr. Zehra MULUK

Üye : Prof. Dr. Soner Gönen

Üye : Doç. Dr. Taslan Ula

Üye : Doç. Dr. M. Aydın ERAR

Üye : Doç. Dr. Hüseyin TATLIDİL

ONAY

Yukarıdaki imzaların, adı geçen öğretim üyelerine ait olduğunu onaylarım.

/ /

Prof. Dr. Dinçer Ülkü
Fen Bilimleri Enstitüsü Müdürü

Biricik Annem'e ve
Babam'a

ÖZET

Bu çalışmanın amacı, kesikli değişkenlerin çok sık rastlandığı anket türü verilerde, diskriminant analizine alternatif olarak önerilen lojistik regresyon analizini uygulamak ve gözlemleri verilerin yapısında bulunan gruplardan birine atayacak en iyi modele ulaşmaktır.

Birinci Bölümde, çok değişkenli verilerin analizinde ayrımsama sorununa değinilerek çalışmanın amacı belirtildi.

İkinci Bölümde, ikili ve çoklu grup lojistik regresyon analizi, kestirim yöntemleri, artık incelemesi ve diskriminant-lojistik regresyon analizi karşılaştırmalarına ilişkin bir derlemenin sunulduğu genel bilgiler verildi.

Üçüncü Bölümde, amaca yönelik olarak, lojistik regresyon analizinin kardiyolojik ve Öğrenci Seçme Sınavına ilişkin verilere uygulaması yapılarak bulguları verildi. Kardiyolojik verilere uydurulan çoklu grup lojistik modelin, amaç kestirim olduğunda ikili modellere tercih edileceği, ayrıca diskriminant fonksiyonundan daha iyi ayrımsama verdiği görüldü. ÖSS verilerinin analizinde lojistik modelin uygun olmadığı, model yetersizliğinin verilerden kaynaklandığı vurgulandı.

Son Bölümde, çoklu grup lojistik modellerin Begg ve Gray yaklaşımı ile elde edilen ikili lojistik modellere göre daha iyi ayrımsama vermemekle birlikte, amaç kestirim olduğunda tercih edilmesi gerektiği, ayrıca diskriminant analizinin kesikli değişkenlerden kaynaklanan varsayım bozulumu nedeni ile yeterli olmadığı belirtildi. Lojistik regresyon analizinin somut ölçümlere dayalı verilerde iyi sonuç verebildiği, somut ölçümlerin elde edilemediği sosyal uygulamalarda yetersiz kalabildiği sonucuna ulaşıldı.

SUMMARY

In this study, the purpose is, in the case of survey sampling, to find the best model that will allocate observations to the groups existing in the data. Since survey sampling data generally contain discrete variables, the normality assumption will not be satisfied and logistic regression analysis is used as an alternative to discriminant analysis.

In the First Section, discrimination problem in analysing the multivariate data and the purpose of the study have been pointed out.

In the Second Section, dichotomous and polytomous logistic regression analysis, estimation techniques, residual diagnostics and comparison of logistic and discriminant analysis have been examined.

In the Third Section, two applications of logistic regression analysis, to cardiologic and to student election exam data have been examined and results have been summarized. In the first application, it has been seen that, multiple group logistic model with main effects which has been fitted to cardiologic data is preferred to two group logistic models for the purpose of estimation and gives better discrimination than discriminant function. In the second application, it has been mentioned that, logistic model is not suitable for student election exam data and this insufficiency is due to the data set.

In the last Section, depending on the results, it has been pointed out that, despite multiple group logistic models don't have better discrimination power, they should be preferred to Begg and Gray's individualized logistic models approximation for the purpose of estimation. On the other hand it has been seen that since the assumptions are not satisfied because of discrete variables, discriminant analysis doesn't have sufficient discrimination power.

Lastly it has been emphasized that logistic regression analysis gives good results for data sets which depends on concrete measurements while it is insufficient for social applications.

TEŞEKKÜR

Bu tez konusuna beni yönlendiren ve çalışmam boyunca değerli katkı ve eleştirileri ile bana güç veren çok sevgili danışmanım Sayın Doç. Dr. Hüseyin Tatlıdil'e, medikal uygulamanın her aşamasında değerli yardımlarını esirgemeyen Gülhane Askeri Tıp Akademisi Kardiyoloji Bölümü uzmanlık öğrencisi Sayın Dr. Cihangir Uyan'a, bilgisayar çalışmalarımda bana değerli zamanlarını ayırarak, destek olan Bilgisayar Mühendisliği Bölümü öğretim elemanlarından Sayın Yrd. Doç. Dr. İbrahim Sinir'e, çalışmama yapıcı uyarılarda bulunan Sayın Doç. Dr. Aydın Erar'a, Bölüm Başkanım Sayın Prof. Dr. Ceyhan İnal'a, Tez yazımında değerli katkılarından dolayı Sayın Rafiye Güngör'e ve çalışmam boyunca gösterdikleri sabır, anlayış ve desteklerinden dolayı tüm bölüm arkadaşlarıma ve değerli aileme sonsuz teşekkürlerimi sunarım.

İÇİNDEKİLER DİZİNİ

İÇİNDEKİLER DİZİNİ (Devam ediyor)

Sayfa

İÇİNDEKİLER DİZİNİ (Devam ediyor)

ŞEKİLLER DİZİNİ

ÇİZELGELER DİZİNİ

ÇİZELGELER DİZİNİ (Devam ediyor)

1. GİRİŞ

Çok değişkenli istatistiksel verilerin sınıflandırılması, bu verilere uygulanabilecek çeşitli istatistiksel yöntemler için gerekli ve yararlı bilgiler verecektir. Gözlemleri verilerin yapısında bulunan olası gruplara atamak için kullanılan yöntemlerden üç tanesi, kümeleme (clustering), diskriminant (discriminant) ve lojistik regresyon (logistic regression) analizidir.
Kümeleme analizinde verilerin yapısındaki grup (küme) sayısı bilinmemekte, gözlemler uzaklık (distance) ya da benzerlik (similarity) ölçütlerine göre kümelenmektedir. Burada amaç yalnızca gözlemlerin oluşturduğu küme yapısını bulmaktır. Diskriminant ve lojistik regresyon analizinde ise verilerin yapısındaki grup sayısı bilinmekte ve bu verilerden yararlanarak bir ayrımsama modeli elde edilmektedir. Kurulan bu model yardımı ile veri kümesine yeni alınan gözlemlerin gruplara atamaları yapılmaktadır.

Bu üç yöntemden kümeleme ve diskriminant analizi şimdiye dek çok geniş olarak incelenmiş yöntemlerdir. Logistik regresyon ise daha çok son yıllarda yoğun bir şekilde kullanılmaya başlanmıştır. Yöntem genelde çeşitli varsayım bozulumları durumunda diskriminant analizi ve çapraz tablolara bir alternatif olarak uygulanmaktadır. Kullanım nedeni olarak en temel yaklaşım doğrusal regresyon analizinden yapılabilir: Bağımlı değişken 0,1 gibi iki (binary) ya da ikiden çok düzey içeren kesikli (polytomous) değişken olduğunda normallik varsayımı bozulmakta ve doğrusal regresyon analizi uygulanamamaktadır. Böyle durumlarda lojistik regresyon analizi önerilmektedir.

Öte yandan, değişkenlerin normal dağıldığı ve grupların ortak varyans-kovaryans matrisine sahip olduğu varsayımlarına dayanan diskriminant analizi yönteminde bazı değişkenlerin kesikli olması, yöntemin yanlı kestirim vermesine neden olacaktır. Bu durumda da lojistik regresyon analizi daha etkin ve güçlü (robust) kestirimler vermekte, gözlemlerin birbirlerinin etkisini maskeleme durumlarını beklenmeyen kestirim değerleri vererek ortaya çıkartabilmektedir. Böylece alternatif bir yöntem olarak gelişen ve genelleştirilmiş lineer modeller (GLM) ailesinin özel bir durumu olan lojistik modeli araştıran lojistik regresyon analizi son yıllarda medikal uygulamalarda, özellikle epidemiolojide, ekonomide çok sık kullanılmaktadır. Çözümlemede elde edilen modelin matematiksel olarak çok esnek ve kolay yorumlanabilir olması ve anlamlı yorumlara götürülebilmesi, yönteme olan eğilimin artmasında önemli faktörlerdir.

Bu ön bilgiler ışığında çalışmanın amacı, uygulamada çok sık karşılaşılan anket türü verilerin değerlendirilmesi sırasında, verilerin yapısında var olan gruplara gözlemleri en doğru şekilde atayacak modeli bulmaktır. Bu tür verilerin genelde kesikli değişken içermesi ve normallik varsayımından söz edilememesi nedeniyle çözümlemede lojistik regresyon analizi kullanılmaktadır.

Çalışmada,lojistik regresyon analizi kardiyolojik ve öğrenci seçme sınavına (ÖSS) ilişkin verilere uygulanarak yöntemin bu veriler üzerindeki performansı incelenecektir. İlk uygulamada çeşitli türde kalp hastalığı olan hastalara ilişkin bilgiler yardımı ile en iyi lojistik model bulunmaya çalışılacak, bu model yardımı ile hastaların kalp hastalık gruplarından hangisine gireceği incelenerek hekimlerin tanılarının doğruluğu sınanacaktır. Ayrıca mümkün olduğunca az bilgi yardımı ile doğru tanı koyma yolları araştırılacak, ikili ve çoklu grup lojistik regresyon analiz yöntemleri karşılaştırılacaktır. Bunun yanı sıra diskriminant analiz yöntemi de uygulanarak lojistik regresyon analizi sonuçları ile kıyaslaması yapılacaktır.

İkinci uygulamada üniversite sınavına giren adayların çeşitli bilgileri kullanılarak kaçıncı tercihlerini kazanabileceklerini tahmin etmek amacı ile lojistik modelden yararlanılacaktır. Bu tahmini mümkün olduğunca az bilgi kullanarak aynı doğrulukta yapmanın yolları da araştırılacaktır. Bu amaçlara yönelik olarak çalışmanın izleyen ikinci bölümünde lojistik regresyon analizine ilişkin literatür taramasından elde edilen genel bilgiler özetlenmektedir. Bu bölümde ikili ve çoklu grup lojistik regresyon analizi ele alınmakta, kestirim yöntemleri, uyum iyiliği ölçütleri ve değişken seçim yöntemlerine değinilmektedir. Artık (residual) incelemesinin yanı sıra lojistik regresyon ve diskriminant analizi karşılaştırması da konu edilmektedir.

Üçüncü bölümde Gülhane Askeri Tıp Akademisi Hastanesi Kardiyoloji Bölümünde yatan kalp hastalarının kayıtlarından alınan veriler ile geniş bir anket çalışması sonucu elde edilen öğrenci seçme sınavı verilerine lojistik model uygulaması detaylı bir şekilde incelenmektedir.

Son bölümde elde edilen sonuçlar yorumlanmakta, daha sonraki araştırmalarda üzerinde durulması gereken henüz açıklığa kavuşturulmamış bazı konuların yanı sıra çalışmada ele alınmayan noktalar ve nedenleri tartışılmaktadır.

2. GENEL BİLGİLER

2.1. Genel Doğrusal Regresyon ve Lojistik Modelleri

Genel doğrusal regresyon modeli*,

$$Y=X\beta+\varepsilon \tag{2.1}$$

şeklinde verilmekte olup modelde n, gözlem sayısı; p, açıklayıcı değişken sayısı olmak üzere, Y, 1xn boyutlu yanıt (bağımlı değişken) vektörü, X, nx(p+1) boyutlu sabit açıklayıcı değişkenler matrisi; ß (p+1)x1 boyutlu bilinmeyen parametreler vektörüdür; ε ise 1xn boyutlu hata vektörüdür. Yanıt değişkeni Y'nin sürekli olduğu ve başta artıklar üzerinde normallik olmak üzere çeşitli varsayımların yapıldığı regresyon modelinde, yanıt değişkeninin kesikli olması durumunda -ikili ya da ikiden çok düzeyli- varsayımlar bozulmakta ve hata teriminin binom (multinomial) dağılım göstermesi nedeni ile gözlem varyansları eşit olmamaktadır. Bu durumda, veriler bilinen regresyon analizi ile incelenemediğinden lojistik regresyon analizi önerilmektedir.

Çok değişkenli istatistiksel analiz yöntemlerinden diskriminant analizi, değişkenlerin en az bir tanesi kesikli olduğunda varsayım bozulumu nedeni ile uygulanamamakta, bu durumda açıklayıcı değişkenlerin kesikli ve sürekli olmalarına ilişkin hiçbir kısıtlama getirmeyen, diskriminant analizi gibi gözlemlerin geleceğe yönelik atanması amacı ile kullanılabilen lojistik regresyon analizi alternatif olarak önerilmektedir (Press and Wilson, 1978).

Öte yandan, çapraz çizelgelerle incelenen kesikli değişkenlerin düzey sayıları arttıkça değişken kesikliden çok sürekli hale gelmekte ve bu durumda elde edilen çözüm iyi olmamaktadır. Ayrıca örneklem genişliği sabit tutu-

---

* Çalışmada aksi belirtilmedikçe matris ve vektörler büyük harflerle, sayısal değerler ise küçük harflerle verilmektedir.

lurken değişkenin düzey sayısının artması halinde çapraz tabloların bazı gözelerinde hiçbir değer gözlenemeyebilmekte ya da tek bir gözlem bulunabilmektedir. Bu durumda elde edilen model beklenen sonuçlara götürmeyebilecektir. Bunun yanı sıra çok kez tekrarlanan deneylerde bile gözlemlerin bazı gözelere düşme olasılığı sıfır olabilmektedir (structural zero). Burada etkilerin bağımsızlık testi bu gözeler dışındaki gözelere uygulanmakta ve test tablonun bir alt kümesi ile sınırlanmakta, bu sınırlama beklenmeyen sonuçlar yaratabilmektedir. Tüm bu kısıtlar durumunda çapraz tablolara alternatif bir yöntem olarak yine lojistik regresyon analizi önerilmektedir. Analizin güncel olmasının nedenleri şöyle özetlenebilmektedir:

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

ii) Lojistik modelin parametreleri epidemioloji'de yapılan ölçümlere benzediği için yorumları kolay olmaktadır. Epidemiolojide katsayıların exponansiyeli hastalık riski olarak yorumlanmaktadır (Hosmer and Lemeshow, 1989b).

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

iv) Lojistik modele dayalı analizler için standart paket programlar vardır.

v) Açıklayıcı değişkenlerin olasılık fonksiyonlarının dağılımı üzerinde kısıt olmaması (yarı parametrik) nedeni ile çeşitli testler uygulanabilmektedir (Anderson, 1982).

Epidemioloji ve diğer medikal uygulamalarının yanı sıra deneysel verilerin analizinde, askeri konularda, meteorolojide v.b. alanlarda sıkça kullanılan lojistik regresyon analizi farklı varsayımlar durumunda aynı lojistik formülasyona götürdüğü için varsayım bozumlarına karşı daha güçlü bir yöntemdir.

2.2. Doğrusal Olasılık Modeli

(2.1) formülasyonu ile verilen genel doğrusal regresyon modeli her gözlem için i=1,...,n olmak üzere,

$$E(y_i/x_{i1},...,x_{ip})=\sum_{k=0}^{p} \beta_k x_{ik} \tag{2.2}$$

şeklinde koşullu beklenen değer olarak da tanımlanabilir. Bu modelde açıklayıcı değişkenler üzerinde kısıt yok iken Y yanıt değişkeninin süreklilik koşulu vardır. (2.1)'de verilen model her i gözlemi için,

$$y_i=\sum_{k=0}^{p} \beta_k x_{ik}+u_i \tag{2.3}$$

olarak ifade edilebilir. Modelde açıklayıcı değişkenler üzerinde bir kısıt olmadığından y_i yanıt değerleri (-∞, +∞) arasındadır.

Yanıt değişkeninin ikili değerler (0,1) aldığı durumda bu yaklaşım bozulacak ve $P(y_i=1)$, i'nci gözlemin 1 değerini alma olasılığı olmak üzere, beklenen değer,

$$\begin{aligned} E(y_i)&=1\times P(y_i=1)+0\times P(y_i=0) \\ &=P(y_i=1) \end{aligned} \tag{2.4}$$

olacaktır. Bu sonuç regresyon denklemi olarak yazıldığında,

$$E(y_i)=P(y_i=1)=\sum_{k=0}^{p} \beta_k x_{ik} \tag{2.5}$$

elde edilir. Denklemin sağ tarafı olasılık olarak yorumlanmalı ve 0-1 arasında değerler almalıdır. Bu nedenle y_i yanıt değerlerinin ikili olduğu regresyon modeline "doğrusal olasılık modeli (linear probability model)" adı verilmektedir.

(2.3) denkleminde y_i'nin iki ayrı değeri için u_i hata terimi i=1,...,n için, $y_i=0$ ve $y_i=1$ için sıra ile,

$$\sum_{k=0}^{p} \beta_k x_{ik}+u_i=0 \Rightarrow u_i=-\sum_{k=0}^{p} \beta_k x_{ik}$$

$$\sum_{k=0}^{p} \beta_k x_{ik}+u_i=1 \Rightarrow u_i=1-\sum_{k=0}^{p} \beta_k x_{ik} \qquad (2.6)$$

şeklinde ifade edilmektedir. Hata terimleri üzerindeki varsayımlardan beklenen değeri,

$$E(u_i)=P(y_i=0)(-\sum_{k=0}^{p} \beta_k x_{ik})+P(y_i=1)(1-\sum_{k=0}^{p} \beta_k x_{ik})$$

$$=0 \qquad (2.7)$$

ve varyansı,

$$V(u_i)=E(u_i^2)=P(y_i=0)(-\sum_{k=0}^{p} \beta_k x_{ik})^2+P(y_i=1)(1-\sum_{k=0}^{p} \beta_k x_{ik})^2$$

$$=(\sum_{k=0}^{p} \beta_k x_{ik})(1-\sum_{k=0}^{p} \beta_k x_{ik}) \qquad (2.8)$$

olarak elde edilmektedir. u_i'lerin varyansı açıklayıcı değişkenlere bağlı olarak değişmektedir. Böylece β_k katsayılarının En Küçük Kareler (EKK) kestirimi yansız ancak en iyi olmayacaktır.

Değişen varyanslılık nedeni ile doğrusal olasılık modeli ağırlıklandırılarak varyanslar sabit hale getirilmelidir. Bunun için önce Y'lerin X'ler üzerindeki regresyonundan $\hat{\beta}$ kestirimleri bulunur. Buradan ağırlık değerleri,

$$w_i=1/(\sum_{k=0}^{p} \hat{\beta}_k x_{ik})(1-\sum_{k=0}^{p} \hat{\beta}_k x_{ik}) \qquad (2.9)$$

olarak elde edilir. (2.3) ile verilen model bu ağırlık ile ağırlıklandırıldığında yeni hata terimi $w_i\ u_i$, sabit varyanslı olacak ve minimum varyanslı, yansız kestirim değerleri elde edilecektir.

(2.3) ile verilen denklemin sağ tarafı 0-1 aralığı dışında olursa doğrusal olasılık modelinin geçerliliğinden şüphe edilir (Aldrich and Nelson, 1986).

2.2.1. Tekrarlı verilerde doğrusal olasılık modeli

(2.5) eşitliği ile verilen modelde $P(y_i=1)$ olasılığı X'e bağlı olarak değişeceğinden her bir yanıt değeri farklı Bernoulli dağılımı, n bağımsız tekrarlı yanıt değişkeni ise Binom dağılımı gösterecektir.

Veri kümesinde M grup ve her grupta n_j $(j=1,..,M)$ gözlem olduğunda, j'nci grupta bulunan n_j farklı y_i gözlemi y_{ij} ile gösterildiğinde,

$$y_j = \sum_{i=1}^{n_j} y_{ij} \tag{2.10}$$

elde edilir. Burada $y_{ij} \sim$ Bernoulli (P_j) ve $y_j \sim$ Binom (P_j, n_j) dağılımlıdır. j'nci grupta n_j gözlemden y_j tanesinde olumlu yanıt $(y_{ij}=1)$ elde edildiğinde ve başarı (olumlu yanıt) oranı $f_j=y_j/n_j$ olarak verildiğinde doğrusal olasılık modeli, $j=1,..,M$ için,

$$f_j = \sum_{k=0}^{p} \beta_k x_{jk} + u_j \tag{2.11}$$

şeklinde tanımlanmaktadır. Buradan elde edilecek katsayı kestirimleri yansız olacaktır ancak hata terimleri değişen varyanslıdır. Bu durumu gidermek için yapılan ağırlıklandırmada ağırlık değerleri,

$$w_j = n_j/f_j(1-f_j) \tag{2.12}$$

olarak elde edilmektedir (Aldrich and Nelson, 1986).

2.3. İki Grup Lojistik Modeller

(2.5) eşitliği ile verilen doğrusal olasılık modelinde denklemin sol tarafı 0-1 arasında sınırlı olasılık değerleri olup bu değerler, aralık dışında sonsuz değerler alabilen açıklayıcı değişkenlerle ilişkilendirildiği için eşitlik her zaman sağlanamamaktadır. Bu durumda yanıt değişkeni olarak verilen olasılık değerleri, üzerinde çeşitli dönüşümlerle $(-\infty, +\infty)$ arasında tanımlı hale getirilmelidir. Lojistik (lojit) ve probit (probability unit) dönüşümler bunlardan iki tanesi olup genelde birbirlerine yakın sonuçlar vermektedir (Aldrich and Nelson, 1986).

Lojistik modelin biyolojik deneyler (bioassay) analizi için kullanımı ilk olarak Berkson (1944) tarafından önerilmiş, Cox (1970) bu modeli gözden geçirerek çeşitli uygulamalarını yapmış, özet gelişmeler ise Anderson (1979, 1983) tarafından verilmiştir. Ayrıca verilerin lojistik modele uyumu ile ilgili bir çok çalışmalar da yapılmıştır. Bunlar arasında Aranda-Ordaz (1981), Brown (1982), Copas (1983) ve Johnson (1985) tarafından yapılanlar en önemlileridir. Pregibon (1981) iki grup lojistik modellerde etkin (influential), aykırı (outlier) gözlemleri ve teşhis ölçütlerini (diagnostics), Lesaffre (1986), Lesaffre ve Albert (1989a) ise çoklu grup lojistik modellerde etkin ve aykırı gözlemlerle teşhis ölçütlerini incelemişlerdir.

2.3.1. Lojistik model

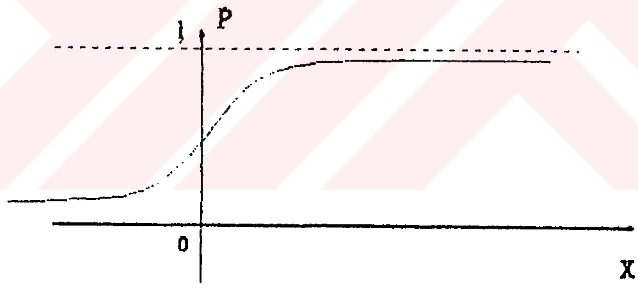
(2.5) ile verilen olasılık modelinde, olasılık değerleri üzerinde yapılan $P/(1-P)$ dönüşümü yanıt değişkeninin sınırlarını $(0,\infty)$ arasında yapacaktır. Sınırları $(-\infty,\infty)$ yapmak için bu oranın doğal logaritması alınır. Yeni yanıt değişkeni X'lerin doğrusal bir fonksiyonu olarak yazılabilir. Böylece $P_i=P(y_i=1)$ ve $i=1..,n$ için,

$$E(y_i)=L_i=Ln(P_i/(1-P_i))=\sum_{k=0}^{p} \beta_k x_{ik} \tag{2.13}$$

yazılabilir. Bu modele "lojistik model (lojit)" denmektedir. P_i olasılık değeri,

$$P_i=\exp(\sum_{k=0}^{p} \beta_k x_{ik})/(1+\exp(\sum_{k=0}^{p} \beta_k x_{ik})) \tag{2.14}$$

şeklinde tanımlanmakta olup lojistik fonksiyon adını alır. Yanıt değişkeninin açıklayıcı değişkene karşı yayılım (scatter) çiziminden ilişkileri, yanıt değişkeninin kesikliliği nedeni ile açıkça görülemez. Bunun için yanıt değişkeni yerine P_i olasılık değerine karşı çizim yapılır ve lojistik fonksiyon çizimi adını alıp S şeklinde sigmoid bir eğri verir. Fonksiyon sürekli olup 0-1 arasında değerler alır ve Şekil 2.1'de gösterildiği gibidir.



Şekil 2.1. Lojistik fonksiyon (sigmoid eğri)

Yanıt değişkeninin ikili değerler olması nedeni ile hata terimleri sıfır ortalama ve P(1-P) varyanslıdır ((2.7) ve (2.8) eşitliklerinden). Böylece hatalar binom dağılımlı olup analiz bu temele dayandırılmaktadır.

P başarı, (1-P) başarısızlık olasılığı olarak düşünüldüğünde lojistik model, başarı olasılığının başarısızlık olasılığına oranının logaritmasını açıklayıcı değişkenlere doğrusal olarak bağlayan modeldir. İki grup lojistik modele ilişkin varsayımlar şöyledir:

i) $y_i \in (0,1)$, $i=1..,n$

ii) $P(y_i=1/X_i)=P_i$, (2.14) eşitliğinde tanımlandığı gibidir.

iii) $y_1,...,y_n$'ler istatistiksel olarak bağımsızdır.

iv) Açıklayıcı değişkenler birbirinden bağımsızdır.

Gözlemlerin bağımsızlığına ilişkin 3 nolu varsayım bozulduğunda, gözlemlerin alt grupları arasındaki ilişkiyi modelleyen Polya Eggenberger dağılımı önerilmiştir (Qu, et all., 1987).

Lojistik modelin yanıt değişkenini oluşturan lojit dönüşümün (Lojit(P)=Ln P/(1-P)) özellikleri şöyle sıralanabilir:

i) P arttıkça lojit (P) de artar.

ii) P, 0-1 arasında iken lojit (P) tüm gerçel doğru üzerinde değerler alır.

iii) Eğer $P<0.5$ ise lojit $(P)<0$ ve $P>0.5$ ise lojit $(P)>0$'dır. Gözlemlerin sınıflara atamaları en basit şekilde bu kurala göre yapılabilmektedir (Aldrich and Nelson, 1986).

Açıklayıcı değişkenler üzerine kısıt getirmeyen lojistik modelde, bu değişkenlerin hepsinin sürekli, hepsinin kesikli ya da kesikli ve sürekli karışık olduğu durumda model,Bishop tarafından ayrı ayrı aşağıdaki gibi tanımlanmıştır:

i) Açıklayıcı değişkenlerin hepsi kesikli ise lojistik model (2.13) eşitliğinde tanımlandığı gibidir.

ii) Açıklayıcı değişkenlerin hepsi sürekli ise, $P(X_1,...,X_p)$, p açıklayıcı değişken üzerinde koşullu başarı olasılığı olduğunda, lojistik model,

$$Ln\left(\frac{P(X_1,\ldots,X_p)}{1-P(X_1,\ldots,X_p)}\right)=\beta_0+\sum_{k=1}^{p}\beta_k x_{ik} \qquad (2.15)$$

olarak tanımlanmaktadır ve katsayıların kestirim yöntemi olarak 2.4 alt bölümünde ele alınan yöntemlerden minimum lojit ki-kare yöntemi kullanılmaktadır.

iii) Açıklayıcı değişkenlerin bazılarının sürekli bazılarının kesikli olduğu durumda çok değişkenli frekans dağılımı, başarı durumu için $f_1(X_1,\ldots,X_p)$ ve başarısızlık durumu için $f_0(X_1,\ldots,X_p)$ şeklinde tanımlanıp lojistik model,

$$Ln\left(\frac{P\ f_1(X_1,\ldots,X_p)}{(1-P)f_0(X_1,\ldots,X_p)}\right)=\beta_0+\sum_{k=1}^{p}\beta_k x_{ik} \qquad (2.16)$$

olarak verilmektedir. Burada β katsayıları diskriminant fonksiyonunun katsayılarıdır ve gözlemleri f_0 ve f_1 fonksiyonlarına karşılık gelerek ayırır. P ise yanıt değişkeninin 1 değerini almasının önsel olasılığıdır. Bu modelde kestirim, 2.4 alt bölümünde verilen yöntemlerden Ağırlıklı En Küçük Kareler yöntemlerinden birisi ile yapılmakta ve diskriminant fonksiyonu katsayıları başlangıç değer olarak alınmaktadır (Bishop, et all., 1975).

Böylece üç ayrı şekilde ancak aynı yapıda yazılabilen lojistik model, çeşitli model ailelerinin özel bir durumu olarak tanımlanmıştır (Prentice, 1976; Aranda-Ordaz, 1981; Guerro and Johnson, 1982; Stukel, 1988; Taylor, 1988). Prentice, 1976, lojistik modeli,

$$P(X)=\int_0^{R(X)} Z^{m_1-1}\ (1-Z)^{m_2-1}\ dZ/B(m_1,m_2) \qquad (2.17)$$

şeklinde tanımlanan genel parametrik ailenin özel bir türü olarak incelenmiştir. Burada $m_1,m_2>0$, $B(m_1,m_2)$ beta fonksiyonu ve $\beta'=(\beta_0,\ldots,\beta_p)$ olmak üzere R(X) (2.14) eşitliğindeki gibi ancak matris formunda,

$$R(X)=\exp(\beta'X)/(1+\exp(\beta'X')) \tag{2.18}$$

şeklinde tanımlanmaktadır. $m_1=m_2=1$ durumu lojistik modele karşılık gelmektedir (Brown, 1982).

Bu şekilde tanımlanabilen çeşitli modellerin genellemesi "Genelleştirilmiş Lineer Modeller" adını almaktadır. Modellerde yanıt değişkeni açıklayıcı değişkenlere doğrusal bir yapı ile bağlıdır. Genelleştirilmiş lineer modeller üç bölümden oluşur: İlk bölümde yanıt değişkeni, bilinen sayıdaki denemelerde başarı sayısının bilinen binom sayımlarından oluşur. İkinci bölümde başarı olasılığı, açıklayıcı değişkenlere sadece bir doğrusal yapı ile bağlıdır. Üçüncü bölümde açıklayıcı değişkenlerin doğrusal yapısını binom sayımlarının beklenen değerine bağlayan bir fonksiyon vardır ve bu fonksiyon lojistik model için lojit dönüşüm olup bağ fonksiyonu (link function) adını alır. Diğer bir deyişle bağ fonksiyonu, yanıt değişkeninin hangi fonksiyonunun açıklayıcı değişkenlerin doğrusal bir fonksiyonunu verdiğini belirler. Gözlemlerin varyansı ortalama ile fonksiyonel olarak ilişkilidir (McCullagh and Nelder, 1983; Weisberg, 1985).

McCullagh ve Nelder (1983) tarafından detaylı olarak incelenen (GLM içerisinde yer alan) çeşitli modeller, lojistik modele çeşitli durumlarda alternatif olarak kullanılabilmektedir. Bu modeller, probit, log-log, bütünleyici log-log (complementary) ve çift üstel (double exponensial) modellerdir. Modellerin her biri için bağ fonksiyonu değişmektedir (Kay and Little, 1986). Bunlardan yalnızca probit model, lojistik model ile birlikte en sık kullanılanı olması nedeni ile izleyen alt bölümde incelenmektedir.

2.3.2. Probit analizi ve probit model

Finney'in (1971) geliştirdiği ve doz-yanıt (dose-response) çalışmalarına uygulanan probit analizi,diskriminant analizi gibi bir modelleme tekniğidir. Teorisi şöyle özetlenebilir:

Eğer doz λ ile ölçülürse, tolerans gösterilebilecek doz miktarına ilişkin dağılım $dP=f(\lambda)d\lambda$ ile tanımlanabilir. Buna göre dP oranı, toleransı λ ile $\lambda+d\lambda$ arasında olan bireyleri içerir. Eğer tüm kitleye λ_0 doz miktarı verilirse tüm bireyler toleransı λ_0'dan az olan yanıt verecektir ve bunun oranı P olup,

$$P=\int_0^{\lambda_0} f(\lambda)d\lambda \tag{2.19}$$

şeklinde tanımlanmaktadır.

$v \sim N(0,1)$ olmak üzere $y_i=\sum_{k=0}^{p} \beta_k x_{ik}$ şeklindeki yanıt (başarı) olasılığı, Φ standart normal dağılımını gösterdiğinde, $i=1,...,n$ için,

$$P(y_i=1/X)=P_i=(1/\sqrt{2\Pi})\int_{-\infty}^{y_i} \exp(-1/2\ v^2)dv$$

$$=\Phi(y_i) \tag{2.20}$$

şeklinde verilmekte ve

$$\Phi^{-1}(P_i)=\sum_{k=0}^{p} \beta_k x_{ik}=y_i \tag{2.21}$$

doğrusal modeli elde edilmektedir.

Burada Φ fonksiyonu probit modelin özelliği olup, lojistik modelde bunun yerine (2.14) eşitliğindeki gibi exponansiyel fonksiyonlar gelmektedir.

Probit ve lojistik modeller için fonksiyonel temeller aynı ancak karşılık gelen dağılımların kuyrukları bir miktar farklıdır. Genelde gözlem sayısı fazla olmadıkça birbirine yakın sonuçlar veren bu modeller extrem olasılıklarda çok sayıda gözlem olduğunda oldukça farklılaşırlar. Modellerin yanıt değişkeni tanımı ve varyans formülleri değişiktir. Katsayı kestirimleri aralarında ölçek farkı olması nedeni ile birbirleri ile doğrudan kıyaslanamazlar (Ashton, 1972; Aldrich and Nelson, 1986).

Amemiya (1981), doğrusal olasılık model katsayı kestirimi ile lojistik ve probit modellerin katsayı kestirimleri arasında ilişkiyi incelemiş ve bu ilişkileri basit doğrusal ilişkiler olarak göstermiştir.

Maddala (1988), katsayılar kestirildikten sonra açıklayıcı değişkenlerdeki değişimlerin,gözlemlerin gruplara atanma olasılık kestirimleri üzerindeki etkisini lojistik, probit ve doğrusal olasılık modelleri için incelemiştir.

Prentice, gerek probit gerekse lojistik modellerin doz-yanıt eğrileri için genellemesini yapmıştır. Probit model, (2.17) ile tanımlanan genel modelde sıra ile çarpıklık ve basıklık katsayıları olan m_1 ve m_2'nin farklı değerleri aldığı duruma karşılık gelmektedir (Prentice, 1976).

2.3.3. Lojistik ayrımsama

Yanıt değişkeninin 0,1 değerlerine karşılık gelen H_1 ve H_2 grupları, $X_1,...,X_p$ açıklayıcı değişkenlerine dayanılarak ayırt edilmek istensin. Gruplardaki gözlem sayısı sıra ile n_1 ve n_2 olduğunda $n=n_1+n_2$ gözleme dayalı ayrımsama kuralının oluşumu $f_s(X_1,...,X_p)$ şeklindeki olasılık fonksiyonunun fonksiyonel yapısına ilişkin varsayımlara dayanır. Fonksiyon yapısı için üç tür varsayım söz konusudur:

i) Çok değişkenli normal dağılım fonksiyonu (fully distributional)

ii) Lojistik ayrımsama fonksiyonu (partially distributional)

iii) Dağılımdan bağımsız Kernel ayrımsama fonksiyonu (distribution free).

Lojistik ayrımsama fonksiyonu söz konusu olduğunda $X_0=1$ iken $f_s(X)=f_s(X_1,...,X_p)$, H_s (s=1,2) grubunun olasılık yoğunluk fonksiyonu olarak tanımlanır. Lojistik varsayımda, $\beta'=(\beta_0,\beta_1,...,\beta_p)$ için,

$$f_1(X)/f_2(X)=\exp(\beta'X') \tag{2.22}$$

ya da logaritması alınarak,

$$Ln(f_1(X)/f_2(X))=\beta'X' \tag{2.23}$$

şeklinde tanımlanmaktadır (Lesaffre, 1986). Bu son eşitlik log-olabilirlik oranı olup X'lerde doğrusaldır. $f_s(X)$ fonksiyonu üzerinde,tanım bölgesi içersinde integralinin 1 olması dışında bir kısıt yoktur.

(2.22) eşitliği,

$$f_1(X)/f_2(X)=\exp(h(X,\beta)) \tag{2.24}$$

formunda genellenebilmektedir. Burada h fonksiyonu β'da doğrusal, X'lerde doğrusal olmayabilen bir fonksiyondur. Lojistik varsayım bilinmeyen β parametrelerini içermektedir. Her bir gözlem için X koşulu altında gruplardan birine atanma olasılığı olarak tanımlanan sonsal olasılıkları hesaplamak için β kestirimleri gerekmektedir. Bunun için lojistik varsayım altında örneklemin olabilirlik fonksiyonu belirlenmelidir. Olabilirlik fonksiyonu ise gözlemlerin örneklenme türlerinden etkilenmektedir. Örnekleme türleri ayrımsama kuralının oluşturulmasında ve böylece gözlemleri H_1 ve H_2 gruplarından birine atamada etkili olmaktadır (Lesaffre, 1986). Katsayı kestirimlerinin her biri için değiştiği üç tür örnekleme vardır. Bunlar karışık (mixture or prospective), koşullu (conditional) ve ayrı (separate or retrospective) örneklemedir.

Karışık örneklemede gözlemler (X,H) bileşik dağılımından örneklenmekte yani gözlemler hangi gruptan olduğu bilinmeksizin rastgele seçilmektedir. Burada H grup üyeliğini gösteren değişken olup iki grup olduğunda H_1 ve H_2 şeklinde gösterilmektedir. Bu örnekleme türüne ilişkin olabilirlik fonksiyonu ve fonksiyonun koşullu örneklemenin olabilirlik fonksiyonu ile ilişkisi Lesaffre (1986) ve Albert ve Lesaffre (1986) tarafından incelenmiştir.

Koşullu örneklemede H'nın X koşulu altında dağılımı incelenmekte, gözlemler bu koşullu dağılımdan örneklenmektedir. Biyolojik deney analizinde çok sık kullanılan bu örnekleme türüne ilişkin olabilirlik fonksiyonu diğer örnekleme türlerinin olabilirlik fonksiyonuna temel teşkil etmektedir (Lesaffre, 1986).

Ayrı örneklemede ise X'in H koşulu altında dağılımından örnekleme yapılmaktadır. Anderson (1972; 1982) tarafından detaylı olarak incelenen bu örnekleme türünün genelde uygulaması zor olup geçmişe yönelik çalışmalarda (retrospective) uygulanmaktadır (Kay and Little, 1986).

2.4. İki Grup Lojistik Modellerde Kestirim Yöntemleri ve Uyum İyiliği Ölçütleri

İki grup lojistik modelin katsayılarını kestirmede kullanılan yöntemler, En Çok Olabilirlik (Maximum Likelihood, EÇO), Yeniden Ağırlıklandırılmış İteratif En Küçük Kareler (Reweighted Iterative Least Square, RILS) yöntemleri ile tekrarlı veri durumunda kullanılan Minimum Lojit Ki-Kare (Minimum Logit Chi-Square, MLCS) yöntemleridir.

2.4.1. En çok olabilirlik yöntemi

Başarı olasılığı $P_i=P(y_i=1/X)$ olarak tanımlandığında başarısızlık olasılığı $1-P_i$ olup olasılık genel olarak $i=1,...,n$ için,

$$P(y_i/X_i)=P_i^{y_i}(1-P_i)^{1-y_i} \tag{2.25}$$

şeklinde yazılabilmektedir. n gözlem için bu olasılık,

$$L(Y/X)=P(Y/X)=\prod_{i=1}^{n} p_i^{y_i}\,(1-P_i)^{1-y_i} \tag{2.26}$$

olarak genellenmekte olup olabilirlik fonksiyonunun tanımını vermektedir.

EÇO yönteminin kuralı, p açıklayıcı değişkene ilişkin β kestirimini, Y yanıtını gözleme olabilirliğini mümkün olduğunca büyük kılacak şekilde bulmaktır. $L(Y/X,\beta)$ olabilirlik fonksiyonu olmak üzere, yöntem $\hat{\beta}$'yı bu fonksiyonu maximum yapacak şekilde seçmektedir. Lojistik modelin olabilirlik fonksiyonu, (2.25) eşitliğinde P_i yerine (2.14) ile verilen açık ifadesi konularak elde edilir. Olabilirlik fonksiyonunun logaritması,

$$\text{Ln } L(Y/X,\beta)=\sum_{i=1}^{n}(y_i \text{Ln}P_i+(1-y_i)\text{Ln}(1-P_i)) \tag{2.27}$$

şeklinde olup bunun β'ya göre birinci türevi, j=1.., p için

$$\sum_{i=1}^{n}(y_i-P_i)x_{ij}=0 \tag{2.28}$$

olabilirlik denklemini vermektedir. Bu denklemin çözümü ile $\hat{\beta}$ kestirim değerleri elde edilmektedir. P_i'nin

(2.14) ile tanımlandığı gibi üstel olması nedeni ile denklem β'da doğrusal değildir. Bu nedenle EÇO yöntemi ile tek adımda kesin çözüm elde edilemez, iteratif çözümleme gerekmektedir (Aldrich and Nelson, 1986).

İteratif çözümlemede β'ların başlangıç değerleri verilerek elde edilen ilk kestirimlerinde, her adımda δ kadar küçük miktarda ayarlamalar yapılarak türevler alınır ve EÇO kestirimleri bulunur. İteratif işlemler yakınsama sağlayıncaya dek devam eder. Yakınsama ancak δ düzeltme terimlerinin iterasyon değerlerini değiştirmediği noktada sağlanır ve süreç durur.

Lojistik modelin EÇO kestirimlerini bulmak için iterasyona başlarken başlangıç değerlerini vermenin çeşitli yolları vardır. Bunlardan iki tanesi diskriminant fonksiyonunun katsayılarını kullanmak ve de 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. EÇO yönteminde her adımda yapılacak düzeltme miktarı, tek açıklayıcı değişken durumunda lojit tablolarından elde edilebilmektedir (Berkson, 1957). Veriler birbirinden çok ayrık olduğunda EÇO kestirimlerinde yakınsama elde edilememektedir (McCullagh and Nelder, 1983).

Gözlemlerin J tane uzak değeri ($J \leq n$) verilerin J grupta tekrar edildiği duruma karşılık gelmektedir. Bu uzak gruplara "birliktelik örüntüsü (covariate pattern)" adı verilmektedir. y_j, n_j gözlem üzerinden y_{ij} yanıt değerlerinin toplamını gösterdiğinde log-olabilirlik fonksiyonu,

$$L(\beta)=\sum_{j=1}^{J} y_j \, Ln(n_j \, P_j)+(n_j-y_j)L_n(n_j(1-P_j)) \qquad (2.29)$$

olarak verilmektedir. Burada $P_j=P(y_j=1/X_j)$ olasılığı olup j, her bir gözlemi değil her bir gözlem grubunu göstermektedir. Her grupta n_j gözlem bulunmaktadır. Bu durumda katsayıların EÇO kestirimleri bu olabilirlik fonksiyonunun maximizasyonu ile elde edilmektedir (Hosmer and Lemeshow, 1989a).

2.4.2. Yeniden ağırlıklandırılmış iteratif en küçük kareler yöntemi

Gruplandırılmış verilerde J grubun her birinde n_j denemeden r_j başarı elde edildiğinde j=1.., J için başarı oranı $P_j=r_j/n_j$ olarak tanımlanmaktadır. $Var(r_j/n_j)=P_j(1-P_j)/n_j$ olduğundan her binom dağılımlı gözlem için varyans değişmektedir. Bu durumda lojit(r_j/n_j)'nin açıklayıcı değişkenler üzerinde $w_j=n_j/P_j(1-P_j)$ ağırlığı ile ağırlıklandırılmış regresyonu uygun olacaktır. Ancak w_j ağırlık değerleri de P_j'nin bir fonksiyonu olduğu için EKK yöntemi iteratif olarak uygulanacak, her adımda ağırlıklar kestirimlere bağlı olarak yeniden elde edilecektir. Weisberg (1985) bu yöntemin aşamalarını ayrıntılı olarak açıklamıştır.

Bu yöntemde yakınsamanın elde edilmesi 2.3.5 alt bölümünde değinilen örnekleme türlerine bağlı olmaktadır (Lesaffre, 1986).

2.4.3. Minimum lojit ki-kare yöntemi

Log-doğrusal modeli test etmede kullanılan ve ağırlıklı EKK kestiriminin özel bir türü olarak 2XJ çapraz tablolarında Berkson (1955) tarafından önerilen yöntemde, beklenen ve gözlenen lojit değerleri arasındaki farkdan yararlanılmaktadır (Bishop, et all., 1975). Yöntemin lojistik regresyon analizinde kullanımı tekrarlı veriler olduğu durumda söz konusudur. Yukarıda, RILS yönteminde değinilen P_j başarı olasılığı (2.14) eşitliğinde tanımlandığı gibidir. Bu olasılık üzerinde yapılan lojit dönüşüm MLCS yönteminde yanıt değişkenini oluşturmaktadır. Kestirimde kullanılacak ağırlık değeri $n_j\ P_j(1-P_j)$ olarak elde edilmektedir (Smith, et all., 1984).

Yöntem, lojit değeri olarak tanımlanan yanıt değişkeninin açıklayıcı değişkenler üzerinde yukarıda değinilen ağırlık değeri ile ağırlıklandırılmış regresyonundan EKK kestirimlerini elde etmeye dayanmaktadır. Buradan tek adımda bulunan ağırlıklı EKK kestirimleri MLCS kestirimleri adını almaktadır (Aldrich and Nelson, 1986).

Olasılık değerinin 0 ya da 1 olduğu durumda lojit değeri tanımlı olmayacağı için P_j yerine $P_j+1/2n_j$ değerinin konulduğu ayarlanmış lojit ki-kare yöntemi kullanılmaktadır (All-Sarraf and Young, 1986).

2.4.4. Diğer kestirim yöntemleri

Yukarıda değinilen kestirim yöntemleri dışında iteratif olmayan ağırlıklı EKK ve diskriminant fonksiyonu yöntemleri de söz konusudur. İteratif olmayan EKK yönteminin getirdiği bir kısıt, veri kümesindeki çoğu gözlem için olasılık kestirim değerinin sıfırdan farklı olmasıdır. Bu kısıt açıklayıcı değişken sayısı çok olduğunda sağlanamayacağı için yöntem, tek açıklayıcı değişken durumu dışında kullanışlı değildir.

Diskriminant fonksiyonu kestirim yöntemi ise açıklayıcı değişkenlerin yanıt değişkeni koşulu altında normal dağıldığı varsayımına dayanmaktadır. Açıklayıcı değişkenler arasında kesikli olanlarına çok sık rastlandığından normallik varsayımı nadiren sağlanmaktadır. Bu varsayım bozulumunda elde edilen diskriminant fonksiyonu kestirimleri yanlı olacağından, yöntem bir ön analiz olarak ve lojistik modele başlangıç kestirimleri vermek amacı ile kullanılabilmektedir (Hosmer and Lemeshow, 1989b).

Öte yandan beklenen ve gözlenen olasılık değerlerinin farkına dayalı minimum ki-kare yöntemi (Al-Sarraf and Young, 1986) ile tek açıklayıcı değişken durumunda Fourier dönüşümü yardımı ile geliştirilen algoritmayı kullanarak kestirim veren yöntemler de önerilmiştir (Tritchler, 1984).

2.4.5. Kestirim yöntemlerinin karşılaştırılması

Lojistik modelin parametre kestiriminde kullanılan yöntemlerden EÇO yöntemi her zaman tutarlı, etkin ve yeterli kestirimler verir ancak bu kestirimler her zaman yansız olmayabilir, asimtotik olarak yansızdır. Ayrıca bu kestirimler asimtotik olarak normal dağılımlıdır. Doğrusal olasılık modelinin ağırlıklı EKK kestirimi ile lojistik modelin EÇO kestirimi benzer istatistiksel özelliklere sahiptir. Varsayımlar sağlandığı sürece EKK ve EÇO kestirimleri aynı özelliktedirler. Tek fark EÇO yönteminin olabilirlik denkleminin doğrusal olmayıp iteratif türevlemeler gerektirmesidir. Bu da hesaplama maliyetini artırmaktadır. Her iki kestirimin ortak özelliği yansız, etkin ve normal kestirimler vermeleridir.

RILS kestirimlerinin yansızlık, etkinlik ve normallik özellikleri büyük örneklemlerde asimtotik olarak sağlanmaktadır.

MLCS kestirimleri asimtotik olarak etkin ve yeterlidir (Aldrich and Nelson, 1986).

Lojistik modelin kestirimlerini elde etmede hangi yöntemin daha yararlı olacağını sınamak için yapılan simülasyon çalışmalarına göre, amaç istatistiksel çıkarsama olduğunda EÇO kestirimleri MLCS kestirimlerine tercih edilmektedir. Bazı lojistik ve probit modeller için MLCS kestirimi EÇO kestirimlerine tercih edilmektedir ancak bu modeller için MLCS kestirimlerinin daha iyi olduğuna dair teorik bir kanıt yoktur (Smith, et all., 1984).

Bu iki kestirim yönteminin karşılaştırıldığı diğer bir simülasyon çalışması Smith ve arkadaşları (1984) tarafından yapılmıştır. Basıklık ve çarpıklık katsayılarının da dikkate alındığı çalışmada MLCS kestirimlerinin EÇO'den daha yanlı olduğu ve varyansının teorik değerinden çok saptığı gözlenmiştir. Nokta kestirimi için MLCS yöntemi tercih edilirken çıkarsama için EÇO kestirimlerinin daha iyi olduğu sonucuna varılmıştır. Basıklık katsayısının

EÇO yönteminde daha büyük sapma göstermesi MLCS yöntemi lehine tek sonuç olarak gözlenmiştir.

Öte yandan açıklayıcı değişkenlerin hatalı ölçüldüğü durumda Carrol ve arkadaşları koşullu EÇO kestirimini ve özelliklerini araştırmışlardır. Ayrıca yaptıkları simülasyon çalışması sonucunda büyük örneklemlerde yan'ın etkin olup koşullu EÇO yönteminin tercih edileceğini göstermişlerdir (Carrol, et all., 1984).

2.4.6. Uyum iyiliği ve sapma ölçütleri

Lojistik varsayım çok değişkenli dağılımların geniş bir kesimi için sağlansa da kurulan lojistik modelin geçerliliği sınanmalıdır. Modelde bulunması gereken tüm değişkenler ele alındıktan sonra modelin yanıt değişkenini açıklamadaki etkinliğini araştırmaya "uyum iyiliği araştırması" denmektedir. İki grup lojistik model için lojistik varsayım (2.23) eşitliğinde verildiği gibidir. Bu varsayım bozulumu çeşitli şekillerde olabilmektedir:

i) Logaritmik dönüşüm yerine bir başka dönüşüm kullanılabilir.

ii) Logaritmik dönüşüm doğru olsa bile açıklayıcı değişkenler uygun olmayabilir. Örneğin önemli değişkenler ya da gerekli etkileşim terimleri modele alınmadığında varsayım bozulacaktır.

iii) Tüm değişkenler modelde bulunur ancak ölçek yanlıştır.

iv) Modelde aykırı değer vardır.

Bu tür varsayım bozulumlarını araştırmak için modelin uyum iyiliği incelemesinde grafiksel analiz ve istatistiksel ölçütlerin kullanımı yararlı olmaktadır. Lojistik modelin uyum iyiliğini araştırmada kullanılabilecek ölçütler şunlardır:

i) Tüm değişkenleri içeren model ile kestirilen modele ilişkin olabilirlik oran değerlerinin farkına dayanan

Artık Kareler Toplamına (AKT) benzer ölçütler ki-kare dağılımlı olup modelin geçerliliğini sınamada kullanılırlar. Bu yol ile modele eklenen karesel terimin etkisi de sınanabilmektedir (Cox, 1970).

ii) Artık değerleri hesaplanarak bunların X'e ya da olasılık değerlerine karşı çiziminden aykırı değerler araştırılır. Bulunacak bu aykırı değer model uyumundaki sorunun göstergesi olacaktır.

iii) AKT ve olabilirlik oranına dayalı R^2 türü ölçütler vardır. Uyum testi için yapay (pseudo) R^2 ölçütleri incelenebilmektedir (Maddala, 1988).

iv) Lojistik model ayrımsama amacı ile kullanıldığından modelin doğru sınıflandırma yüzdeleri de birer uyum iyiliği ölçütü olabilir. Bu yüzdeler yararlı bir ölçüt olmasına rağmen ayrımsama gücünün yeterli bir göstergesi değildir; ancak,her problemde incelenmesinde destek olması açısından yarar vardır.

Lojistik modelde normallik varsayımı sağlamadığı için modelin uyum iyiliği testlerinde t ve F ölçütleri yerine ki-kare, G^2 vb. gibi parametrik olmayan ölçütler kullanılmalıdır. Bu ölçütler izleyen alt bölümlerde ele alınmaktadır.

2.4.6.a. ki-kare, G^2, sapma ölçütleri

Parametrik olmayan en basit uyum iyiliği ölçütleri ki-kare ve G^2 olup "O" gözlenen, "E" beklenen değerleri, OxLog O ve OxLog E ise sıra ile gözlenen ve beklenen olabilirlikleri gösterdiğinde,

$$\chi^2 = \Sigma(O-E)^2/E \qquad (2.30)$$

ve

$$G^2 = -2\Sigma O \text{ Log}(E/O) = 2\Sigma O \text{ Log}(O/E)$$
$$G^2 = -2(\Sigma O \text{ Log } E - \Sigma O \text{ Log } O)$$
$$G^2 = -2(L(E)-L(O))$$

şeklinde tanımlanmaktadır (Burada log=Ln dir) (Bishop, et all., 1975). Lojistik modelin uyum iyiliği testi için de kullanılabilen bu ölçütlerden ki-kare, veriler J grupta tekrar edildiğinde $\hat{P}_i$, olasılık kestirim değeri olmak üzere,

$$\chi^2=\sum_{i=1}^{J}(y_i-n_i\hat{P}_i)^2/n_i\hat{P}_i(1-\hat{P}_i) \tag{2.31}$$

şeklinde verilmektedir. Tekrar olmadığı durumda formüldeki n_i terimi kalkacaktır ve J=n olacaktır. Ancak bu ölçüt P=0 ve P=1 olduğunda tanımsız olacağı için P'nin 0 ve 1 değerleri yerine sırasıyla 1/2n ve (1-1/2n) değerleri konularak ayarlama yapılmaktadır (Ashton, 1972; Pregibon, 1981; Weisberg, 1985).

Lojistik regresyon analizinde gözlenen ve beklenen modellerin karşılaştırılması kesiklilikten dolayı log-olabilirlik fonksiyonlarına dayanmaktadır (Burada gözlenen ile mevcut model, beklenen ile hipotetik model kastedilmektedir). Gözlem sayısı kadar parametre içeren modele doymuş (saturated) model denildiğinde, bunun kestirilen model ile karşılaştırılması G^2 ölçütüne karşılık gelen sapma ölçütü ile yapılmakta olup bu ölçüt,

$$D=-2Ln\left(\frac{\text{Kestirilen model olabilirliği}}{\text{Doymuş model olabilirliği}}\right) \tag{2.32}$$

şeklinde tanımlanmaktadır. Burada -2 katsayısının alınma nedeni matematiksel olup bilinen dağılıma ulaşmak içindir. Olabilirlik fonksiyonları yerine konulduğunda lojistik modelin sapma ölçütü,

$$D=\sum_{i=1}^{n}d_i^2=-2\sum_{i=1}^{n}(y_iLn(\hat{P}_i/y_i)+(1-y_i)Ln((1-\hat{P}_i)/(1-y_i))) \tag{2.33}$$

olarak elde edilmektedir. (n-p) serbestlik dereceli ki-kare dağılımı gösteren bu ölçüt çeşitli modeller arasından seçim yapmada kullanılmakta, en küçük sapma değerli model en iyi model olarak alınabilmektedir.

Veriler J grupta tekrar edildiğinde d_j ve bundan yararlanarak bulunan D ölçütü,

$$d_j=\pm\sqrt{2}(y_j Ln(y_j/n_j\hat{P}_j)+(n_j-y_j)Ln((n_j-y_j)/n_j(1-\hat{P}_j)))^{1/2}$$

(2.34)

$$D=\sum_{j=1}^{J} d_j^2$$

şeklindedir. Bu ölçüt (J-p-1) serbestlik dereceli ki-kare dağılmaktadır (Hosmer and Lemeshow, 1989a; 1989b).

2.4.6.b. diğer uyum iyiliği ölçütleri

Doğrusal regresyon analizinde model katsayılarının anlamlılığı tümel F testi ile araştırılmaktadır. Lojistik regresyon analizinde bu teste karşılık olarak (2.32) ile tanımlanan sapma ölçütü kullanılabilmektedir. Pay'da tek bir sabit terim içeren modelin olabilirlik değerinin (L_0) payda'da ise elde edilen modelin olabilirlik değerinin (L_1) yer aldığı ölçüt,

$$C=-2Ln(L_0/L_1)=-2(LnL_0-LnL_1) \qquad (2.35)$$

şeklinde olup (p-1) serbestlik dereceli ki-kare dağılımı göstermektedir.

Öte yandan C'ye benzer olarak en iyi modele ulaşmak için çeşitli modellerin olabilirlik değerleri birbirleri ile (2.32) ölçütü yardımı ile kıyaslanabilmektedir. Modellerin içerdikleri değişken sayıları arasındaki fark, ki-kare dağılımlı ölçütün serbestlik derecesini oluşturmaktadır (Stopher and Meyburg, 1979).

Lojistik modelin uyum iyiliği testinde kullanılabilecek en basit ölçütlerden biri, doğrusal regresyonda modelin açıklanma miktarı R^2'ye karşılık gelen yapay (pseudo) R^2 olup,

$$\text{pseudo-}R^2 \doteq C/(C+n) \qquad (2.36)$$

şeklinde tanımlanmaktadır. Burada C değeri (2.35)'de verildiği gibidir. 0-1 arasında değerler alan bu ölçütün sıfıra yakın değeri modelin uyum eksikliğinin bir göstergesi olmaktadır.

Bunun yanı sıra önerilen diğer bir ölçüt, modelin log-olabilirlik değerinin -2 katı şeklinde olup T1 ölçütü adını almakta ve (n-p-1) serbestlik dereceli ki-kare dağılımı göstermektedir (Kay and Little, 1986).

Uyum iyiliğinin aksine uyum eksikliğini, verileri gruplara bölerek test etmeye yönelik ölçütler de bulunmaktadır. Bunlardan bir tanesi yukarıda değinilen T1 ölçütünün bir versiyonudur ve aralıklara bölünerek kesikli hale getirilen sürekli açıklayıcı değişkene ilişkin bölünmeden önce ve sonraki olabilirlik fonksiyonlarının farkı olarak tanımlanmaktadır. Ki-kare dağılımlı ölçüt Landwehr ve arkadaşlarının (1984) önerdiği lokal-ortalama sapma (local mean deviance) adlı grafiksel yaklaşımın özel bir durumudur (Kay and Little, 1986). Bu grafiksel yaklaşımda ise veri kümesi gruplara bölünmekte, her grubun sapma ölçütüne lokal etkisi araştırılmaktadır. Bu etkiler tümel (global) etki ile kıyaslanarak modelin uyum eksikliği olup olmadığına karar verilmektedir (Landwehr, et all., 1984).

Diğer bir uyum eksikliği ölçütü Tsiatis (1980) tarafından önerilmiştir. Açıklayıcı değişkenleri gruplara bölerek hesaplanan ve grup sayısı-1 serbestlik dereceli ki-kare dağılan bu ölçüt lokal ortalama sapma çizimlerine karşılık gelmekte ve "skor ölçütü" adını almaktadır (Lesaffre, 1986).

Benzer olarak, Hosmer ve Lemeshow, açıklayıcı değişkenlerin gruplara bölünmesini olasılık kestirimlerine dayanarak yapan Hosmer-Lemeshow test ölçütünü önermişlerdir (Hosmer and Lemeshow, 1980, 1989b; Lemeshow and Hosmer, 1982). Ölçüt, veriler g gruba bölündüğünde 2xg tablosundan Pearson ki-kare değerlerini hesaplamaya dayanmakta olup (g-2) serbestlik dereceli ki-kare dağılmaktadır. Kolay yorumlanması ölçütün bir avantajı iken verileri gruplandırmadan kaynaklanan bilgi kaybının uyumdan sapmaları göz ardı etmeye neden olabilmesi de dezavantajıdır. Hosmer ve Lemeshow (1980), bu ölçüte benzer daha genel uyum eksikliği (iyiliği) ölçütlerinden de söz etmektedirler.

Tüm bu ölçütlerden başka elde edilen çeşitli modeller arasında seçim yapmaya yönelik bir ölçüt de modellerin doğru sınıflandırma olasılıklarına dayanmaktadır. Hilden ve arkadaşları bu amaçla,

$$T7=\frac{1}{n}\sum_{i=1}^{n}(y_i\hat{P}_i+(1-y_i)Ln(1-\hat{P}_i)) \tag{2.37}$$

ölçütünü kullanmışlardır. Bu ölçüte göre, en büyük T7 değerini veren modeli en iyi model olarak tanımlamışlardır (Kay and Little, 1986).

Ayrıca modelin sınıflandırma çizelgesinden elde edilen doğru/yanlış sınıflandırma oranı da başlı başına bir uyum iyiliği ölçütü olarak değerlendirilebilmektedir. Çizelgede yanıt değişkeninin gerçek ve kestirim değerleri çaprazlanmaktadır. Kestirim değeri, 0.5 değerini aşarsa 1 aksi taktirde 0 grubuna atanmaktadır. Doğru/yanlış sınıflandırma oranını uyum iyiliği kriteri olarak kullanmanın bazı dezavantajları söz konusudur: Sürekli olarak kestirilen yanıt değişkeni kritik değerler yardımı ile kesikli hale getirilmektedir. Pratikte P=0.48 ve P=0.52 arasında çok az fark vardır. 0.5 kritik değeri ile karşılaştırıldığında yakın olan bu değerler farklı gruplara atanacaktır. Öte yandan sınıflandırma çizelgesinde daha çok gözlemli gruba atama yapılmakta olup bu da model uyumundan bağımsızdır. Amaç sınıflandırma olduğunda bu çizelgenin kullanımı uygun iken uyum iyiliği testine sadece bir katkı sağlamaktadır (Hosmer and Lemeshow, 1989b).

Lojistik modelin uyum iyiliği ölçütlerinin yanı sıra, katsayıların anlamlılığını ve etkisini test etmeye yönelik, gerek analitik gerekse grafiksel ölçütler bulunmaktadır. Analitik ölçütlerden biri Wald ölçütü olup katsayı kestirim değerlerinin standart hatalarına oranı olarak tanımlanmakta ve Z değeri ile karşılaştırılmaktadır. Ölçütün 2'den büyük değerleri için değişkenlerin önemli olduğu sonucuna varılmaktadır. İkiden çok açıklayıcı değişken durumunda katsayı kestirimlerinin iteratif olarak elde edilmesi nedeni ile ölçütün kullanımı zorlaşmaktadır.

Copas (1983), açıklayıcı değişkenlerin modeldeki yapısını ve anlamlılığını incelemek amacı ile olasılık kestirimlerinin çizimini önermiştir. Çizimden elde edilen eğrinin şekline göre açıklayıcı değişkenin doğrusal, karesel, vb. etkilerine karar verilebilmektedir. Bu amaca yönelik bir diğer çizim de kısmi artık çizimleridir (Fienberg and Gong, 1984; Landwehr, et all. 1984).

Modelde bulunan açıklayıcı değişkenlerin anlamlılığı bu şekilde sınandıktan sonra etkileşim* (interaction) terim-

---

* Etkileşim ile karışım (confounding) terimleri karıştırılmamalıdır. Bunlar farklı ifadeler olmalarına rağmen her ikisi de iki ya da daha çok değişken arasındaki ilişkiyi, bu ilişkiyi etkileyecek ek değişkenleri kontrol ettikten sonra değerlendirmektedir. Eğer değişkenler arasındaki ilişki kontrol edilen değişkenin ihmali ile anlamlı olarak değişiyorsa karışım söz konusudur. Eğer kontrol edi-

./.

lerinin varlığı ya da değişkenlerin karesel etkilerinin söz konusu olup olmadığı incelenmesi gereken bir diğer durumdur ve bir anlamda uyum iyiliği testine karşılık gelmektedir. Bu test etkileşim terimlerini içeren model için yukarıda tanımlanan uyum iyiliği ölçütlerinden D ile yapılarak modelin doğrusal ya da karesel olduğuna karar verilebilmektedir (Hosmer and Lemeshow, 1989b).

Bütün bunların yanı sıra Jennings (1986a), değişkenlerin karesel ve kübik etkilerinin dikkate alındığı modelde uyum iyiliği ölçütlerini incelemiş, Brown (1982) ise lojistik modeli genel parametrik ailenin bir elemanı olarak ele almış ve bu genel modelden yararlanarak lojistik modelin uyum iyiliğini inceleyen bir ölçüt önermiştir.

2.5. Çoklu Grup Lojistik Modeller

Yanıt değişkeninin ikiden çok düzey içeren kesikli değişken olduğu durumda bu değişkenin açıklayıcı değişkenler üzerindeki regresyon modeli çoklu grup (multigroup or polychotomous) lojistik model adını almaktadır.

İki grup lojistik model üzerinde bazı değişimler ile çoklu gruba uyarlanabilmektedir. İlk olarak yanıt değişkeninin 0, 1 ve 2 gibi üç düzeyli olduğu durum ele alınsın. Bu durumda iki ayrı iki grup lojistik model söz konusudur: Biri Y=1'e karşı Y=0 için, diğeri ise Y=2'ye karşı Y=0 içindir. Böylece Y=0 grubu temel (reference, base) gruptur. Y=2'ye karşı Y=1'i kıyaslayan lojistik fonksiyon yukarıda tanımlı iki karşılaştırmaya ilişkin lojistik fonksiyonların farklarından elde edilmektedir.

---

len değişkenin farklı düzeyleri için değişkenlerarası ilişki farklı ise etkileşim söz konusudur. Etkileşim karışımdan önce denetlenmelidir. Kontrol edilen değişken sayısı birden çok olduğunda bunların hepsi yerine bir alt kümesini kontrol etmek sorunu doğar. Kontrol edilen değişkeni içeren kuvvetli etkileşim söz konusu olduğunda bu değişken için karışımı test etmek uygun değildir ve önemli etkileşimler olduğunda karışımın denetimi önerilmez (Kleinbaum, et all., 1988).

$X=(X_0,\ldots,X_p)$ (p+1) boyutlu açıklayıcı değişkenler vektörü olarak alındığında lojistik fonksiyonlar;

$$g_1(X)=\mathrm{Ln}\left(\frac{P(Y=1/X)}{P(Y=0/X)}\right)=\beta_{10}+\beta_{11}X_1+\ldots+\beta_{1p}X_p$$

(2.38)

$$g_2(X)=\mathrm{Ln}\left(\frac{P(Y=2/X)}{P(Y=0/X)}\right)=\beta_{20}+\beta_{21}X_1+\ldots+\beta_{2p}X_p$$

şeklinde tanımlanmaktadır. Buradan yanıt kategorileri için koşullu olasılıklar üç grup durumunda genel olarak j=0,1,2 olmak üzere

$$P_j(X)=P(Y=j/X)=\exp(g_j(X))\Big/\sum_{k=0}^{2}\exp(g_k(X)) \qquad (2.39)$$

şeklinde elde edilmektedir. Modelde $\beta_0=0$ ve böylece $g_0(X)=0$ dır. Her bir $P_j(X)$ olasılıkları 2(p+1) parametrenin fonksiyonudur.

Olabilirlik fonksiyonunu oluşturmak için grup üyeliğini belirlemede üç tane ikili değişken formülasyonu kullanılır. Bu değişkenler, Y=0 için $y_0=1$, $y_1=0$, $y_2=0$; Y=1 için $y_0=0$, $y_1=1$, $y_2=0$; Y=2 için $y_0=0$, $y_1=0$, $y_2=1$ şeklinde kodlanmaktadır. Y'nin tüm değerleri için $\sum_j y_j=1$'dir.

n bağımsız gözlemli örneklem için koşullu olabilirlik fonksiyonu,

$$L(\beta)=\prod_{i=1}^{n}\left(P_0(X_i)^{y_{0i}}\,P_1(X_i)^{y_{1i}}\,P_2(X_i)^{y_{2i}}\right) \qquad (2.40)$$

şeklindedir (Hosmer and Lemeshow, 1989b).

Yanıt değişkeni g kategorili olduğunda grupları ikişerli karşılaştıran (g-1) tane lojistik modele gereksinim duyulmaktadır. Bu durumda (2.39) ile verilen olasılıklar,

$\beta_j' = (\beta_{j0}, \ldots, \beta_{jp})$ için $g_j(X) = \beta_j' X'$ eşitliğinden yararlanarak $j=1,..,g$ olmak üzere,

$$P_j(X) = P_j(Y=j/X) = \exp(\beta_j' X') / \sum_{k=0}^{g-1} \exp(\beta_k' X') \qquad (2.41)$$

şeklinde g gruba genellenmektedir (Lesaffre, 1986).

Fienberg (1979), iki açıklayıcı değişken durumunda (g-1) tane lojistik modelin 5 ayrı yazım türünü tanımlamıştır.

2.5.1. Tekrarlı veriler için çoklu grup lojistik modeller

Yanıt değişkeninin g kategorisinde gözlemler J grupta tekrarlı olduğunda lojistik modelin olasılıkları multinomial dağılımlıdır. Bu olasılıklar $j=1,\ldots,g$, $i=1,\ldots,J$ için,

$$P_{ji} = P(Y=j/X_i) = \exp(\beta_j' X_i') / B_i \qquad (2.42)$$

olup burada B_i,

$$B_i = \sum_{j=1}^{g} (\exp(\beta_j' X_i')) \qquad (2.43)$$

şeklindedir.

P_{ji} üzerinde lojistik dönüşüm ile çoklu grup lojistik model elde edilmekte olup bu modele ilişkin varsayımlar şöyledir:

i) p açıklayıcı değişkene ilişkin n gözlem olup $n \geq p$ dir.

ii) Yanıt değişkeni, $g>2$ olmak üzere her bir olası g sınıfa düşen gözlemlerin yanıt sayısı olarak ölçülür.

iii) Yanıt değişkenindeki tekrarlar, açıklayıcı değişkenlerin her bir düzeyi içerisinde incelendiğinde, oluşan gözeler içerisinde ve arasında bağımsızdır.

iv) Her X_i gözlemi için j grubunda olma olasılığı (2.42) formülü ile elde edilmektedir (Aldrich and Nelson, 1986).

2.5.2. Çoklu grup lojistik varsayımı

H_s grup üyeliği göstergesi, $f_s(X)$ $(s=1,..,g)$, gözlemlerin s'nci grup için olasılık yoğunluk fonksiyonu olduğunda çoklu grup lojistik varsayımı, $\beta_s'=(\beta_{0s}, \beta_{1s},..., \beta_{ps})$ s'nci olasılık yoğunluk fonksiyonuna ilişkin bilinmeyen parametreler vektörü olduğunda $s=1,..,g-1$ için,

$$f_s(X)/f_g(X)=\exp(\beta_s'X') \tag{2.44}$$

şeklinde verilmektedir. H_g sınıfı temel alındığında β_g=0'dır ve tüm sınıflar temel sınıf ile karşılaştırılırlar.

(2.44) eşitliğinin logaritmasından lojistik model,

$$\mathrm{Ln}\ f_s(X)/f_g(X)=\beta_s'X' \tag{2.45}$$

olarak elde edilir. H_g'den farklı bir sınıf, örneğin H_t temel alındığında bu model, $t,s=1,...,g(t\neq s)$ için,

$$\mathrm{Ln}\ f_s(X)/f_t(X)=(\beta_s-\beta_t)'X' \tag{2.46}$$

şeklini almaktadır. Lojistik varsayım burada da geçerlidir ancak model farklı bir parametre kümesine sahiptir (Lesaffre, 1986).

Çoklu grup lojistik varsayımın genellenmesi iki grup için (2.24) ile verilen forma benzer olarak yapılabilmektedir.

2.5.3. Lojistik ayrımsama

$H_1,..,H_g$ sınıflarından sıra ile $n_1,..,n_g$ gözlemler alındığında toplam $n=\sum_{s=1}^{g} n_s$ gözleme dayalı ayrımsama fonksiyonunu oluşturmak için $\beta_s(s=1,..,g-1)$ katsayı kestirimleri elde edilmelidir. Diskriminant analizinde olduğu gibi g grubu birbirinden ayırmak için (g-1) tane lojistik fonksiyon gerektiğinden toplam (g-1)(p+1) tane kestirim değeri elde edilmelidir. Kestirimler olabilirlik fonksiyonlarından elde edilmekte olup bu fonksiyonlar iki grup lojistik modellerde olduğu gibi örnekleme yöntemlerine göre değişmektedir. İki grup durumunda değinilen koşullu ve karışık örnekleme türleri için sonsal olasılıklar çoklu grup durumunda $s=1,..,g-1$ için,

$$P_s=P(H_s/X)=\exp(\beta_s'X')/\sum_{t=1}^{g}\exp(\beta_t'X') \qquad (2.47)$$

şeklinde tanımlanmaktadır. Burada H_g temel sınıftır. $\beta_s=(\beta^*_{0s},\beta_{1s},...,\beta_{ps})$ şeklinde olup, sabit terim için,

$$\exp(\beta^*_{0s})=\exp(\beta_{0s})(\tilde{P}_s/\tilde{P}_g) \qquad (2.48)$$

ilişkisi söz konusudur. $\tilde{P}_s$ ve $\tilde{P}_g$'ler sıra ile s ve g'nci gruplar için önsel olasılıklar olup, n_s ve n_g bu grupların gözlem sayısı ve $n=n_s+n_g$ olduğunda n_s/n ve n_g/n oranlarına eşittir. Ancak genelde karışık örnekleme için bu önsel olasılıklar eşit kabul edilerek $\exp(\beta^*_{0s})=\exp(\beta_{0s})$ alınmaktadır.

H_g ve H_t'nin temel sınıf alındığı durumlarda lojistik modeller (2.45) ve (2.46) eşitliklerine benzer olarak,

$$\text{Ln } P(H_s/X)/P(H_g/X)=\beta_s'X' \qquad (2.49)$$

ve

$$\text{Ln } P(H_s/X)/P(H_t/X)=(\beta_s-\beta_t)'X' \qquad (2.50)$$

şeklinde elde edilmektedirler.

Örnekleme durumlarına göre olabilirlik fonksiyonları iki grup lojistik modellere ilişkin fonksiyonların g gruba genellemesi olmaktadır.

Koşullu örnekleme durumunda n gözlemli örneklemin olabilirlik fonksiyonu, s(i), s'nci gruptaki i'nci gözlemi gösterdiğinde,

$$L_c = L_c(X,\beta) = \prod_{i=1}^{n} P(H_{s(i)}/X_i)$$

$$= \prod_{i=1}^{n} (\exp(\beta_{s(i)}{}'X_i'))/\sum_{t=1}^{g} \exp(\beta_t'X') \quad (2.51)$$

şeklinde tanımlanmaktadır. Bu fonksiyon,alternatif olarak,

$$L(X,\beta) = \prod_{i=1}^{n} \exp(\sum_{t=1}^{g} y_{ti} \operatorname{Ln} P(H_t/X_i)) \quad (2.52)$$

şeklinde de yazılabilmektedir. Burada y_{ti} i'nci gözlemin H_t grubunda olup olmamasına göre 1 ya da 0 değerini almaktadır. Karışık örnekleme türü için olabilirlik fonksiyonu (2.51) ile aynı olmaktadır (Lesaffre, 1986).

β_{sj}'nin EÇO kestirimi (2.51) ya da (2.52) ile verilen olabilirlik fonksiyonlarının maximizasyonu ile elde edilmektedir.

$\hat{\beta}$, β'nın EÇO kestirimi olduğunda çoklu grup lojistik varsayıma dayalı ayrımsama (sınıflandırma) kuralı, eğer yalnız ve yalnız tüm $s \neq t=1,..,g$ için,

$$P(H_s/X) = \max_{1 \leq t \leq g} P(H_t/X)$$

ya da (2.53)

$$(\hat{\beta}_s-\hat{\beta}_t)'X_i'>0$$

ise i'nci gözlemin H_s grubuna atanacağı şeklindedir. Burada s=1,..,g-1 için $\beta_s'X'$ vektörüne "ayrımsama skor vektörü" adı verilmektedir (Lesaffre, 1986).

2.6. Çoklu Grup Lojistik Modellerde Kestirim Yöntemleri

Çoklu grup lojistik modellerde kestirim yöntemleri iki grup durumunda uygulanan yöntemlerin genellemesidir. Bunlardan EÇO yöntemi iteratif türevlemeler gerektirdiğinden çoklu grup durumunda, çeşitli yöntemleri içeren iteratif yöntemler içerisinde incelenecektir. MLCS yönteminin yanı sıra EÇO kestirimlerinin Begg ve Gray'in (1984) ikili regresyon (individulized regression) modelleri yaklaşımı ile incelenmesi de konu edilecektir.

2.6.1. En çok olabilirlik yöntemi

Yanıt değişkeninin g düzeyli olduğu çoklu grup lojistik modelde açıklayıcı değişkenler matrisi n(g-1)x(g-1)(p+1) boyutludur. $X=(X_1',\ldots,X_n')$ için i'inci gözleme karşılık gelen (g-1)x(g-1)(p+1) boyutlu matris,

$$X_i' = \left|\begin{array}{cccccccccc} x_{oi} & \ldots & x_{pi} & 0 & \ldots & 0 & \ldots 0 & & \ldots & 0 \\ 0 & \ldots & 0 & x_{oi} & \ldots & x_{pi} & \ldots 0 & & & 0 \\ . & & & . & & . & & & & \\ . & & & . & & . & & & & \\ . & & & . & & . & & & & \\ 0 & \ldots & 0 & 0 & \ldots & 0 & \ldots x_{oi} & & \ldots & x_{pi} \end{array}\right| \qquad (2.54)$$

şeklindedir. Gözlenen ve beklenen değerlerin farkından elde edilen artık vektörü $r'=(r_1',\ldots,r_n')$ olup s=1,..,g-1 için $r_i=y_{si}-\hat{P}_{si}$ şeklinde tanımlanmaktadır. Burada $\hat{P}_{si}$ (2.47) eşitliği kullanılarak i'nci gözlem için elde edilen P_{si} olasılık değerinin kestirimidir. EÇO kestirimi, (2.51) ile verilen log-olabilirlik fonksiyonunun β'ya göre birinci türevi ile,

$$\frac{\partial LnL_c(X,\beta)}{\partial \beta} = r'X \qquad (2.55)$$

şeklinde bulunan skor vektörünün sıfıra eşitlenerek çözümünden elde edilmektedir.

İkinci derece türevler ise matris formunda,

$$\frac{\partial^2 LnL_c(X,\beta)}{\partial \beta' \partial \beta} = -(X'VX) \qquad (2.56)$$

olarak bulunur. Burada yarı-köşegen (quasi-diagonal) matris olan V, i'nci gözlem için,

$$V_i = \begin{vmatrix} P_1(1-P_1) & -P_1P_2 & \ldots & -P_1P_{g-1} \\ & P_2(1-P_2) & \ldots & -P_2P_{g-1} \\ S & & & \\ & & & P_{g-1}(1-P_{g-1}) \end{vmatrix} \qquad (2.57)$$

şeklinde olup (g-1)x(g-1) boyutludur. Burada $P_j=P(H_j/X_i)$ olasılık değeridir.

(2.56) ile verilen ikinci türevler matrisine bilgi matrisi (information matrix) denmektedir. EÇO kestirimlerinin varyans-kovaryans matrisi ise,

$$\omega(\hat{\beta})=E\left(-\frac{\partial^2 LnL_c(X,\beta)}{\partial \beta' \partial \beta}\right)=(X'VX)^{-1} \qquad (2.58)$$

olarak elde edilmektedir. Matris Y'ye bağlı olmadığı için

$$-E\left(\frac{LnL_c(X,\beta)}{\partial\beta'\partial\beta}\right)^{-1} = -\left(\frac{LnL_c(X,\beta)}{\partial\beta'\partial\beta}\right)^{-1}$$

$$=(X'VX)^{-1} \qquad (2.59)$$

eşitliği söz konusudur (Lesaffre, 1986).

2.6.1.a. yeniden ağırlıklandırılmış iteratif en küçük kareler yöntemleri

Ağırlıklandırmış yöntemler serisi olarak ele alınan Newton-Raphson ve Fisher-skorlama (Fisher-scoring) yöntemleri genel olarak RILS yöntemleri başlığı altında incelenmektedirler. Newton-Raphson yönteminde, olabilirlik fonksiyonunun t'nci adımdaki EÇO kestirimi β(t) kullanılarak t+1'nci adımdaki kestirim değeri,

$$\beta(t+1)=\beta(t)-\left(\frac{\partial^2 LnL(X,\beta)}{\partial\beta'\partial\beta}\right)^{-1}\Bigg|_{\beta=\beta(t)} \frac{\partial LnL(X,\beta)}{\partial\beta'\partial\beta}\Bigg|_{\beta=\beta(t)} \qquad (2.60)$$

eşitliğinden elde edilmektedir. Yöntemde iterasyon sayısı arttıkça ıraksama olabileceği gibi hızlı bir biçimde yakınsama da olabilmektedir. Yakınsama β'nın başlangıç kestirimine bağlıdır.

Fisher-skorlama yönteminin Newton-Raphson'dan farkı (2.60) eşitliğinde ikinci derece türevler matrisi yerine beklenen değerinin kullanılmasıdır. (2.59) ile verilen eşitlikten matrisin beklenen değerinin kendisine eşit olması nedeni ile bu iki yöntem aynıdır. Bu sonuç lojistik model dışında çeşitli model aileleri için de sağlamaktadır.

Böylece genel olarak RILS yöntemi, (2.60) eşitliğinde yer alan birinci ve ikinci türev değerleri yerine (2.55) ve (2.56) ile verilen değerleri konulduğunda, t+1'nci adımdaki kestirimi,

$$\beta(t+1)=\beta(t)+(X'V(t)X)^{-1}\ X'r(t) \qquad (2.61)$$

olarak elde etmektedir. Burada $V(t)=V(\beta(t))$, $r(t)$ ise $\beta(t)$'de bulunan artık vektörüdür.

İlk kez iki grup lojistik model için Walker ve Duncan (1967) tarafından önerilen yöntemin, Newton Raphson ve Fisher-skorlama yöntemleri ile aynı olduğu söylenebilir (Lesaffre, 1986).

Yöntemlerin yakınsamasında başlangıç değerleri ve örnekleme türü önemli rol oynamaktadır. Başlangıç değeri olarak sıfır değerinin alınmasının yanı sıra klasik diskriminant analizi katsayıları da kullanılabilmekte ve daha çabuk yakınsama sağlayabilmektedir. Örnekleme türleri içerisinde ayrı örnekleme söz konusu olduğunda ise sabit terim kestirimi ve kestirimin varyansı üzerinde ayarlama yapmak gerekmektedir (Lesaffre, 1986).

2.6.2. Minimum lojit ki-kare yöntemi

Veriler J grupta tekrar edildiğinde ve her grupta tekrar sayısı çok olduğunda katsayı kestirimleri ağırlıklı EKK ya da MLCS yöntemleri ile elde edilebilmektedir.

Çoklu grup lojistik modelde, (g-1) tane iki grup lojistik modellerin her biri için 2.4.3 alt bölümünde değinilen MLCS yöntemi ayrı ayrı uygulanmaktadır. Ancak burada incelenen modellerin hata terimleri ilişkili olacaktır. İlişkinin nedeni,modellerin P_{ij} olasılıklarına dayanması ve bu olasılıkların toplamının 1 olması zorunluğudur. Bu ilişkileri dikkate alan kestirim yöntemleri daha iyi kestirimler vermekle birlikte daha karmaşıktırlar (Aldrich and Nelson, 1986)

2.6.3. Çoklu grup lojistik modellerde en çok olabilirlik kestirimlerinin varlığı

Çoklu grup lojistik modellerin EÇO kestirimleri iteratif olarak hesaplanırken bazen tek bir maximum değere ulaşılamamaktadır. Bu durumu yaratan veri konfigürasyonlarının belirlenmesi optimizasyon sürecinde gereksiz iterasyonları önlemesi açısından gerekmektedir. Albert ve Anderson (1984), bu sorunu ele almış ve EÇO kestirimlerinin varlığına ilişkin teoremleri ispatlamışlardır. Üç grupta topladıkları olası veri konfigürasyonları şöyledir:

i) Tam/Yarı-tam bölünme (complete/quasi-complete separation)

ii) Kısmi/Yarı-kısmi bölünme (partial/quasi-partial separation)

iii) Taşma (overlap) durumu

Albert ve Lesaffre (1986) bu veri konfigürasyonları için EÇO kestirimlerinin varlık ve tekliğine ilişkin teoremleri incelemişlerdir. Varlık, EÇO kestiriminin sonlu olduğu anlamında kullanılmaktadır. Böylece EÇO kestiriminin olmaması sonlu maximumun elde edilememesinin bir göstergesidir. Bu veri konfigürasyonları izleyen alt bölümlerde ele alınmaktadır.

2.6.3.a. tam ve yarı-tam bölünme

Lojistik varsayıma dayalı sınıflandırma kuralı (2.53) eşitliğinde verilmektedir. Bu kurala dayanarak Albert ve Anderson'a (1984) göre $j,k=1,..,g(j\neq k)$ ve $i=1,..,n$ için,

$$(\beta_j-\beta_k)'X_i'>0 \qquad (2.62)$$

ise $H_1,...,H_g$ gruplarını oluşturan gözlemler "tam bölünmüştür" (Lesaffre, 1986). Bu veri konfigürasyonu için $\sup_\beta Ln\ L(X,\beta)=0$ olup β'nın EÇO kestirimi sonlu değildir.

Eğer (2.62) ile verilen eşitsizlik en az bir i,j,k ölçüsü için eşitlik halinde ise ($\geq$) örneklemdeki gözlemler "yarı-tam bölünmüştür" denmektedir. Bu bölünme durumu için de $\sup_\beta \mathrm{Ln}\ L(X,\beta)<0$ olup EÇO kestirimi sonlu değildir. Silvapulle (1981), tam ve yarı-tam bölünme durumunda EÇO kestirimlerine ilişkin teoremleri iki grup lojistik modeller için incelemiştir.

EÇO kestirimlerinin iteratif olarak maximizasyonunda ıraksama yaratan bu veri konfigürasyonlarını belirlemek için Albert ve Anderson (1984), şöyle bir yaklaşım önermişlerdir: Her iterasyonda veri kümesi içersinde doğru atanmış en yüksek sonsal olasılık değerine sahip gözlem aranır. Bu olasılığın $(1-\varepsilon)$ gibi bir sınırı aşması, örneklemde en az bir gözlem için doğru sınıflandırma olasılığının 1'e çok yaklaştığının göstergesidir. Böyle bir durum iki nedenden kaynaklanabilmektedir. Ya bu gözlem aykırıdır ki bu durumda iterasyon maximuma ulaştığında durur ya da veri kümesinde yarı-tam bölünme vardır. Bu durumda gözlem tam ayrılanlar arasındadır ve asimtotik varyans-kovaryans matrisinin öğeleri çok büyüktür. Neden yarı-tam bölünme olduğunda önerilen çözüm standartlaştırılmış veriler üzerinden lojistik modelin tekrar araştırılması ve iteratif kestirimler sırasında varyans-kovaryans matrisinin herhangi bir köşegen ögesi 10^3 değerini aştığı adımdaki kestirimlerin alınması şeklindedir. Yaklaşım, değişkenler arasında çoklubağlantı (multicollinearity) olmadıkça daima yarı-tam bölünmeyi belirlemektedir (Albert and Lesaffre, 1986).

2.6.3.b. kısmi ve yarı-kısmi bölünme

Eğer verilerdeki g grup (2.62) ile verildiği anlamda tam bölünmüş q tane $C_1,\ldots,C_q$ kümelerine gruplanabiliyorsa $(q<g)$ örneklemdeki gözlemler "kısmi bölünmüştür" denmektedir ve $v=(p+1)(q-1)$ için her bir C_s grubundaki katsayı kestirimleri $\gamma \in R^v$ olup $s,t=1,\ldots,q$ $(s\neq t)$ tüm gruplar üzerinden, $i=1,\ldots,n$ için,

$$(\gamma_s-\gamma_t)'X_i'>0 \qquad (2.63)$$

eşitliğini sağlamaktadır (Lesaffre and Albert, 1989b). Bu koşulu sağlayan $C_1,...,C_q$ kümelerinin yanı sıra $D_1,...,D_q$ gibi kümeler de bulunabilmektedir.

Tam bölünme kısmi bölünmenin q=g olan özel bir durumudur. Böylece bir yandan yarı-tam bölünmenin özel bir durumu olan kısmi bölünme diğer yandan tam bölünmenin de genellemesidir. Bu nedenle Albert ve Anderson'un teoremine göre β'nın EÇO kestirimi sonlu değildir (Lesaffre, 1986). Albert ve Lesaffre (1986), ne tam ne de yarı-tam bölünme olmayıp ara bir durum olarak adlandırdıkları kısmi bölünmeye bir örnek olarak üç grup durumunda iki grubun birbiri üzerine taştığı ve üçüncü grubun bu iki gruptan tamamen ayrıldığı durumu göstermişlerdir.

Eğer verilerde var olan g grup yarı-tam bölünmüş olan $C_1,...,C_q$ gibi q tane kümede gruplandırılabiliyorsa gözlemlerin $C_1,...,C_q$ kümelerinde yarı-kısmi bölündükleri söylenmektedir. Bu durumda (2.63) eşitsizliğinde en az bir (i,s,t) üçlüsü için eşitlik (≧) söz konusudur. Kısmi bölünmede olduğu gibi bu koşulu sağlayan $C_1,...,C_q$ kümeleri tek değildir.

Yarı-kısmi bölünme bir örneklem özelliğidir ve yarı-tam bölünmenin genellemesi olarak yorumlanabilmektedir. Böylece bu bölünme türü için de EÇO kestirimleri sonlu değildir (Lesaffre, 1986).

2.6.3.c. taşma durumu

Veri kümesinde ne tam/yarı-tam ne de kısmi/yarı-kısmi bölünme yoksa muhtemelen taşma durumu vardır. Bu durumda j,k=1,...,g (j≠k) ve i=1,...,n için,

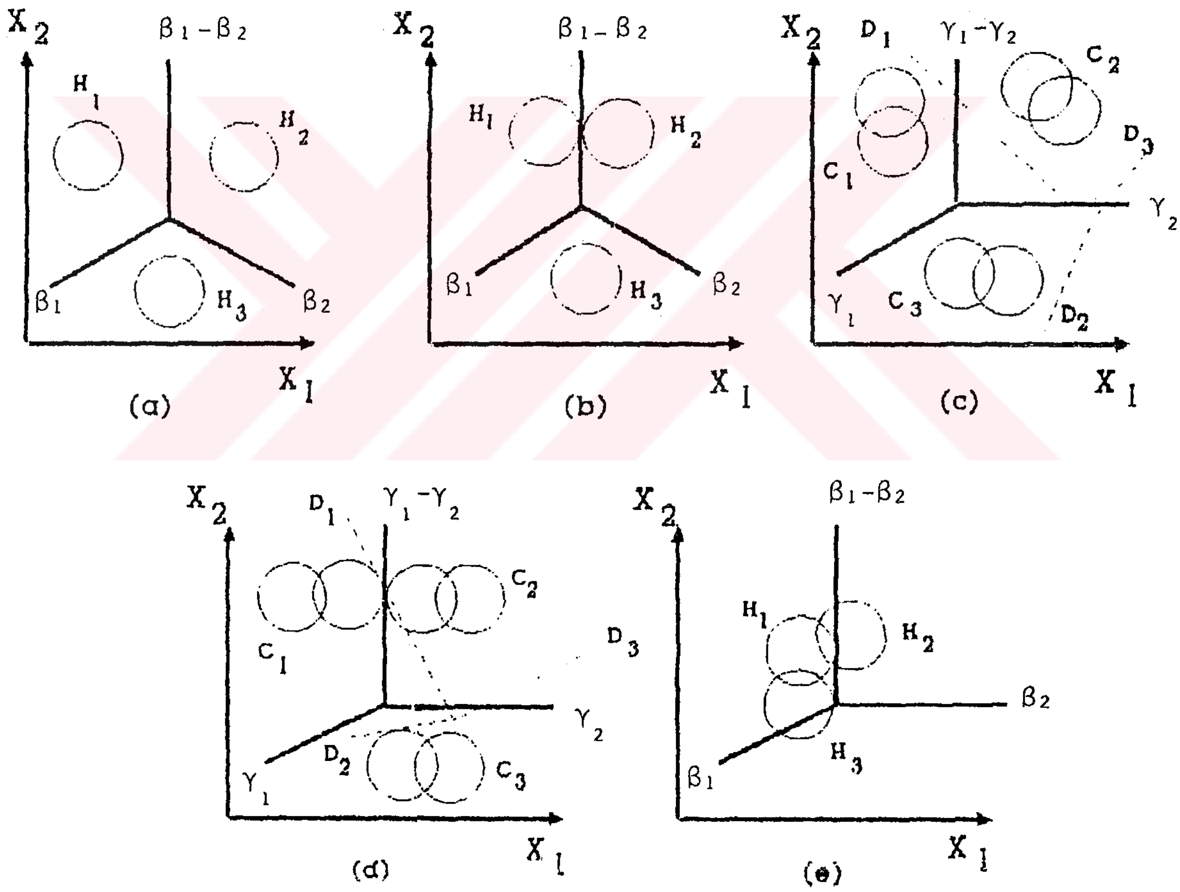
$$(\beta_j-\beta_k)'X_i'<0 \qquad (2.64)$$

eşitsizliği söz konusudur (Lesaffre, 1986).

Taşma durumu olması için tüm grupların birbiri üzerine taşması gerekmemektedir. Örneğin üç grup olduğunda, farklı iki grubun üçüncü grup üzerine taşması da bu durumun bir göstergesidir.

Lojistik varsayımlar altında β'nın EÇO kestirimi yalnız ve yalnız veri kümesinde taşma olduğunda vardır ve tektir (Albert and Anderson, 1984).

Yukarıda incelenen veri konfigürasyonlarının grafiksel gösterimi aşağıda Şekil 2.2 ile verilmektedir.



Şekil 2.2. Verilerin çeşitli bölünme durumları: a) tam bölünme, b) yarı-tam bölünme, c) kısmi bölünme, d) yarı-kısmi bölünme, e) taşma (Lesaffre, 1986).

Bu şekilde iki değişken ve üç grup ele alınmaktadır. (c) ve (d) çizimlerinde alternatif kümeler olan D_1, D_2, D_3 kırık çizgilerle verilmektedir.

2.6.3.d. bölünme türünün belirlenmesi

Anderson (1972), lojistik modelin olabilirlik fonksiyonunun maximum değeri 0.5'den büyük olduğunda tam bölünmenin söz konusu olduğunu, bu kuralın kısmi bölünme için geçerli olmadığını vurgulamıştır.

Tam, yarı-tam ve yarı-kısmi bölünme durumları küçük örneklem problemleridir. Kısmi bölünmenin ise örneklem genişliği ile ilgisi olmayıp küçük örneklemlerde özellikle medikal uygulamalarda sık görülmektedir (Albert and Lesaffre, 1986). Yanıt değişkeni iki düzeyli olduğunda (g=2) bu dört bölünme türü birbirinin aynı olacaktır.

Bu bilgiler doğrultusunda çoklu grup karar modelleri içerisinde yer alan çoklu grup lojistik modellerde bilinmeyen β katsayılarının EÇO kestirimlerinin hesaplanması pratikte anlamsız görülmektedir. Oysa bu doğru değildir. Tam, yarı-tam ve yarı-kısmi bölünme durumlarına pratikte pek rastlanmamaktadır. Kısmi bölünme ise çok görülen bir durumdur. Burada tanımlananların yanı sıra oluşabilecek diğer veri konfigürasyonları için de EÇO kestirimlerinin sonlu olmaması istenmeyen bir özellik değildir. Örneğin tam bölünme durumunda %100 doğru sınıflandırma sağlayan sonsuz sayıda EÇO kestirimi elde edilebilmektedir. Bu duruma diskriminant analizi gibi diğer çoklu grup karar modellerinde rastlanmayabilinir.

Sonuçta, çeşitli veri konfigürasyonları için EÇO kestirimlerinin sonlu olmamasına rağmen iyi ayrımsama sağlaması, bu kestirimlerin kullanımı için yeterli bir koşuldur (Lesaffre, 1986).

2.6.4. En çok olabilirlik kestirimlerine Begg ve Gray yaklaşımı

Çoklu grup lojistik modelin katsayılarının EÇO kestirimlerini, büyük örneklemlerde ve değişken sayının çok olduğu durumlarda, bilgisayar belleği ve zaman kısıtı nedeni ile elde etmek güçleşmektedir. Ayrıca yukarıda değinilen veri konfigürasyonları için de sonlu kestirimler elde edilememektedir. Bu durumda işlemlerde kolaylık sağlamak amacı ile Begg ve Gray (1984), yanıt değişkeninin her bir düzeyini temel alınan sınıfla bireysel olarak iki grup lojistik modellerle karşılaştıran ikili lojistik regresyon analiz yöntemini önermişlerdir. Yöntemin işlem yükü daha az olup değişken seçimi uygulaması daha kolaydır ve standart paket programları vardır. (g-1) tane iki grup lojistik modelin incelenmesi temeline dayanan yöntem şöyle açıklanmaktadır:

g düzeyli yanıt değişkeninin g'nci düzeyi temel sınıf alındığında ve z_{di} gösterge değişkeni i'nci gözlem d'nci grupta iken 1 aksi takdirde 0 değerini aldığında d=1,..,g için değişkenin 1 değerini alma olasılığı,

$$P_{di}=P(z_{di}=1/X_i) \qquad (2.65)$$

şeklinde tanımlanmaktadır. Çoklu grup lojistik model, d=1,...,g-1 için,

$$\text{Ln } P_{di}/P_{gi}=\beta_d'X_i' \qquad (2.66)$$

varsayımına dayanmaktadır. Burada $\beta_d'=(\beta_{d_0},..,\beta_{dp})$ katsayılar vektörüdür.

Her bir d sınıfını g temel sınıfı ile karşılaştıran ikili lojistik regresyon yönteminde ise olasılıklar,

$$\Theta_{di}=P(z_{di}=1/X_i,\ z_{gi}+z_{di}=1)$$
$$\Theta_{gi}=P(z_{di}=0/X_i,\ z_{gi}+z_{di}=1) \qquad (2.67)$$

şeklinde tanımlanmakta olup lojistik modeller $d=1,..,g-1$ için,

$$\mathrm{Ln}\ \Theta_{di}/\Theta_{gi}=\tilde{\beta}_d'\ X_i' \qquad (2.68)$$

olarak elde edilmektedir. Burada $\tilde{\beta}_d$,(p+1) boyutlu bilinmeyen katsayılar vektörüdür. (2.66) ve (2.68) ile verilen modeller parametrik olarak eşittir. Böylece $d=1,..,g-1$ için $\tilde{\beta}_d=\beta_d$ yazılabilmektedir. Bayes teoreminden yararlanarak,

$$\Theta_{di}=P_{di}/(P_{gi}+P_{di}) \qquad (2.69)$$

eşitliği elde edilebilmekte ve $P_{di}/P_{gi}=\Theta_{di}/(1-\Theta_{di})$ yazılabilmektedir.

İkili lojistik regresyon yönteminde (g-1) tane lojistik modelin olabilirlik fonksiyonları ayrı ayrı maximize edilmektedir. Her birinin olabilirlik fonksiyonu $d=1...,g-1$ için,

$$L_d=\prod_{c_d}\ \Theta_{di}^{z_{di}}\ (1-\Theta_{di})^{z_{gi}} \qquad (2.70)$$

şeklinde verilmektedir. Burada $c_d=\{i:z_{gi}+z_{di}=1\}$ olarak tanımlanmaktadır. Katsayıların EÇO kestirimi,

$$\frac{\partial \mathrm{Ln}L_d}{\partial \beta_{dj}}=\sum_{i=1}^{n} x_{ij}(z_{di}-(z_{di}+z_{gi})\Theta_{di}) \qquad (2.71)$$

şeklindeki skor değerinin sıfıra eşitlenerek çözümünden elde edilmektedir. $\tilde{\beta}_d$ $(d=1,..,g-1)$ ile gösterilen kestirim değerleri (2.66) eşitliğindeki olasılık kestirimlerini elde etmek için kullanılabilmektedirler.

Begg ve Gray (1984), asimtotik olarak yansız olan $\tilde{\beta}_d$ kestirimlerini, çoklu grup lojistik model kestirimleri (β_d) ile karşılaştırarak etkinliklerini araştırmışlar, her iki kestirimin de (varyansları farklı da olsa) normal dağıldığını göstermişlerdir. Yaptıkları simülasyon çalışmasında $\tilde{\beta}$ kestirimlerini tutarlı ve etkin bulmuşlardır. Açıklayıcı değişken (p) ve grup sayısı (g) artarsa etkinliğin azalacağını ve H_g temel sınıfının genişliğinin etkinlik üzerinde etkili olduğunu belirtmişlerdir (Lesaffre, 1986).

İkili lojistik modellerin ayrımsama performansı çoklu grup lojistik model kadar iyi olup seçilen temel sınıfa, değişkenlerin marjinal dağılımına, grup ve değişken sayısına göre değişmektedir. Ancak çıkarsama yaparken, çoklu grup lojistik modellerin tercih edilmesi önerilmektedir (Hosmer and Lemeshow, 1989b).

İkili lojistik regresyon analizinde temel sınıfın 2.7.3 alt bölümünde değinildiği gibi uygun olarak seçilmesi yöntemin etkinliğini artırabilmektedir. Uygun bir sınıf bulunamadığı taktirde en geçerli sınıf temel alınmaktadır. Öte yandan tam bölünme durumunda yöntem çoklu grup lojistik regresyon analizine göre daha düşük etkinlikte kestirimler vermektedir (Lesaffre, 1986).

2.7. Çoklu Grup Lojistik Modellerde Uyum İyiliği Testleri, Değişken Seçimi ve Temel Sınıfın Belirlenmesi

Çoklu grup lojistik modeller için (2.44) eşitliği ile tanımlanan lojistik varsayım çok değişkenli dağılımların geniş bir sınıfı için sağlasa da 2.4.6 alt bölümünde iki grup lojistik modeller için değinilen varsayım bozulumları burada da söz konusu edilebilmektedir. Bu durumda elde edilen modelin geçerliliği yine aynı alt bölümde değinilen ölçütleri, çoklu gruba genelleyerek sınanabilmektir. İzleyen alt bölümlerde uyum iyiliği testleri, karışık örnekleme durumunda değişken seçimi ve temel sınıfın seçim yolları üzerinde durulmaktadır.

2.7.1. Uyum iyiliği testleri

En temel uyum iyiliği ölçütleri olan ki-kare ve sapma, g grup için sıra ile,

$$X^2=\sum_{i=1}^{n} X_i^2=\sum_{i=1}^{n}\sum_{t=1}^{g}(y_{ti}-\hat{P}_{ti})^2/\hat{P}_{ti}(1-\hat{P}_{ti}) \qquad (2.72)$$

$$D=-2\sum_{i=1}^{n} d_i^2=-2\sum_{i=1}^{n}\sum_{t=1}^{g} y_{ti}\mathrm{Ln}\hat{P}_{ti} \qquad (2.73)$$

şeklinde genellenmektedir. Burada $\hat{P}_{ti}=P(H_t/X_i)$ olup (2.47) eşitliğinde j yerine t konarak i'nci gözlem için elde edilen olasılık değerinin kestirimidir. y_{ti} ise i'nci gözlem t'nci grupta ise 1, aksi taktirde 0 değerini almaktadır. Ölçütlerin yorumları iki grup lojistik model incelemesinde verildiği gibidir. Ancak D ölçütü burada n-p(g-1) serbestlik dereceli ki-kare dağılımı göstermektedir.

Çoklu grup lojistik modelde gerek tüm değişkenlere gerekse bunların bir alt kümesine ilişkin kurulan hipotezler, 2.4.6 alt bölümünde, iki grup lojistik modeller için önerilen uyum iyiliği ölçütleri g grup için genellenerek test edilebilmektedirler.

Lesaffre (1986), uyum iyiliği testi için önerilen ölçütlerden modellerin olabilirlik fonksiyonlarına dayanan olabilirlik oran ölçütünü (LR), modeldeki açıklayıcı değişkenlerin katsayılarının anlamlılığını test eden Wald ölçütü (W) ile skor ölçütlerini (R) ve Begg ve Gray'in yaklaşımı ile verilen test ölçütlerini incelemiştir. Bartalucci ve Frazer (1977), yaptıkları simülasyon çalışmasında bu ölçütleri birbirleri ile karşılaştırmışlardır.

Çoklu grup lojistik modelde yer alan katsayıların anlamlılığı yukarıda değinilen ölçütlerden biri ile sınandıktan sonra karar verilen son modele etkileşim terimleri eklenerek bunların etkileri de araştırılabilmektedir. Ancak burada karşılaşılan bir sorun açıklayıcı değişkenlerin tüm (g-1) model için değilde bazıları için önemli olabilmesidir. Değişken sayısı minimum olacak şekilde modelleme yapılıyorsa bazı katsayıları sıfır olması için zorlanacak ve kestirimleri diğer modeller için elde edilecektir (Hosmer and Lemeshow, 1989b).

Lesaffre'in (1986) ele aldığı Begg ve Gray yaklaşımı için uyum iyiliği testleri, H_g temel sınıfı ile karşılaştırılan her (g-1) lojistik model için ayrı ayrı yapılabilmektedir. 2.4.6 alt bölümünde değinilen tüm uyum iyiliği ölçütlerinin uygulanabildiği bu yaklaşımın bazı dezavantajları söz konusudur: Örneğin her bir ikili lojistik model için Hosmer-Lemeshow ölçütünün kullanımı birinci tür hatanın belirlenmesini güçleştirmektedir. Ayrıca herhangi ikili lojistik model için aykırı ya da etkin olan gözlemler tüm model üzerinde aykırı ya da etkin olmayabilmektedir (Albert and Lesaffre, 1986).

2.7.2. Değişken seçimi

Çoklu grup lojistik modellerde değişkenlerin modele seçimi ya da modelden çıkartılması için 2.7.1 alt bölümünde değinilen LR, W, R ölçütlerinden biri kullanılabilmektedir.

Modele yeni bir değişkenin seçilmesi (çıkartılması), bu yeni değişkenin varlığı (yokluğu), (g-1) ekstra katsayının eklenmesi (çıkartılması) anlamına gelmektedir. Alternatif olarak önemli değişkenler yerine önemli katsayılar seçilebilmektedir (çıkartılabilmektedir). Bu tür bir seçim $s=1,..,g$ ve $\beta_g=0$ için,

$$P(H_s/X)=\exp(\beta_s'X_s')/\sum_{t=1}^{g}\exp(\beta_t'X_t') \qquad (2.74)$$

olmak üzere

$$\mathrm{Ln}\ P(H_s/X)/P(H_g/X)=\beta_s'X_s' \qquad (2.75)$$

şeklinde tanımlanan lojistik modele götürmektedir.

Değişken seçim yöntemleri küçük ve büyük örneklem genişliklerine göre değişmektedir.

Küçük örneklemlerde (2.49) ve (2.75) ile tanımlanan lojistik modeller için doğrusal ya da doğrusal olmayan regresyondaki klasik değişken seçimi yöntemlerinden herhangi biri uygulanabilmektedir. Seçim stratejisi daima k ve k' parametre sayıları olmak üzere M_k ve $M_{k'}$ gibi iki istatistiksel modeli karşılaştırır ve en iyi uyum sağlayan modeli seçer. Aşamalı modeller için $M_k \subset M_{k'}$ dür ($k'=k+r$, $r \geq 0$).

Lesaffre (1986), lojistik regresyon analizinde değişken seçim yöntemleri olarak doğrusal regresyon analizinde temel olan ileriye doğru seçim (forward selection), geriye doğru çıkarma (backward elimination), adımsal seçim (stepwise selection) ve tüm olası alt kümeler seçim yöntemlerini (all subsets selection) incelemiştir.

Hosmer ve arkadaşları (1989), yeni bir seçim algoritması geliştirmişler, Miller (1984) ise doğrusal regresyon modeli için geliştirdiği değişken seçim algoritmasının lojistik modele uygulanabileceğini vurgulamıştır.

Bunların yanı sıra bazı araştırıcılar orjinal değişkenler yerine çıkarsanan (hipotetik) değişkenleri kullanarak değişken seçimine gidilmesini önermişlerdir. Bunlardan

D'Agostino ve Pozen (1982), değişkenlerin kümelenmesini takiben adımsal seçim yönteminin uygulanmasını önermektedirler. İlk aşama olan kümeleme işlemi istatistiksel yöntemlere ve mantıksal nedenlere dayanmaktadır. Bu da aynı olguyu temsil eden değişkenlerin bir küme oluşturması demektir. İkinci aşamada her kümedeki birinci temel bileşenler üzerinde standart seçim yöntemi uygulanmaktadır (Lesaffre, 1986).

Örneklem genişliğinin çok büyük olduğu durumlarda yukarıda değinilen yöntemlerden hiç biri kullanılamamaktadır. Bu durumda değişken seçimi uygulamanın bir yolu veri kümesini rastgele alt gruplara bölmek ve bunların her birine standart seçim yöntemlerinden birini uygulamaktır. Değişken sayısı az olduğunda yöntem seçilen kümede tekrar edilebilmektedir.

Örneklem büyüklüğünü azaltmanın bir diğer yolu olası tüm ikili grup karşılaştırmalarına standart seçim yöntemlerinden birini uygulamaktır. Bu yöntem Begg ve Gray'ın yaklaşımını kullanarak kolaylaştırılabilmektedir. Bu yolla g-1 karşılaştırma için değişken seçimi uygulanmaktadır (Lesaffre, 1986).

Öte yandan temel seçim yöntemlerini karşılaştırmak amacı ile çeşitli çalışmalar yapılmıştır. Bunlardan birinde Berk (1978), yalnızca ileriye doğru seçim ve geriye doğru çıkarma yöntemlerinin tüm olası alt kümeler seçim yönteminden çok farklı olduklarında aynı sonuca götüreceklerini göstermiştir. Ayrıca büyük örneklem durumunda ileriye doğru seçim yönteminin, tüm olası alt kümeler seçim yöntemine yakın sonuçlar vereceğini belirtmiştir.

Yukarıda değinilen yöntemlerden biri ile açıklayıcı değişkenlerin bir alt kümesini seçmek ve katsayı kestirimlerini aynı verileri kullanarak yapmak kestirimlerin yanlı olmasına neden olacaktır. Miller (1984), bu yan'ı, ihmal (ommission bias), rekabet (competation or selection bias) ve durdurma kuralı (stopping rule bias) olmak üzere üç grupta incelemiştir. Yan sorununa şu anda getirilen tek çözüm örneklemi deney ve test gruplarına bölerek değişken seçimini deney, katsayı kestirimlerini de test kümesinde yapmaktır. Ancak bu durumda bilgi kaybı büyük olacağı için, büyük örneklem durumlarında uygulanabilecek bir yaklaşımdır (Lesaffre, 1986).

2.7.3. Temel sınıfın seçimi

Temel sınıfın herhangi bir sınıf olarak her bir seçimi ayrı bir lojistik modeli temsil etmektedir. H_g sınıfı temel alındığında lojistik varsayım sağlıyorsa, $H_t(t \neq g)$ gibi farklı bir sınıf temel alındığında da sağlayacaktır ve lojistik model bu durumda (2.50) eşitliğinde verildiği gibi tanımlanmaktadır. Bu model ile incelemeye başlandığında ve katsayısı sıfır olan değişkenler çıkartıldığında (2.75) ile tanımlanan daha basit modellere ulaşılabilinmektedir. Böylece temel sınıfın uygun seçimi daha iyi lojistik modellere götürmektedir.

Eğer herhangi bir $H_s(s=1,..,g)$ sınıfı temel alındığında modelde minimum sayıda parametre bulunuyorsa, bu sınıf temel olarak alınmaya en uygun olanıdır ve "minimal temel sınıf (minimal base class)" adını almaktadır. Ancak minimum sayıda parametre içeren modelin belirlenmesi önemli olmakla beraber çok zordur, çünkü pratikte böyle bir model tanımlanamamaktadır. Bununla beraber bu amaca yönelik olarak ileri sürülen çeşitli yaklaşımlar söz konusudur. Yaklaşımlardan birinde (2.75) ile tanımlanan modele temel sınıfın her seçimi için ileriye (geriye) doğru seçim (çıkarma) yöntemleri uygulayıp, en az değişken içeren modelin sonuç model olarak alınması önerilmektedir. Bu yaklaşım büyük bir örneklem durumunda çok zaman alıcıdır.

Bir diğer yaklaşım da, $l_s-p(s)$ $(s=1,..,g)$ şeklinde tanımlanan Akeike bilgi kriterini (Akeike's information criterion) kullanarak bu kriteri maximum olan modeli seçmektir. Burada l_s ve $p(s)$ sıra ile H_s'nin temel alındığı modelin log-olabilirlik değeri ve parametre sayısıdır. Bu yaklaşım büyük örneklemlere uygulanabilen basit bir yaklaşımdır. Ancak değişkenlere ayrı ayrı uygulandığı için bazen iyi sonuç veremeyebilmektedir. Bu durumda değişkenlerin çıkartılması için yanılma düzeyinin daha tutucu (conservative) bir değer olarak $\alpha=0.10$ veya 0.15 alınması önerilmektedir (Lesaffre, 1986).

2.8. Diskriminant Analizi Yöntemi

Gözlemleri veri kümesinin yapısında varolan gruplardan birisine sınıflandırma amacına yönelik çok değişkenli istatistiksel yöntemlerden birisi de diskriminant analizi olup uygulamada çok sık kullanılmaktadır.

İlk kez Fisher tarafından varyans analizinde gruplararası ayrımsamayı maximize etmek amacı ile iki grup için önerilen diskriminant analizi yönteminin temel varsayımları, değişkenlerin çok değişkenli normal dağılması ve grupların ortak varyans-kovaryans matrisli olmalarıdır.

Veriler iki grupta (g=2) toplandığında Fisher doğrusal diskriminant fonksiyonu (FLDF),

$$(\bar{X}^{(1)}-\bar{X}^{(2)})' S^{-1} X$$

şeklinde iki grubu mümkün olduğunca ayıran p değişkenin doğrusal kombinasyonu olarak önermiştir. Burada $\bar{X}^{(1)}$ ve $\bar{X}^{(2)}$, n_1 ve n_2 genişliğinde iki bağımsız grubun ortalama vektörü, S ise j=1,..,p için,

$$S=\sum_{g=1}^{2}\sum_{i=1}^{n_g}\frac{(x_{ji}^{(g)}-\bar{X}_j^{(g)})(x_{j'i}^{(g)}-\bar{X}_{j'}^{(g)})}{n_1+n_2-2} \qquad (2.76)$$

olarak tanımlanan toplanmış örneklem varyans-kovaryans matrisidir (Gnanadesikan, et all., 1989).

Bu fonksiyon,

$$\ell=b_1X_1+\ldots+b_pX_p \qquad (2.77)$$

şeklinde de gösterilmekte olup gözlemleri gruplardan birine atamada kullanılmaktadır. Atama için sınır değer (2.77)'den yararlanarak en basit yaklaşımla $(\overline{\ell}_1+\overline{\ell}_2)/2$ olarak elde edilmektedir.

Burada,

$$\overline{\ell}_1 = b_1\overline{x}_{11} + \ldots + b_p\overline{x}_{1p}$$

$$\overline{\ell}_2 = b_1\overline{x}_{21} + \ldots + b_p\overline{x}_{2p} \tag{2.78}$$

şeklinde tanımlanmaktadır. Böylece FLDF $b_0=-(\overline{\ell}_1+\overline{\ell}_2)/2$ olmak üzere,

$$\ell^{*} = b_0 + b_1X_1 + \ldots + b_pX_p \tag{2.79}$$

şeklinde elde edilmekte ve $\ell^{*}>0$ ise gözlem H_1 aksi taktirde H_2 grubuna atanmaktadır.

FLDF yukarıda değinilen varsayımlar bozulduğunda yeterli olmayacaktır. Bu durumda çeşitli alternatif fonksiyonlar önerilmektedir. Ortak varyans-kovaryans matrisi varsayımı bozulduğunda en iyi doğrusal diskriminant fonksiyonu (best linear discriminant function, BLDF), değişkenlerin karesel etkisi söz konusu olduğunda ise karesel diskriminant fonksiyonunun (quadratic discriminant function, QDF) kullanımı önerilmektedir (Gnanadesikan, et all., 1989).

BLDF yaklaşımı ilk kez Clunies ve Riffenburg (1960) tarafından tanımlanmış, Anderson ve Bahadur (1962) tarafından geliştirilmiştir. Marks ve Jean (1974), Gnanadesikan ve arkadaşları (1989), çeşitli varsayım bozulumları durumunda bu üç diskriminant fonksiyonunu karşılaştırmışlardır. Knoke (1982), bu fonksiyonlara ek olarak daha yüksek dereceden diskriminant fonksiyonlarını da incelemiştir.

2.8.1. Diskriminant ve lojistik regresyon analiz yöntemlerinin karşılaştırılması

Diskriminant analizinde varsayımlar bozulduğunda yukarıda değinilen diskriminant fonksiyonlarının yanı sıra son yıllarda çok sık kullanılan bir yöntemin de lojistik regresyon olduğundan söz edilmişti. Yöntem açıklayıcı değişkenlerin kesikli olduğu durumda da uygulanabildiği için diskriminant analizinden daha güçlü kestirimler vermekte dolayısı ile daha iyi modele götürebilmektedir (Knoke, 1982).

(2.79) eşitliği ile özetlenen diskriminant fonksiyonuna ilişkin olasılık değeri,

$$P(\ell^{*})=1/(1+e^{-\ell^{*}}) \qquad (2.80)$$

şeklinde tanımlanmakta olup bu değer lojistik modele ilişkin olasılık değeridir (Kleinbaum, et all., 1988).

İki yöntem arasındaki ilişki iki grup durumunda şöyle açıklanabilmektedir: Diskriminant analizi, Y koşulu altında P(X/Y=1) olasılığını X'lerin dağılım fonksiyonu yardımı ile tanımlamaktadır.

Bayes teoreminden,

$$P(Y=1/X)=P(Y=1)\ \frac{P(X/Y=1)}{P(X/Y=1)+P(X/Y=0)} \qquad (2.81)$$

yazılabilmektedir. Burada paydadaki olasılıklar, ortak varyans-kovaryans matrisli ve normal dağılımlı verilerden elde edildiğinde diskriminant analizine karşılık gelmektedir. P(Y=1/X) koşullu olasılığı ise lojistik model olasılıklarını tanımlamaktadır. Böylece lojistik regresyon analizi diskriminant analizinin aksine X koşulu altında Y'lerin dağılım fonksiyonunu kullanarak bu olasılıkları elde etmektedir. Eğer ortalama vektörleri ve varyans-kovaryans matrisleri biliniyorsa (2.81) eşitliği

(2.80)'e karşılık gelmektedir (Dietz, 1987).

Diskriminant analizi varsayım bozulumlarında yanlı kestirimler vermektedir. Bu durumda,alternatif olan lojistik regresyon analizi ile elde edilen kestirimler yeterlidir. Ayrıca diskriminant analizi ile gözlemlerin maskeleme etkileri gözlenemez iken lojistik modelde bu durum beklenmeyen,normalin dışında kestirimlerle kendisini göstermektedir (Press and Wilson, 1978).

Dietz (1987), doğrusal ve karesel diskriminant analizi ile lojistik regresyon analizini karşılaştırmış, diskriminant analizi varsayımları sağladığında doğrusal diskriminant ve lojistik regresyon analizlerinin karesel diskriminantdan daha az hata içerdiğini göstermiştir. Normallik varsayımı bozulumunda lojistik regresyon analizinin tercih edildiğini vurgularken, modele etkileşim terimleri eklendiğinde karesel diskriminant fonksiyonunun, genişletilmiş (augmented) diskriminant fonksiyonu ve lojistik model ile karşılaştırılması gerektiği üzerinde durmuştur.

İzleyen alt bölümlerde lojistik regresyon ve diskriminant analizinin çeşitli varsayımların sağlandığı durumlara göre karşılaştırması yapılmaktadır.

2.8.1.a. diskriminant analizi varsayımlarının sağlandığı durum

Diskriminant analizinin varsayımları sağladığında lojistik regresyon analizi ile karşılaştırması için, iki grup durumunda,

$$D^2=(\overline{X}^{(1)}-\overline{X}^{(2)})'\,S^{-1}(\overline{X}^{(1)}-\overline{X}^{(2)}) \qquad (2.82)$$

şeklinde tanımlanan Mahallanobis uzaklığının kullanılması önerilmektedir. Burada S^{-1},(2.76) ile tanımlanan varyans-kovaryans matrisinin tersi olup ortalama vektörleri 2.8 bölümünde tanımlandığı gibidir. Böylece,

i) $D^2<2$ ise lojistik modelin diskriminant fonksiyonuna göre etkinliğinin %90 olduğu,

ii) $D^2>3$ ise, $D^2=3$ için bu etkinliğin %65; $D^2=3.5$ için %48 ve $D^2=4$ için %34'e düştüğü vurgulanmaktadır.

Bayne ve arkadaşlarına göre yanlış sınıflandırma olasılığı küçük (<0.15) olduğunda ve küçük örneklemlerde (n<25) lojistik modelin performansı düşüktür (Lesaffre, 1986).

2.8.1.b. lojistik varsayımın sağlandığı durum

Dağılım ailesinin geniş bir kesimi lojistik varsayımı sağlamaktadır. Bu tür dağılımlar için lojistik modelden elde edilen EÇO kestirimleri tutarlı olmaktadır. Doğrusal diskriminant analizinin ayrımsama performansı ise normallik varsayımından belirli sapmalara bağlıdır. Bazen yapılacak bir dönüşüm ile bu performans artırılabilmektedir (Lesaffre, 1986).

2.8.1.c. açıklayıcı değişkenlerin yalnız ikili ya da ikili ve sürekli oldukları durum

Açıklayıcı değişkenlerin ikili ya da ikili ve sürekli değişkenlerin karışımı olduğu durumda diskriminant analizi varsayım bozulumu nedeni ile yanlı kestirimler verirken lojistik modelin ayrımsama performansı da düşebilmektedir. Bu durumda kesikli ve sürekli değişkenler arasında var olabilecek ilişkiler nedeni ile etkileşimlerin de katıldığı eklemeli modellerin incelenmesi önerilmektedir.

Sonuçta denebilir ki, amaç gözlemleri g gruptan birine atamak olduğunda doğrusal diskriminant analizi,etkileşim terimleri ile genişletildiğinde ve yanlış sınıflandırma oranı kriter alındığında varsayımlar sağlamasa bile iyi performans göstermektedir. Öte yandan, amaç sonsal olasılık değerlerinin doğru olarak kestirimi olduğunda ve lojistik varsayımlarında ciddi bir bozulum olmadığında lojistik regresyon analizi tercih edilmektedir (Lesaffre, 1986).

2.9. Lojistik Regresyon Analizinde Artıklar ve Teşhis Ölçütleri

Doğrusal regresyon modelinde Y-E(Y/X) şeklinde tanımlanan artıklar sıfır ortalama ve σ^2 varyansı ile normal dağılmakta olup, X'e bağlı değillerdir. Lojistik regresyon modelinde ise yanıt değişkeni P=P(Y=1/X) başarı olasılığı ile binom dağılımından artık değerleri Y-P şeklinde olup P aracılığı ile X'e bağlıdırlar. Bu da artık incelemesini oldukça güçleştirmektedir. Artıklar modelin verileri ne kadar iyi temsil ettiğini gösteren değerler olarak düşünülürse, lojistik modelde artık benzeri ölçütler ki-kare ve sapma değerleri olarak alınabilmektedir (Jennings, 1986b).

Yanıt değişkeni ikili olduğunda ve verilerde tekrar olmadığında ki-kare ölçütü (2.31)'e benzer olarak,

$$X^2 = \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} (y_i - \hat{P}_i)^2 / \hat{P}_i (1-\hat{P}_i) \qquad (2.83)$$

şeklindedir. Sapma ölçütü ise (2.33) eşitliğinde tanımlandığı gibidir. Bu eşitliklerde her bir gözlem için tanımlanan X_i^2 ve d_i^2 değerlerinden ilki gözlenen ve beklenen değerler farkının ölçeklenmiş ölçüsü, diğeri ise bu farka göre sapma ölçütüne gelen katkı miktarı olup ölçütler lojistik regresyon analizinde teşhis ölçütlerinin hesaplanmasında ve grafiksel yöntemlerde kullanılmaktadırlar (Lesaffre, 1986).

Doğrusal regresyon modellerinde aykırı değer, veri kümesindeki diğer gözlemlere uzak olan büyük artıklı değer; uç değer (extreme observation), açıklayıcı değişkenin extrem olan ve dağılımın iki ucunda yer alan değeri; etkin gözlem (influential observation) ise model kestiriminde büyük etki yaratan gözlem olarak tanımlanmaktadır (Cook and Weisberg, 1980).

Lojistik modellerde ise etkin gözlem ve uç değer tanımları aynı iken aykırı değer, olasılık değeri gerçekte büyük (küçük) iken çok küçük (büyük) olarak kestirilen değer şeklinde tanımlanmaktadır. Diğer bir deyişle gruplara büyük olasılık değerleri ile yanlış olarak atanan gözlem değerleridir. Böylece lojistik model kestirimlerinin normal artık değerlerinden çok aykırı değerlerce belirlendiği söylenebilmektedir.

Çizim ya da ölçütler yardımı ile bulunan aykırı değerler veri kümesinden çıkartıldığında, doğrusal regresyon modellerinin aksine, dağılımın yalnızca bir kuyruğundan kesme olacağı için modelin sistematik kısmı değişmektedir. Bu durumda bulunacak kestirimler yanlı olduğu için aykırı değerlerin modelde kalması gerektiği söylenmektedir (Jennings, 1986b).

2.9.1. İki grup lojistik modellerde teşhis ölçütleri

Yanıt değişkeninin binom dağıldığı lojistik modellerin elemanı olduğu üstel ailenin olabilirlik fonksiyonu,

$$f(Y;\Theta)=\exp(Y\Theta-a(\Theta)+b(Y)) \qquad (2.84)$$

şeklinde verilmektedir. Lojistik model için $Ln(P/(1-P))=\Theta$, $a(\Theta)=nLn(1+\exp(\Theta))$ ve $b(Y)=Ln\binom{n}{Y}$ şeklinde tanımlanmaktadır.

Bu olabilirlik fonksiyonunun logaritmasının Θ'ya göre türevlerini kullanarak artık benzeri ölçütler elde edilebilmektedir (Pregibon, 1981).

Bunlardan ki-kare ölçütü (2.83)'de, D ölçütü ise (2.33)'de tanımlanmakta olup büyük değerleri aykırı gözlemlere karşılık gelmektedir.

Bu ölçütlerin yanı sıra gözlemlerin veri kümesinin merkezine uzaklığının bir ölçüsü olan h ölçütü (leverage) ve buradan m=1-h olarak elde edilen m ölçütü (projection) aykırı değer ve etkin gözlemleri incelemede kullanılmaktadır. İki grup durumunda $P_i=P(y_i=1/X_i)$ olmak üzere, nxn boyutlu V matrisi, $V=kşg(v_1,...,v_n)$ şeklinde tanımlanmaktadır. Matrisin elemanları $v_i = \hat{P}_i(1-\hat{P}_i)$ şeklindedir ($i=1,..,n$).

Böylece nxn boyutlu H matrisi,

$$H=V^{1/2}X(X'VX)^{-1}X'V^{1/2} \tag{2.85}$$

şeklinde elde edilmektedir. M=I-H ise projeksiyon matrisi adını alan simetrik ve idempotent matristir. H(M) matrisinin büyük (küçük) olan köşegen öğeleri $h_{ii}(m_{ii})$, uç değere karşılık gelmektedir. Lojistik model uç değerlere karşı korumalıdır; çünkü,katsayılar kestirilirken varyans değerleri ile ağırlıklandırma otomatik olarak yapılmaktadır (Pregibon, 1981).

Aykırı gözlem ve uç değerleri belirlemek için bu ölçütlerin yanı sıra önerilen çizim yöntemleri arasında Pregibon (1981), i gözlem indisi olmak üzere χ_i-i; d_i-i ve m_{ii}-i çizimlerini, Albert ve Lesaffre (1986) ise $\chi_i/\sqrt{m_{ii}}$-i ve $d_i/\sqrt{m_{ii}}$-i çizimlerini önermişlerdir. Bunların yanı sıra model uygunluğunu sınamada kullanılan deneysel olasılık (emprical probability) ve kısmi artık (partial residual) çizimleri de aykırı değerlerin incelenebileceği çizim yöntemleri arasındadır (Landwehr, et all., 1984; Fienberg and Gong, 1984; Jennings, 1986b).

Öte yandan Pregibon (1981) gözlemlerin D ve χ^2 ölçütleri ile $\hat{\beta}$ kestirimleri üzerindeki etkisini incelemiştir. ℓ'nci gözlem veri kümesinden çıkartıldığında bu ölçütlerdeki değişim sıra ile $\Delta_\ell D$, $\Delta_\ell \chi^2$ ve $\Delta_\ell \hat{\beta}$ olarak tanımlamış, bu ölçütlerin gözlem indislerine karşı çiziminden etkin gözlemlerin belirlenebileceğini vurgulamıştır. Albert ve Lesaffre (1986), $\Delta_\ell \hat{\beta}$/standart hata $(\hat{\beta})$-i çiziminin katsayı kestiriminde etkili olan gözlemi gösterdiğini belirtmişlerdir.

Johnson, yukarıda değinilen ölçütlere alternatif olarak önerdiği ölçütlerde olasılıkların belirlenmesi ve gözlemlerin sınıflandırılmasına göre etkinlikleri incelemiştir. Lojistik regresyonda gözlemler elde edilen olasılık değerlerinin büyüklüğüne göre sınıflandırıldığından, olasılık kestirimine göre etkili olan gözlemlerin sınıflandırmada da etkili olacağını belirtmiştir. Ayrıca gözlemlerin olasılık değerlerini üzerindeki etkisini $\hat{\beta}$'nın belirlediğini, amaç olasılık kestirimi ise $\hat{\beta}$'dan çok olasılıkların incelenmesi gerektiğini ve bunların 0-1 arasında olmasından dolayı daha kolay irdelenebileceğini vurgulamıştır (Johnson, 1985).

2.9.2. Tekrarlı veriler için teşhis ölçütleri

Verilerde J<n grup olduğunda ve her grupta n_j(j=1,..,J) tekrar olduğunda artık benzeri ölçütlerden ki-kare (2.31) de verildiği gibi olup, burada sınırlar n yerine J'ye kadar gitmektedir. Sapma ölçütü ise (2.34)'de verildiği gibidir. Bunların yanı sıra Al-Sarraf ve Young (1986) Pearson artık değerini tanımlamışlardır. Elemanları uç değerleri belirlemede kullanılan H matrisi, (2.85) eşitliğinde tanımlandığı gibidir. Ancak burada tekrarlı veriler söz konusu olduğu için JXJ boyutlu V matrisi, $V=kşg(v_1,...,v_J)$ olarak tanımlanmaktadır. Matrisin elemanları $v_j=n_j\hat{P}_j(1-\hat{P}_j)$ şeklindedir (j=1,..,J). M matrisi yine H matrisinden elde edilmektedir. Her iki matrisin köşegen elemanlarının yorumu (m_{jj}, h_{jj}) yukarıda 2.9.1. bölümünde verildiği gibidir.

Hosmer ve Lemeshow (1989b), türettikleri verilere ilişkin olasılık kestirimine karşı h_{jj} çizimini incelemişler ve beklenenin aksine extrem olasılıklarda h_{jj} değerini küçük bulmuşlardır.

Bu ölçütlerin dışında etkin gözlemleri belirlemeye yönelik ölçütler ve çizimler Hosmer ve Lemeshow (1989b) tarafından incelenmiştir.

2.9.3. Çoklu grup lojistik modeller için teşhis ölçütleri

Çoklu grup lojistik regresyon analizinde H_g temel sınıfına karşı (g-1) tane grubun ikili karşılaştırmaları söz konusudur. Artıkları incelemek için yukarıda değinilen ölçütler her bir H_s-H_g (s=1,..,g-1) karşılaştırması için ayrı ayrı uygulanabilmektedir. Ancak herhangi iki grup karşılaştırması için aykırı ya da etkin bulunan gözlemler tüm gruplar üzerinden incelendiğinde etkin bulunmayabilinmektedirler (Albert and Lesaffre, 1986).

Tüm olası etkin gözlemleri belirlemek için tüm H_s-H_t (t,s=1..,g; s≠t) karşılaştırmaları yapılmalıdır. Böylece g(g-1)/2 karşılaştırma söz konusudur.

Çoklu grup lojistik modellerde artık benzeri ölçütler iki grup ölçütlerinin genellemesidir. Lesaffre ve Albert (1989a), bunlardan ki-kare ve sapma ölçütlerini (2.72) ve (2.73) eşitliklerinde verildiği gibi tanımlamışlardır.

H matrisi (2.85) eşitliğinde verildiği gibi olup n(g-1)xn(g-1) boyutludur. Ancak burada n(g-1)×n(g-1) boyutlu V matrisi $V=kşg(V_1,...,V_n)$ olup V_i, i=1,..,n (2.57) eşitliğinde tanımlanan matristir. M=I-H matrisinin köşegen blokları olan (g-1)x(g-1) boyutlu M_{ii} matrisinin determinant değeri etkili uç değerleri belirlemede kullanılmaktadır. Bu ölçüt aynı zamanda gözlemler çalışmadan çıkartıldığında veri kümesinin hacmindeki değişimin bir ölçüsü olarak da yorumlanmaktadır (Lesaffre and Albert, 1989a). 0-1 arasında değerler alabilen ölçütün sıfıra (bire) yakın değeri i'nci gözlemin kestirimler üzerinde büyük (küçük) etkiye sahip olduğunu (etkili uç değer olduğunu) göstermektedir. Ancak bu etkiye karar vermek için bir kritik değer tanımlanmakta ve $|M_{ii}|\leq 1-2(p+1)(g-1)/N$ ise i'nci gözlemin etkili uç değer olduğu söylenebilmektedir.

Lesaffre ve Albert (1989a), Pregibon'ın (1981), katsayı kestirimi için önerdiği bir-adım kestirim yaklaşımını kullanarak gözlemlerin katsayılar ve sapma üzerindeki etkisini iki grup lojistik modellere benzer olarak elde etmişlerdir. Ayrıca, gözlemleri sınıflara atama sınırlarının etkin gözlemlerden nasıl etkilendiğini görmek ve bu gözlemleri çıkartmadan ya da üzerinde dönüşüm yapmadan önce her birinin sınıflara atama olasılıkları üzerindeki etkisini araştırmak için sınıflandırma teşhis ölçütlerini incelemişlerdir. Etkin gözlemin,bütün gözlemlerin olasılıkları üzerindeki etkisini araştıran ölçütlerin yanı sıra alternatif teşhis ölçütleri de tanımlamışlardır. Pregibon'nın (1981) iki grup lojistik modeller için önerdiği çizimleri çoklu gruba genellemişler ve bunlara ek bazı çizim yöntemleri önermişlerdir.

Buraya dek incelenen ölçütler etkin/aykırı gözlemleri belirlemeye yöneliktir. Çalışmanın amacı risk faktörünü belirlemek olduğunda,etkinliğine karar verilen gözlemler önemli bir durumun göstergesi olabilecekleri için veri kümesinede kalıp dikkatlice incelenmelidirler. Özellikle lojistik modeller için etkin gözlemlerin modelde bulunmasında yarar vardır (Hosmer and Lemeshow, 1989b).

3. ÇALIŞMA

Çalışmanın bu bölümünde kardiyolojik verilerle, öğrenci seçme sınavına ilişkin verilere lojistik regresyon analizi uygulanmaktadır. Uygulamalarda Albert (1985, 1986), tarafından yazılan LOGDIS (program for multiple group logistic discrimination) ve SELLOG (main program for multiple group stepwise logistic discrimination) programları kullanılmıştır. Bunlardan ilki lojistik modelin EÇO kestirimlerini Newton Raphson iteratif yöntemi ile elde etmeye yönelik olup,ikincisi modele en iyi değişkenleri adımsal yöntemle seçen programdır.

Programların işletimi DATA GENERAL MV 10000 bilgisayarlarında yapılmıştır. Ek sonuçlar IBM PC XT'de SYSTAT ve BMDP paket programları kullanılarak elde edilmiştir.

İzleyen alt bölümlerde iki ayrı veri kümesine ilişkin uygulamalar adım adım incelenmektedir.

3.1. Kardiyolojik Verilere Lojistik Model Uygulaması

Epidemilojide geniş bir kullanım alanına sahip olan lojistik model son yıllarda kanser araştırmaları ve kardiyolojik uygulamalarda yaygın bir şekilde kullanılmaya başlanmıştır. Bu çalışmalardan biri kalp hastaları üzerine uygulama şeklinde Lesaffre (1986) tarafından gerçekleştirilmiştir. Elektrokardiogram (EKG) ve vektörkardiogram (VKG) ölçümlerine ilişkin açıklayıcı değişkenlerden oluşan iki modelin kurulduğu çalışmada, ayrımsama güçleri karşılaştırılan modeller arasında fark görülememiştir.

Kalp hastalarının teşhisinde kullanılan ve Türkiye'de de yaygın olan EKG'ye ilişkin gerekli bilgiler şöyle özetlenebilmektedir: Oniki derivasyonlu EKG (12-lead ECG), kalbin vücuttaki elektriksel aktivitesinin elektroşeride (kağıda) yazdırılması olarak tanımlanmaktadır. Bölümleri D1, D2, D3, AVR, AVL, AVF, V1, V2, V3, V4, V5, V6 şeklindedir. Bunlardan ilk altı tanesi kalbi yatay, diğerleri dikey gören bölümlerdir. Bu bölümler yardımı ile kalbin bozuk olduğu kısım görülebilmektedir. Bu oniki bölüm birbiri ile ilişkilidir.

Üç derivasyonlu VKG (3-lead VCG) ise, üç derivasyonlu vektörler yardımı ile kalpteki elektriksel değişikliklerin kalbin ana vektörünün elektroşeride yazdırılması olarak tanımlanmaktadır. VKG bulgularının değerlendirilmesinde matematiksel işlemlerin getirdiği zorluk söz konusudur.

Çalışmada, temelde dört grupta toplanabilecek kalp hastalığı türleri için kurulacak lojistik model yardımı ile hastaların hangi tür kalp hastalığına sahip olduklarını tahmin etmek hedeflenmiştir. Bu tahmini yapmadaki amaç kardiyologların hastalara koydukları tanıların doğruluğunu istatistiksel olarak sınamaktır. Çünkü bazı durumlarda kardiyologlar aynı hastaya farklı tanılar koyabilmektedirler.

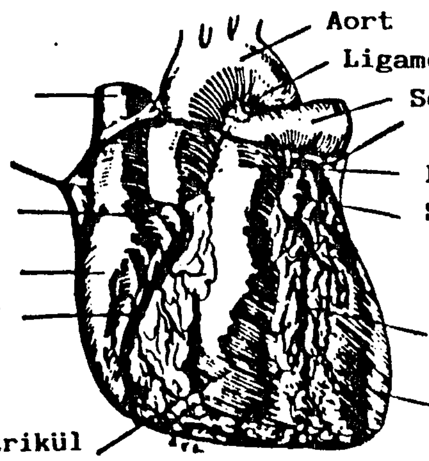
Öte yandan kurulan lojistik modelde değişken seçimine giderek, mümkün olduğunca az test ve/veya tahlil bulgularını kullanarak hastanın hastalık grubunu en doğru şekilde tahmin edecek en iyi modelin bulunması amaçlanmıştır. Böylece hastalardan az sayıda bilgi alınmasının işlem kolaylığı sağlayıp zamandan kazanılacağı düşünülmüştür.

Bu amaçla, düzenli hasta kayıtlarının elde edilebileceği Gülhane Askeri Tıp Akademisi Hastahanesi (GATA) Kardiyoloji Bölümü'nde yatan kalp hastalarından rastgele 250 tanesinin dosyaları incelenerek veriler kaydedilmiştir. Açıklayıcı değişken olarak nelerin alınabileceğine (kalp hastalıklarını etkileyeceği düşünülen faktörler olarak) bir hekim eşliğinde karar verilmiştir. Bu değişkenlerin tanımları ve çalışmaya alınma nedenleri 3.1.2 alt bölümünde verilmektedir. Değişkenlerin tanımlarına geçmeden önce, çalışmanın daha anlaşılır olması için kalbin anatomik yapısına ilişkin bazı temel bilgilerin verilmesinin yarar sağlayacağı düşünülmüştür. İzleyen alt bölümde kalbin kısa bir anatomisi verilmektedir.

3.1.1. Kalbin anatomisi

Kalp, vücudun en hareketli ve hareketi en sürekli olan organı olup 250-300 gr. ağırlığındadır. Göğüste yer alır ve akciğerlerle çevrilidir. Bütün bir yaşam boyu bir emme-basma tulumba olarak çalışan kalbe, ritmik hareketini, myokard kasılma ve gevşemesi vermektedir. Bu hareketin sonucunda meydana gelen atrium ve ventriküllerin, sistol ve diastolü ile atrioventriküler ve semilüner kapakların açılıp kapanması sağlanır. Sürekli hareketler, anatomik yapısının özelliğinden dolayı, kalbin yeri göğüs içerisinde önemli bir değişikliğe uğramadan ve kalp boşluklarının birbirlerine ve komşu bulundukları organlara göre konumu bozulmadan devam eder. Bunu sağlayan iki anatomik yapı vardır. Bunlardan birisi kalbi dıştan çevreleyip tutan ve onu göğüs içerisinde komşu organlara ve göğüs kafesine bağlayan perikard, diğeri de kalbin bir anlamda iç eksenini (bağ dokusundan yapılmış iç iskeleti) oluşturan santral fibröz yapıdır.

Kalbi dört odacığa ayıran, birbirine hemen hemen dik konumda iki düzlem vardır. Bunlardan biri atriumları ventriküllerden ayıran, atrioventriküler kapakların üzerinde bulunduğu yapıdır. Diğeri de sağ atrium ve ventrikülü sol atrium ve ventrikülden ayıran ve atriumlar arasında atrial septum, ventriküller arasında ventriküler septumdan oluşan yapıdır (Sokolow and Mcilroy, 1981; Özden, 1983). Kalbin bu bölümlerini gösterir şema Şekil 3.1'de verilmektedir.



Şekil 3.1. Kalbin Önden Görünüşü

3.1.2. Çalışmada ele alınan değişkenler

Genel olarak dört ana grupta toplanabilen kalp hastalıkları yanıt değişkeninin dört düzeyini oluşturmaktadır. Bu düzeyler, iskemik (Y=1), romatizmal (Y=2), konjenital (Y=3) ve diğer kalp hastalıkları (Y=4) olarak tanımlanmaktadır.

İskemik kalp hastalığı kronik arterdeki (atardamar) darlık ve tıkanıklıklara bağlı olarak ortaya çıkarken romatizmal kalp hastalığı akut (ivegen) eklem romatizmasına bağlı olarak kalp kapaklarının bozulması şeklinde kendini göstermektedir. Konjenital kalp hastalığı,delik kalp vb. gibi doğumsal bir hastalıktır. Diğer kalp hastalıkları, bu üç gruba girmeyen tüm kalp hastalıklarını içermektedir. Bunlar arasında hipertansiyon, kalp yetmezliği, kalpteki ileti ve ritm bozuklukları yer almaktadır. Hastanın hangi tür kalp hastalık grubuna girdiğini anlayabilmek için, bazı bulgularına gereksinim duyulur. Kardiyoloji bölümünde yatan hastalardan alınan bilgiler genel olarak özgeçmiş bilgileri, dinleme, ölçme ve test bulguları olarak gruplandırılabilir. Her biri ayrı önemde olan bu bilgiler içerisinde ölçüm bulguları ayrı bir önem taşımaktadır. EKG ve koroner anjiyo (angio), bu bulguların en önemli olanlarından iki tanesidir.

Koroner anjiyo, bir kateter (kablo) yardımı ile kalp damarlarına boyalı madde vererek bunların görüntülenip filme kaydedilmesidir. Anjiyo ile kalp damarlarındaki tıkanıklık, darlık, doğumsal bozukluklar görüntülenebilmektedir. Operasyon öncesi uygulanması şart olan koroner anjiyo en son ve kesin tanıya götürmektedir.

Tıpta, EKG ile %60, eforlu EKG ile %80 ve anjiyo ile %100 doğru tanı konduğu söylenmektedir. Son çare olması nedeni ile anjiyo bulguları çalışmaya alınmayacaktır. Böylece diğer bulgular yardımı ile en doğru tanıya varma yolları araştırılacaktır.

Bu bilgiler ışığında, çalışmada ele alınan açıklayıcı değişkenler ve tanımları şöyle verilmektedir:

X1: Yaş. Sürekli değişkendir. Kalp hastalığı üzerinde etkili olabilir; çünkü,iskemik hastalıkların ileri yaşlarda, konjenital hastalıkların erken yaşlarda, romatizmal hastalıkların ise 20 yaş civarında sık görüldüğü bilinmektedir.

X2: Cinsiyet. Kesikli olup 1: Erkek, 2: Kadın şeklinde kodlanmıştır. Korunma yetersizliği nedeni ile romatizmal hastalıklar bayanlarda daha sık görülmektektedir. Bayanlardaki kadınlık hormonlarının iskemik hastalığa neden olan değişimlerin oluşumunu yavaşlattığından bu hastalık ise erkeklerde daha sık görülmektedir. Bu nedenle cinsiyet kalp hastalığını etkileyebilecek bir faktör olarak alınmıştır.

X3: Akut romatizma varlığı. Kesikli değişken olup 1: Var, 0: Yok şeklinde kodlanmıştır. Akut romatizma, üst solunum yolu enfeksiyonu geçirildikten sonra vücutta ortaya çıkan bağışıklık (iminolojik) sistemindeki değişikliklerin eklem yerlerini, kalbi ve diğer birkaç dokuyu tutması ile ortaya çıkan belirtiler bütünüdür. Bu faktör kapak hastalıklarında da etkilidir. Halk arasında akut eklem romatizması romatizma ile karıştırıldığından genelde hastalardan bu konuda doğru bilgi alınamamaktadır.

X4: Önceden kalp hastalığı geçirilip geçirilmediği. Kesikli olup 1: Geçirdi, 0: Geçirmedi şeklinde kodlanmıştır. Kalpte daha önce görülen bir bozukluk, o bölgeye belirli bir grup mikrop oturmasına ve kalp iç zarı iltihabı oluşmasına neden olur. Hastada kalp bozulumu yoksa, böyle bir durum söz konusu değildir.

X5: Şeker hastalığının varlığı. Kesikli olup, 1:Var, 0: Yok şeklinde kodlanmıştır. Şeker hastalığı vücuttaki tüm damarlar gibi kalp damarlarını da bozarak iskemik hastalığa neden olur. Bu hastalarda kalp ağrısı olmaksızın enfaktüs görülebilir. Çünkü bunlarda ağrı eşiği-ağrı şiddetinin kritik değeri- yüksektir.

X6: Böbrek hastalığı varlığı. Kesikli olup, 1: Var, 0: Yok şeklinde kodlanmıştır. Böbrek hastalığının kalbi etkilemesi daha çok hipertansiyon yolu iledir. Hipertansiyon kalp için bir engeldir. Kalp bu engele karşı çalışmak zorunda kalır. Belirli bir dönem sonra gücü tükenir ve kalp yetersizliği sorunu ortaya çıkar.

X7: Diğer hastalıkların varlığı. Kesikli olup, 1: Var, 0: Yok şeklinde kodlanmıştır. Bazı hastalıklar kalbin kapak ve öteki bölgelerini etkileyerek kalp hastalığının oluşumuna neden olur. Örneğin frengi, aort kapağı tutarak aort yetmezliğine neden olur.

X8: Anne-babada kalp hastalığı varlığı. Kesikli olup, 1: Var, 0: Yok şeklinde kodlanmıştır. İskemik dışındaki hastalıklar kesinlikle irsi değildir. Bazı uzmanlar kesin olmamakla birlikte iskemiğin irsi olabileceğini ileri sürmektedirler.

X9: Diğer akrabalarda kalp hastalığı varlığı. Kesikli olup, 1: Var, 0: Yok olarak kodlanmıştır. Bu değişkenin irsi bir katkısı olmayıp, daha çok ailenin ekonomik durumu,hastalıklardan korunması gibi yaşam şartları hakkında bilgi verir. Genlerdeki bozukluklar bazı irsi kalp hastalıklarına neden olabilmekte, böylece hastalık ailenin tüm fertlerinde görülebilmektedir.

X10: Sigara içilip içilmediği. Kesikli olup, 1: İçiyor, 0: İçmiyor şeklinde kodlanmıştır. Sigara damar sertliğini hızlandırdığı için iskemik kalp hastalığının ortaya çıkmasına neden olabilir. Hamilelik döneminde sigara içen anne çocuğunun kalp hastası doğmasına neden olabilmektedir.

X11: Alkol alınıp alınmadığı. Kesikli olup, 1: Alıyor, 0: Almıyor olarak kodlanmıştır. Sigara ile aynı etkiye sahiptir.

X12: Çay-kahve içilip içilmediği. Kesikli olup, 1: İçiyor, 0: İçmiyor şeklinde kodlanmıştır. Çay ve kahvenin içindeki kimyasal maddeler kalpte ritm bozukluğuna neden olabilmektedir.

X13: Nabız (vuru). Sürekli olup dakika başına vuru şeklinde ölçülür. Kalp bozukluğunun habercisidir. Uzun süre yüksek ya da düşük nabız kalbin beslenmesini bozduğundan kalp hastalığına neden olabilmektedir.

X14: Sistolik kan basıncı (büyük tansiyon). Sürekli olup mm Hg olarak ölçülür. Kalp kasıldığında damar ağında oluşan basınçtır. Normal değeri 150'dir. Basınç (tansiyon) kalbin önünde bir engeldir, ne kadar yüksek olursa kalp kanı vücuda pompalamak için o kadar zorlanır. Normale gelmediği taktirde belli bir süre kalbin gücü kalmaz ve hastalık belirtisi görülür.

X15: Diastolik kan basıncı (küçük tansiyon). Sürekli olup mm Hg olarak ölçülür. Kalp gevşediğinde damar ağında oluşan basınçtır. Normal değeri 90'dır.

X16: Ödem olup olmadığı. Kesikli olup, 1: Var, 0: Yok şeklinde kodlanmıştır. Ödem, kalbin kanı pompalayamaması sonucu geriye doğru kanın göllenmesi ile ayaklarda, cilt altında sıvının toplanması olarak ortaya çıkan tablodur. Kalp, böbrek ve karaciğer hastalıklarında kendini gösterir ve hepsinde etkisi farklıdır.

X17: İkter (sarılık) olup olmadığı. Kesikli değişken olup, 1: Var, 0: Yok şeklinde kodlanmıştır. Daha çok mekanik kalp kapağı takılan hastalarda, kanın şekilli elemanlarının (alyuvarlar- akyuvarlar) bu kapaktan geçerken çarparak parçalanması sonucu bilirubin denilen madde ortaya çıkar. Bu maddenin ciltte toplanması sarılık hastalığına neden olur. İleri derecede kalp yetmezliklerinde bu tip bir mekanizma ile aynı sorun meydana gelir.

X18: Siyanoz (morluk) varlığı. Kesikli değişken olup, 1: Var, 0: Yok şeklinde kodlanmıştır. Kalbin yeterli derecede kan pompalayamaması sonucu uzak dokularda oksijenlenmenin bozulması ile ortaya çıkan belirtidir.

X19: Çomak parmak varlığı. Kesikli olup, 1: Var, 0: Yok şeklinde kodlanmıştır. Doğumsal olarak siyanozlu kalp hastalarının teşhislerinde önemli katkısı vardır. Parmaklar bu durumda kibrit çöpü gibidir.

X20: Asit varlığı. Kesikli olup, 1: Var, 0: Yok şeklinde kodlanmıştır. Asit, kalbin kanı pompalayamaması sonucu kanın geriye doğru göllenmesi (birikmesi) ile karın boşluğu bölgesinde sıvının toplanması durumunda ortaya çıkan tablodur.

X21: Oskültasyon (üfürüm) varlığı. Kesikli olup, 1: Var, 0: Yok şeklinde kodlanmıştır. Kanın basınç farkı olan iki farklı ortamdan geçerken elde edilen dinleme bulgusudur. Kapak hastalıkları kalp deliklerinin belirticisidir. Çünkü bu durumda üfürüm söz konusudur. Romatizmal ve konjenital kalp hastalıklarının tipik bulgusudur.

X22: Trill varlığı. Kesikli olup, 1: Var, 0: Yok şeklinde kodlanmıştır. İki farklı basınç bölgesi arasından geçen kanın dokunma ile hissedilen belirtisidir. Her durumda görülmez, basınç farkı çok fazla ise duyulabilir. Mırlayan kediyi severken duyulan hissi verir.

X23: Kalp sesinin yeri. Kesikli olup, sesin yeri 5 bölgede özetlenebildiği için 5 düzeyli olarak kodlanmıştır; 1: Aort odağı, 2: Pulmoner odak, 3: Mezokardiak odak, 4: Trikuspit odak, 5: Mitral odak. Kalp boşluklarına bağlı bozukluklar, kapakları belirli bir yerde bulunduğundan kalbin belirli bölgelerinde duyulur. Sesin duyulduğu bölge o bölgeyi ilgilendiren kalp boşluğu ya da kapaktaki bozuklukları düşündürür.

X24: Kalp sesinin (üfürümün) şiddeti. Sıralı olup, üfürüm şiddeti en hafiften şiddetliye doğru 1'den 6'ya kadar numaralanmış gruplarda toplanmıştır. Bu değişken,dinleme sırasında duyulan üfürümün şiddet değerinin sayısal ifadesidir. Şiddet ne kadar fazla ise bozukluk (kapak, kalp deliği bozukluğu vb.)o derece ileridir. 3 şiddetine dek hekimlerin değer yargıları değişirken 4 şiddetinden itibaren üfürüm trill ile birlikte olur ve şiddet hakkında yanılma payı daha azalır.

X25: Kalp sesinin yayılımı. Kesikli olup, 1: Var, 0: Yok şeklinde kodlanmıştır. Üfürüm, kalbin basınç farkı olan bölgeleri arasında geçen kanın oluşturduğu bir bulgu olduğu için bunun dinleme sırasında belirli bir bölgeye yayılımı söz konusudur. Yayılımın yayılış yönü kalbin hangi bölgesinde bozulum olduğunu gösterir.

X26: Kalp sesinin türü. Kesikli olup, 0: Yok, 1: Sistolik, 2: Diastolik ya da sistolik-diastolik karışık şekilde kodlanmıştır. Belirli bir bölgedeki bozuklukların kalp kasılırken veya gevşerken dinleme bulgusu vermesinden dolayı sistolik, diastolik ya da karışık üfürüm ortaya çıkar ve bu da tanı koymada yol gösterici olmaktadır.

X27: Karaciğerin durumu. Kesikli olup, karaciğer normal ise 0, büyük ise 1 şeklinde kodlanmıştır. Karaciğerin büyük olması, kalbin kanı pompalama yetersizliğinden (kalp yetmezliği) dolayı kanı geriye doğru bırakması ile karaciğeri büyültmesi sonucu ortaya çıkan belirtidir ve bozulumun habercisidir.

X28: Kalpte ritm problemi varlığı. Kesikli olup, 1: Var, 0: Yok şeklinde kodlanmıştır. Kalp ritminin bozulması belirli kalp hastalıklarının habercisidir. Belirli kalp hastalıklarında, belirli ritm problemleri vardır. Bu da tanıda yardımcı olmaktadır.

X29: Hipertrofi varlığı ve türü: Kesikli olup, hipertrofi varlığına göre 2 gruba ayrılmaktadır. 0: Hipertrofi yok, 1: sol, sağ ya da biventriküler hipertrofi var şeklinde kodlanmıştır. Kalbin belirli bir engele karşı çalışması sonucu ortaya çıkan kalp adalelerindeki büyümedir. Engel sağ kalp boşluğu önünde ise sağ ventriküler (kalbin büyük odası), sol kalp boşluğunda ise sol ventriküler, her ikisinde de varsa biventriküler hipertrofi vardır. Bunlar o bölgedeki bozulumu gösterdiği için tanıda yararlıdır.

X30: Enfaktüs türü. Kesikli olup, enfaktüs 6 grupta toplandığı için 6 düzeyli olarak kodlanmıştır. Kalp damarının tıkanması sonucu o damarın beslediği kalp kasının ölmesi ile ortaya çıkan bozukluktur. Kalbin belirli bölgelerini belirli damarlar beslediği için tıkanıklık hangi damarda ise o bölge bozulmaktadır. Buna göre bu bölgelere ilişkin kodlama, 0: Enfaktüs yok, 1: Anterior, 2: İnferior, 3: Lateral, 4: Posterior, 5: Anterior-Inferior şeklindedir.

X31: Tele bulguları. Kesikli olup, 0: Normal, 1: Anormal (Patolojik) şeklinde kodlanmıştır. Tele çekimi, kalbin X ışınlı röntgen filmi ile görüntülenmesidir. Gözle izlenebilen patolojik durumlar tanıda yararlı olmaktadır. Farklı hekimlerin bu bulgular yardımı ile koyduğu tanılar birbirleri ile tutarlılık göstermektedir.

X32: Laboratuar bulguları: Kesikli olup, 0: normal, 1: Patolojik şeklinde kodlanmıştır. Patolojik bulgular kalbin kanı pompalayamaması sonucu vücudun diğer dokularında ortaya çıkan bozukluklar nedeni ile elde edilen normal sınırlar dışındaki kimyasal değerlerdir ve tanıda yol gösterici olmaktadır.

3.1.3. Uygulama sonuçları

Çalışmada ele alınan açıklayıcı değişkenlerin büyük bir kısmı kategoriktir. Bunlar arasında ikiden çok düzey içeren 23, 26 ve 30'ncu değişkenler için düzey sayısının bir eksiği kadar göstermelik değişken eklenmesi gerekmektedir. Ancak veriler incelendiğinde 23'ncü değişken için dördüncü kategorinin (trikuspit odak) hiç gözlenmediği görülmüştür. Bu nedenle düzey sayısı beşe düşen değişken dört tane göstermelik değişkenle tanımlanmıştır. Benzer olarak 26'ncı değişkenin iki ve üçüncü kategorileri birleştirilmiş, değişken, indirgenen üç düzeyi için iki göstermelik değişkenle ifade edilmiştir. 30'ncu değişken de dört ve beşinci kategorileri birleştirilerek üç göstermelik değişken ile tanımlanmıştır. Öte yandan 19'ncu değişkenin tüm değerlerinin sıfır olduğu gözlenmiş ve çalışmadan çıkartılmıştır. Böylece çalışmada incelenecek toplam değişken sayısı 37 olarak elde edilmiştir. Yeni veri kümesinde 22, 23, 24 ve 25'nci değişkenler 23 nolu kalp sesinin yerini gösteren değişkene; 28 ve 29'ncu değişkenler 26 nolu kalp sesinin türünü gösteren değişkene; 33, 34 ve

35'nci değişkenler ise 30 nolu enfaktüs türünü gösteren değişkene karşılık gelmektedir.

Elde edilen yeni veri kümesine çoklu grup lojistik model uygulaması LOGDIS programı yardımı ile yapılmış, ancak ikinci derece türevler matrisi singüler çıktığı için sonuç elde edilememiştir. Yapılan araştırma sonunda 20, 28 ve 29'ncu değişkenler arasında medikal olarak da açıklanabilen bir ilişkinin (doğrusal bağımlılık) singülerliğe neden olduğu görülmüş ve bu değişkenler de çalışmadan çıkartılmıştır. Elde edilen 34 değişken için tüm tanı grupları üzerinden tümel ortalama ve standart sapma değerleri Çizelge 3.1'de verilmektedir. Her bir tanı grubu için

Çizelge 3.1. Tüm tanı grupları üzerinden ortalama ve standart sapma değerleri

Değişken No	Tümel Ortalama	St.Sapma
1	37.39	19.259
2	1.18	.382
3	.36	.490
4	.43	.496
5	.09	.290
6	.07	.252
7	.19	.395
8	.17	.378
9	.03	.165
10	.39	.488
11	.15	.360
12	.99	.109
13	82.10	11.227
14	129.48	24.767
15	75.90	20.267
16	.08	.272
17	.02	.126
18	.05	.222
19	.03	.165
20	.06	.245
21	.06	.230
22	.07	.259
23	.23	.423
24	.29	.456
25	.99	.987
26	.64	.482
27	.18	.382
28	.27	.444
29	.40	.492
30	.16	.371
31	.08	.266
32	.04	.196
33	.44	.521
34	.30	.459

gözlem sayısı ve önsel olasılıklar şöyledir: Birinci grupta $n_1=54$, $\tilde{P}_1=0.22$, ikinci grupta $n_2=71$, $\tilde{P}_2=0.28$, üçüncü grupta $n_3=64$, $\tilde{P}_3=0.26$ ve dördüncü grupta $n_4=61$, $\tilde{P}_4=0.24$.

Bu çizelge incelendiğinde hastaların yaş ortalamasının 37 olup, genelde erkek oldukları görülmektedir. Bunun nedeni verilerin elde edildiği GATA Hastanesinde, genelde genç askerlerin kayıtlarının bulunmasıdır.

Yanıt değişkeni farklı kalp hastalıklarını içerdiği için tümel olarak verilen bu ortalama değerler yoruma pek katkı sağlamamaktadır. Bu nedenle her bir tanı grubu için ortalama değerler Çizelge 3.2'de verilmektedir.

Çizelge 3.2. Her bir tanı grubu için değişkenlerin ortalama değerleri (sütunlar: gruplar, satırlar: değişkenler)

0	1	2	3	4
1	56.8889	28.8592	23.8906	44.2131
2	1.2598	1.1972	1.0156	1.2459
3	.0556	.9859	.1406	.1475
4	.6111	.4789	.2656	.3934
5	.2593	.0282	.0000	.1148
6	.1667	.0423	.0156	.0656
7	.2593	.1127	.0781	.3443
8	.2222	.1268	.1094	.2459
9	.0741	.0141	.0000	.0328
10	.4259	.3380	.4062	.3934
11	.2407	.1127	.0937	.1803
12	.9815	.9859	1.0000	.9836
13	81.0185	81.9014	81.1875	84.2459
14	131.0185	125.1408	128.0469	134.6721
15	80.4630	67.2535	77.1875	80.5738
16	.1111	.0563	.0156	.1475
17	.0370	.0141	.0000	.0164
18	.0556	.0141	.0156	.1311
19	.0185	.0141	.0156	.0656
20	.0000	.0704	.1719	.0000
21	.0185	.0423	.1250	.0328
22	.0185	.0000	.2500	.0164
23	.0185	.2958	.4375	.1311
24	.1481	.6620	.1719	.1148
25	.3333	.9970	.9860	.4754
26	.1852	1.0000	.9687	.2623
27	.2487	.1408	.0312	.3115
28	.2407	.2113	.0937	.5410
29	.4259	.4930	.2969	.3934
30	.5741	.0141	.0000	.1475
31	.2778	.0000	.0000	.0656
32	.1111	.0000	.0156	.0492
33	.4444	.5070	.3281	.4754
34	.6852	.1690	.0312	.3934

Gerek her bir değişkenin yanıt değişkeni ile çapraz çizelgeleri, gerekse Çizelge 3.2 incelendiğinde, yaş ortalamasının iskemik kalp hastalarında 57, romatizmal kalp hastalarında 29, konjenital kalp hastalarında 24 ve diğer grup kalp hastalarında ise 44 olduğu görülmektedir.

Dört tür kalp hastalığı da genelde erkeklerde gözlenmektedir.

Akut romatizma çoğunlukla romatizmal hastalarda gözlenmekte olup bu grupta önceden kalp rahatsızlığı geçiren hasta sayısı çoğunluktadır.

Şeker hastalığı en sık iskemik hastalarda gözlenmektedir.

Nabız dört grupta da ortalama 81 olup, normal sınırlar olan 60-100 arasındadır. Böylece nabzın kalp hastalıkları üzerinde önemli bir etki yaratmadığı söylenebilir. Bunun nedeni ise hekimlerin bazen nabzı dikkate almaksızın değerlendirmede bulunmaları şeklinde açıklanabilmektedir.

Büyük tansiyon (sistolik kan basıncı) en yüksek diğer grup kalp hastalarında olup 134.7 mm Hg iken en düşük romatizmal hastalarda 125.1 mm Hg'dir. Bu değerler normal sınır olan 150'nin altında olduğu için makul değerlerdir.

Küçük tansiyon yine en yüksek diğer grup hastalarda olup 80.6 mm Hg iken, en düşük romatizmal hastalarda 67.3 mm Hg'dir. Bu değerler normal sınır olan 90'a yakın sayılabilmektedir.

Kalp sesi, konjenital hastalarda mezokardiak odakta, romatizmal ve diğer grup hastalarda mitral odakta yer almaktadır. İskemik grupta ise genelde sese rastlanmamaktadır.

Kalp sesinin şiddeti yine iskemik hastalarda söz konusu değilken kalan üç grup hastalarda 2'dir.

Ritm problemi en sık romatizmal ve diğer grup hastalarda görülürken, hipertrofiye en çok romatizmal hastalarda rastlanmaktadır.

Enfaktüs, iskemik hastalıklarda en sık anterior diğer grup hastalarda ise inferior ve lateral bölgelerde görülmektedir.

Laboratuar bulguları genelde iskemik ve diğer grup hastalarda patolojik çıkarken, kalan grup hastalarda normal elde edilmektedir.

Çoklu grup lojistik regresyon analizi 34 değişkenli veri kümesine uygulandığında yarı-tam bölünme durumu nedeni ile ikinci derece türevler matrisinin bir öğesi çok büyük bulunmuş ve sonlu kestirimler elde edilememiştir. Albert ve Anderson'ın (1984), 2.6.3 alt bölümünde değinilen yaklaşımı doğrultusunda veriler standartlaştırılarak lojistik regresyon analizi yeniden uygulanmıştır. Analiz sonucunda verilerde yine yarı-tam bölünme durumu gözlenmiştir. Böylece EÇO kestirimlerinin sonlu olmadığı, ancak ayrımsama amacı ile kullanılabileceği söylenebilmektedir. Elde edilen modelin sınıflandırma çizelgesi Çizelge 3.3'de verilmektedir.

Çizelge 3.3. 34 değişkenli standartlaştırılmış verilere uydurulan lojistik modelin sınıflandırma çizelgesi (sütunlar: gerçek gruplar, satırlar: atanan gruplar)

	1	2	3	4
1	52	0	0	1
2	0	69	4	0
3	0	2	60	0
4	2	0	0	60
Toplam	54	71	64	61

Bu çizelgeye göre modelin doğru sınıflandırma oranı (52+69+60+60)/250=0.96 olarak elde edilmektedir. Yanlış sınıflandırmalar özellikle 2-3 (romatizmal-konjenital) grupları ile 1-4 (iskemik-diğer) grupları arasında söz konusudur. 2'nci gruptan 3'ncü gruba yanlış atama oranı 0.03 iken 3'ncü gruptan 2'nci gruba bu oran 0.06'dır. Bu iki tür kalp hastalığına ilişkin yanlış tanı koymanın nedeni hastaların romatizma konusunda bilgilerinin yeterli olmaması ve dolayısı ile hekimlere yanlış ya da eksik bilgi vermelerinden kaynaklanmaktadır. Diğer yandan l'nci gruptan 4'ncü gruba yanlış atama oranı 0.04 iken 4'ncü gruptan l'nci gruba bu oran 0.02'dir. Bu iki hastalık türü için yanlış tanılamanın nedeni ise, hastaların önce iskemik kalp hastalığı varken tedavi görmediğinden dolayı hastalıklarının diğer grup kalp hastalıklarından birine çevirmesi olarak yorumlanabilmektedir.

Böylece verilerde yarı-tam bölünme durumuna rağmen oldukça iyi bir ayrımsama elde edilmiştir. Ayrıca gruplararası yanlış atanmalar da medikal olarak oldukça anlamlıdır.

Bunun yanı sıra modelin etkinlik ki-kare değeri (bu değer p(g-1) serbestlik dereceli ki-kare dağılımlı olan ve kestirilen model ile sabit terimi içeren modelin olabilirliklerinin farkına dayanan sapma ölçütüne karşılık gelmektedir). 617.44 olup 102 serbestlik derecesinde anlamlıdır. Bu da 34 değişken üzerinden elde edilen modelin anlamlılığının bir göstergesidir.

Çalışmanın amaçlarından biri, mümkün olabildiğince az bilgi kullanarak doğru tanı koymak ve en önemli etkenleri seçmektir. Bu noktadan hareketle en iyi (parsimonıous) lojistik modeli bulmak hedeflenmektedir. İlk adım olarak 34 değişken içerisinde en önemli olanları seçmek için adımsal değişken seçimi uygulanmıştır. Seçim sonucunda elde edilen değişkenler şunlardır: Yaş (X1), cinsiyet (X2), akut romatizma (X3), önceden geçirilen kalp hastalığı (X4), şeker hastalığı (X5), nabız (X13), kalp sesinin yeri değişkeninin son iki düzeyi (X23, X24), üfürümün şiddeti (X25), karaciğerin durumu (X27), ritm problemi (X28), enfaktüs türü değişkenin düzeylerinin hepsi (X30, X31, X32) ve laboratuar bulguları (X34).

Bu değişkenler 3.1.2 alt bölümünde verilen tanımları doğrultusunda medikal olarak oldukça anlamlı bulunmuştur. Ancak seçilenler arasında kalp sesinin yerini gösteren değişkenin aslında dört göstermelik değişkenle ifade edilmesine rağmen sadece iki düzeyi önemli bulunmuştur. Bu düzeyler 3 ve 5 kodlu mezokardiak ve mitral odağa karşılık gelmektedir. Diğer düzeyler modele anlamlı katkısı olmadığı düşüncesi ile birleştirilmiş ve değişken iki göstermelik değişken olarak yeniden düzenlenmiştir. Böylece 32 değişkene indirgenen veri kümesine, öncelikle çoklu grup lojistik regresyon analizi uygulanmıştır. Yine yarı-tam bölünme durumunun gözlendiği modelin doğru sınıflandırma oranı 0.95 olup, etkinlik ki-kare değeri 598.04 bulunmuştur. Bu da iyi ayrımsama yapan modelin 96 serbestlik derecesinde anlamlı olduğunun bir göstergesidir. Daha sonra bu modele değişken seçimi uygulanmış ve aşağıda tanımlanan değişkenler önemli bulunmuştur:

Değişken No	Adı
1 (X1)	Yaş
2 (X2)	Cinsiyet
3 (X3)	Akut romatizma
4 (X4)	Önceden kalp hastalığı varlığı
5 (X5)	Şeker
6 (X11)	Alkol
7 (X21)	Kalp sesinin yeri
8 (X22)	
9 (X23)	Kalp sesinin şiddeti
10 (X25)	Karaciğerin durumu
11 (X26)	Ritm problemi
12 (X27)	Hipertrofi
13 (X28)	
14 (X29)	Enfaktüs türü
15 (X30)	
16 (X32)	Laboratuar bulguları

Seçilen her bir değişken için modelin log-olabilirlik ve ki-kare değerleri ile birikimli doğru sınıflandırma yüzdeleri Çizelge 3.4'de özetlenmektedir.

Buradan da görüldüğü gibi modelin doğru sınıflandırma oranında artış yaratmayan değişkenler de anlamlı değişken

Çizelge 3.4. 32 değişken üzerinden adımsal değişken seçim sonuçları

Log-olabilirlik	Değişken No	Ki-kare	P	Birikimli doğru sınıflandırma yüzdesi %
-248.684215	23	193.39	0.0000	.44
-179.758263	3	137.85	0.0000	.70
-156.307419	1	46.90	0.0000	.72
-147.699232	26	17.22	0.0006	.76
-138.799791	28	17.80	0.0005	.79
-128.029356	29	21.54	0.0001	.83
-115.864784	30	24.33	0.0000	.84
-107.806962	27	16.12	0.0011	.87
-100.602546	22	14.41	0.0024	.86
-94.521988	4	12.16	0.0069	.87
-88.618631	2	11.81	0.0081	.88
-82.027304	5	13.18	0.0043	.91
-74.112974	21	15.83	0.0012	.91
-69.033718	32	10.16	0.0173	.91
-64.948816	11	8.17	0.0426	.91
-61.842737	25	6.21	0.1017	.93

olarak seçilebilmektedir. Bu da 2.4 alt bölümünde değinildiği gibi doğru sınıflandırma oranının model yeterliliği için tek başına bir kriter olamayacağının göstergesidir.

Lojistik modelin performansını seçilen 16 değişken üzerinde sınamak amacı ile lojistik regresyon analizi uygulanmıştır. Uygulamada elde edilen katsayı kestirimleri ve bunların standart hataları ile modelin sınıflandırma çizelgesi Çizelge 3.5 a ve b'de verilmektedir.

Standartlaştırılmış veriler üzerinden elde edilen lojistik modelde yarı-tam bölünmeden dolayı kestirimler sonlu değildir. Ancak Çizelge 3.5.b'den de görülebileceği gibi doğru sınıflandırma oranı 0.93'dür. Değişken sayısı 32'den 16'ya azalırken bu oranın 0.95'den 0.93'e inmesi,

Çizelge 3.5. 16 değişkenli lojistik modelin

a) Katsayı kestirimleri ve standart hataları (sütunlar: gruplar; satırlar: değişkenler)

	Katsayı kestirimleri		
	1	2	3
0	-2.2458	-6.4864	-4.5242
1	-.6808	-2.0433	-1.0092
2	-.9948	4.7795	-1.0843
3	1.5323	4.7856	1.3876
4	.7899	6.4943	5.2841
5	1.1042	-1.7821	-4.9912
6	-.4243	2.4634	1.2881
7	-3.7634	1.0976	-.4431
8	-2.0994	4.8667	1.9209
9	2.8301	6.0307	7.3095
10	-.2092	-1.7660	-2.0883
11	-.7835	-1.6340	-.9763
12	-1.2854	.8265	-.3807
13	3.8400	-6.3235	-5.6928
14	2.5752	-7.2241	-4.1355
15	1.9628	-5.3752	-.9104
16	-.2308	-4.1230	-2.6536

	Katsayı kestirimlerinin standart hataları		
	1	2	3
0	.8694	15.1643	27.0385
1	.8351	2.0490	1.2745
2	.4727	1.3548	3.1967
3	.8822	1.1220	.9448
4	.5444	1.8893	1.7587
5	.4885	1.2881	50.9127
6	.4758	2.0405	1.9837
7	1.3348	.9127	.5871
8	1.0288	1.5853	1.2296
9	1.1172	1.8165	1.9065
10	.3385	1.0687	1.1276
11	.4617	.9128	.7846
12	.4965	.7701	.7222
13	.9891	2.2905	43.0450
14	.7705	42.0740	36.0560
15	.5395	44.5701	1.7975
16	.5108	1.7876	1.2081

b) Sınıflandırma çizelgesi: sütunlar (gerçek gruplar; satırlar: atanan gruplar)

	1	2	3	4
1	51	0	0	4
2	0	69	7	0
3	1	2	57	2
4	2	0	0	55
Toplam	54	71	64	61

çok daha az sayıda değişkenle modelin ayrımsama performansının iyi olduğunun bir göstergesidir. Ayrıca yanlış sınıflandırmalar 1-4 ve 2-3 grupları arasında söz konusu olup daha önce de değinildiği gibi bu durum medikal olarak açıklanabilmektedir.

Modelin etkinlik ki-kare değeri 567.07 olup, 48 serbestlik derecesinde anlamlıdır. Bu da 16 değişkenin verileri açıklamada yeterli olduğunun bir göstergesidir. Benzer olarak 32 değişkenli ve seçilen 16 değişkenli modellerin log-olabilirlik değerlerinin farkının -2 katı şeklinde elde edilen sapma ölçütü değeri 43.22 olarak bulunmuştur. Bu değer 48 serbestlik dereceli ki-kare tablo değeri ile karşılaştırıldığında modele seçilen değişkenlerin anlamlı olduğu, seçilmeyen diğer 16 tanesinin etkisinin önemsiz olduğu görülmektedir. Böylece daha az parametre ile (parsimonous model) iyi bir ayrımsama modeli elde edilebilmektedir.

Elde edilen modelin geçerliliğini sınamanın diğer yolu da artıkların incelenmesidir. Artık değerleri gözlemlerin her bir gruba atanma olasılıkları yardımı ile elde edilmektedir. Çizelge 3.5.a'da verilen katsayı kestirimleri, elde edilen modelde yerine konarak skor değerleri, bunlar yardımı ile de (2.42) eşitliğinden gözlemlerin gruplara atanma olasılıkları hesaplanmıştır. 250 gözleme ilişkin bu skor değerleri, gerçek ve atandıkları gruplar ile atanma olasılıkları EK-I'de verilmektedir. 2.7.1 alt bölümünde tanımlanan artık benzeri χ_i^2 ve d_i^2 $(i=1,..,n)$ ölçütleri bu olasılık değerlerini kullanarak elde edilmiştir. Ölçüt değerlerinin gözlem numarasına karşı çizimleri yapılmış ve gözlemlerin d_i^2 değerleri birbirinden çok farklı bulunmazken, 5 tane gözlemin diğerlerinden daha büyük χ_i^2 değerine sahip aykırı gözlem olduğu görülmüştür. Bu gözlemler 23, 35, 93, 99 ve 166'ncı gözlemler olup gruplara büyük olasılık değerleri ile yanlış atanmışlardır. Ayrıca bu gözlemlerin açıklayıcı değişken değerleri incelendiğinde, diğer gözlemlere ilişkin değerlerden farklı görülmediği için aykırılıklarının, verilerdeki yarı-tam bölünmeden kaynaklandığı ve bu nedenle de yanlış gruplara atandıkları söylenebilir.

Lojistik modelde aykırı gözlemlerin veri kümesinde kalması gerektiği için bu gözlemlerin bulunduğu model ile çalışmaya devam edilmiştir.

Çalışmada verileri en iyi şekilde açıklayacak lojistik modele ulaşmak amaçlanmıştır. Bunun için elde edilen 16 değişkenli lojistik modeldeki kesikli değişkenler arasında etkileşim olup olmadığı, sürekli değişkenlerin ise doğrusal etkileri dışında karesel ya da daha yüksek dereceden etkilerinin bulunup bulunmadığı araştırılmıştır. Modeldeki iki sürekli değişken, Hosmer ve Lemeshow'un (1989b) önerdiği gibi gruplara bölünerek kategorik hale getirilmiştir*. Bunların her bir düzeyi için diğer değişkenlerin ortalama değerleri her üç lojistik modelde de yerine konarak lojit değerleri elde edilmiştir. Kesikli hale getirilen değişkenlerin düzeylerine karşı lojit çizimleri yapılmış, kategoriler arasında büyük değişimler olmadığı için doğrusal etkileri olduğuna karar verilmiştir.

Kesikli olan diğer değişkenler arasında etkileşim varlı-

---

* Sürekli değişkenlerden yaş değişkeni 20-30; 31-45; 46-60, 61 ve daha yukarı olmak üzere 4, kalp sesinin şiddeti değişkeni ise 0, 1, 2, 3, 4, 5 olarak 6 gruba bölünmüştür.

ğına karar verebilmek için, bir önbilgi olması amacı ile Spearman ilişki katsayıları incelenmiştir. Gerek büyük örneklemler için (n=250) Spearman ilişki katsayısının kritik değerinden gerekse medikal anlamlılıklarından yararlanarak önemli sayılabilecek ilişki miktarları ve bu ilişkiye sahip değişkenler şöyle özetlenmiştir: $r_{12}=0.45$, $r_{14}=0.43$, $r_{15}=0.38$, $r_{19}=0.40$, $r_{1,10}=0.46$, $r_{1,13}=0.49$, $r_{1,16}=0.66$, $r_{2,16}=0.41$, $r_{3,8}=0.42$, $r_{3,9}=0.41$, $r_{4,16}=0.37$, $r_{5,16}=0.45$, $r_{7,8}=0.36$, $r_{7,9}=0.40$, $r_{8,9}=0.36$, $r_{9,13}=0.39$, $r_{9,16}=0.36$, $r_{10,12}=0.31$, $r_{10,16}=0.37$, $r_{12,16}=0.30$, $r_{13,16}=0.41$.

Bu ilişkilere dayanarak, X1xX2, X1xX4, X1xX5, X1xX9, X1xX10, X1xX13, X1,xX16,X2xX16, X3xX8, X3xX9, X4xX16, X5xX16, X7xX8, X7xX9, X8xX9, X9xX13, X9xX16, X10xX12, X10xX16, X12xX16, X13xX16 etkileşim terimleri de modele eklenerek 37 değişken (16+21) üzerinden lojistik model araştırılmıştır. Ancak 29'ncu değişken olan X7xX8 etkileşim terimi tüm gözlemler için sıfır olduğundan çalışmadan çıkartılmıştır. Böylece 36 değişkene indirgenen yeni veri kümesi için lojistik regresyon analizi yeniden uygulanmıştır. Elde edilen modelin katsayı kestirimleri ve bunların z değerleri ile sınıflandırma çizelgesi sıra ile Çizelge 3.6.a ve b'de verilmektedir.

Çizelgenin b kısmı incelendiğinde genişletilmiş modelin doğru sınıflandırma oranının 0.94 olduğu görülmektedir. 16 değişkenli modelde bu oran 0.93 olduğundan etkileşim terimleri modelin ayrımsama gücünde önemli bir katkı yaratmamıştır. Ancak modelin etkinlik ki-kare değeri 605.42 olup, 108 serbestlik derecesinde oldukça anlamlıdır. Öte yandan eklenen terimlerin anlamlılığını test etmek için yeni model ile 16 değişkenli modelin log-olabilirlik değerleri farkına dayalı sapma ölçütü 47.17 olarak elde edilmiştir. 60 serbestlik derecesinde (108-48) ki-kare tablo değeri ile karşılaştırıldığında etkileşim terimlerinin anlamlı olmadığı görülmüştür. Böylece etkileşim terimleri ile genişletilmiş karesel lojistik model

Çizelge 3.6. Etkileşim terimleri ile genişletilen 36 değişkenli lojistik modelin

a) Katsayı kestirimleri ve bunların Z değerleri (sütunlar: gruplar, satırlar: değişkenler)

	Katsayı kestirimleri		
	1	2	3
0	-4.7390	-.1231	1.0498
1	3.9821	5.0026	2.5249
2	9.9718	5.8735	5.4357
3	1.5198	1.4617	-.9353
4	3.0575	-4.6565	12.8842
5	-9.3216	-14.3671	-7.7658
6	.1309	2.7524	1.5424
7	-7.9383	4.7756	5.2975
8	-2.0355	4.8944	5.7417
9	-3.6096	6.7095	10.8968
10	-2.5355	-5.7117	-2.4259
11	-1.4365	-3.0275	-2.0569
12	-.9919	-.4914	-1.1183
13	2.6783	-.8978	-2.8964
14	4.2624	-1.0167	.2230
15	3.4697	-1.9200	-1.2906
16	11.9764	5.0427	-.1404
17	-16.7456	-6.1014	-5.9695
18	-1.7540	-13.8413	-10.5142
19	11.5904	12.5290	8.7857
20	11.2578	4.1178	2.7406
21	3.5788	2.1417	-.2200
22	3.8575	-.6551	4.9305
23	-5.0169	-4.8323	3.8020
24	-4.1034	-.6146	-6.9143
25	-2.3823	.4524	.2315
26	6.6928	5.0043	4.6004
27	-1.1019	1.5389	-3.5176
28	.5151	3.7095	1.3373
29	-2.2166	-6.3063	-8.4699
30	-3.6096	-3.6657	-7.0742
31	2.9174	2.5311	-1.2208
32	-9.2048	-6.2236	-3.4569
33	-2.4529	1.2624	-1.3425
34	-.2279	2.5251	1.5799
35	.7160	.2268	5.3596
36	-.4493	-4.4036	-.5249

	Z değerleri		
	1	2	3
0	-.5274	-.0208	.1650
1	.6059	.2205	.0688
2	1.0827	.3000	.1800
3	.3316	.3582	-.2537
4	.3955	1.4160	1.2688
5	-.3215	-.2804	-.1470
6	.1401	.7312	.4133
7	-.1499	1.0217	1.2640
8	-.4387	.8767	1.1762
9	-.3067	.4509	.7383
10	-.3610	-.3115	-.1134
11	-.9959	-2.1762	-1.6161
12	-.8715	-.5570	-1.4572
13	.4250	-.0334	-.1138
14	1.4969	-.1127	.0262
15	1.5820	-.4359	-.2926
16	.6870	.2767	-.0049
17	-1.0457	-.1671	-.0932
18	-.2094	-.8750	-.6916
19	1.1475	.3635	.2481
20	1.1838	.2757	.1797
21	.5917	.1006	-.0080
22	.4313	-.0184	.1654
23	-.3459	-.1866	.1550
24	-.7881	-.0295	-.2492
25	-.1648	.0605	.0311
26	.8713	.8711	.8465
27	-.2885	.1357	-.3677
28	.0192	.1215	.0430
29	-.0379	-.8386	-1.2230
30	-.4060	-.4284	-.8972
31	1.2084	.3938	-0.1231
32	-1.4717	-.7897	-.3024
33	-.9271	-.1190	-.1059
34	-.0782	.2551	.1095
35	.3078	.0274	.4768
36	-.1540	-.4757	-.0441

b) Sınıflandırma çizelgesi (sütunlar: gerçek gruplar, satırlar: atanan gruplar)

	1	2	3	4
1	53	0	1	2
2	0	69	7	0
3	0	2	56	2
4	1	0	0	57
Toplam	54	71	64	61

her ne kadar anlamlı olsa da ayrımsama gücüne büyük bir katkı getirmemektedir ve doğrusal model ile karşılaştırıldığında tercih edilmemektedir.

Bunun yanı sıra anlamlı bulunan karesel lojistik model için değişken seçimine gidilerek mümkün olduğunca az değişkenli, ayrımsama gücü yüksek model elde etmek hedeflenmiştir. Seçim sonucunda, yaş (X1), akut (X3), önceden geçirilen kalp hastalığı (X4), kalp sesinin şiddeti (X9), karaciğer (X10), ritm problemi (X11), enfaktüs türünün üç düzeyi (X13, X14, X15) ana etkilerinin yanı sıra, cinsiyetxlaboratuar bulgusu (x2xX16), akutxkalp sesinin yeri (X3xX8), akutxkalp sesinin şiddeti (X3xX9), şekerxlaboratuar bulguları (X5xX16), kalp sesinin yerixşiddeti (X7xX9) ve kalp sesinin şiddetixlaboratuar bulguları (X9xX16) etkileşim terimleri anlamlı bulunarak modele alınmıştır. Bu 15 değişkenli veri kümesine uygulanan lojistik regresyon analizi ile elde edilen lojistik modelin doğru sınıflandırma oranı 0.92 olup etkinlik ki-kare değeri 45 serbestlik derecesinde 562.48'dir ve modelin yeterliliğinin bir göstergesidir. 36 değişkenli tam modele göre doğru sınıflandırma oranında 0.02'lik bir azalma vardır. Bu da azalan değişken sayısına rağmen ayrımsamanın iyi olduğunun bir göstergesidir.

Öte yandan modelin, 36 değişkenli modele kıyasla sapma ölçütü hesaplanmış ve 63 serbestlik derecesinde 51.75 olarak bulunmuştur. Bu da seçilmeyen diğer 21 değişkenin modele katkısının çok önemli olmadığının bir göstergesidir. Böylece 15 değişkenli karesel model tam modele oldukça yakın olduğundan tercih edilebilir.

Çoklu grup lojistik model uygulamasında temel olan dört modele ilişkin özet değerler ve bunlar arasında seçimi sağlayacak sapma ölçütleri Çizelge 3.7'de verilmektedir. Buna göre log-olabilirlik, etkinlik ki-kare değerleri ile doğru sınıflandırma oranları kriter alındığında, değişkenlerin seçilen alt kümeleri üzerinden kurulan modeller arasında orjinal 16 değişkenli model en iyisidir. Ayrıca tam modeller arasında da yine orjinal 32 değişkenli model karesel modelden çok farklı olmadığı için tercih edilebilir.

Sonuçta karesel lojistik model yerine doğrusal modelin verileri yeterince temsil edilebildiği söylenebilmektedir. Etkileşim terimlerinin modele getirdiği katkı gözardı edilebilecek kadar azdır.

Çalışmada çoklu grup lojistik regresyon analizi sonuçlarını grupların ikişerli olarak incelendiği lojistik regresyon analizi sonuçları ile karşılaştırmak için 2.6.4. alt bölümünde değinilen Begg ve Gray'in (1984) yaklaşımı ile ikili grup lojistik modeller incelenmiştir. Bunun için, gerek değişkenlerin ortalama değerleri gerekse gözlem sayısı açısından normal değerlere sahip olan yanıt değişkeninin 4'ncü düzeyi (4'ncü grup) temel sınıf olarak alınmış ve 1-4, 2-4, 3-4 karşılaştırmalarına ilişkin lojistik regresyon analizi önce tam model (32 değişken) sonra da seçilen model (16 değişken) için yapılmıştır. Ayrıca bu ikili grup karşılaştırmaları diskriminant analizi ile de yapılarak sonuçlar lojistik model ile kıyaslanmıştır. 32 değişken üzerinden her üç karşılaştırma için yapılan uygulamada elde edilen lojistik modellerin doğru sınıflandırma oranları 1.0 (%100) olarak bulunmuştur. Bu da verilerde tam bölünme durumunun bir göstergesidir ve sonlu katsayı

Çizelge 3.7. Çoklu grup lojistik modele ilişkin özet bulgular

Modeldeki Değişken sayısı (p)	Log-Olabilirlik	Etkinlik ki-kare değeri	Serbestlik Derecesi p(g-1)	Doğru Sınıflandırma oranı %	$D=G^2$(sd)
32	-40.235031	598.04	96	95	
					43.22 (48)
(32 değişken üzerinden seçilen) 16	-61.842877	567.07	48	93	$X^2_{(0.05;48)}=67.5$
					47.17 (60)
36 (16 değişken+21 etkileşim)	-38.257025	605.42	108	94	$X^2_{0.05;60}=79.08$
					51.75 (63)
(36 değişken üzerinden seçilen) 15	-64.13382	562.48	45	92	$X^2_{(0.05;63)}=79.08$

kestirimlerinin bulunamamasına neden olmaktadır. Böylece çoklu grup lojistik modellerde gözlenemeyen tam ayrılmış gruplar ikili karşılaştırmalarla elde edilmiştir. Ancak her biri için tam bölünme durumu nedeni ile kestirimler,çoklu grup modellere nazaran etkinlik açısından daha zayıftır.

Her üç ikili karşılaştırma için 16 değişken üzerinden,lojistik regresyon ve diskriminant analizi ile elde edilen katsayı kestirimleri,bunların t değerleri ve modellerin sınıflandırma çizelgeleri Çizelge 3.8'de sıra ile a, b ve c'de verilmektedir.

Çizelge 3.8. 16 değişken üzerinden ikili lojistik modellerin ve diskriminant fonksiyonlarının katsayı kestirimleri, t değerleri ve sınıflandırma çizelgeleri (sütunlar, gerçek gruplar, satırlar, atanan gruplar)

a) 1-4 karşılaştırması için

	Lojistik modelin			Diskriminant fonksiyonunun	
	Katsayı kestirimleri	t değerleri		Katsayı kestirimleri	t değerleri
0	.1811	.0788	0	-1.7173	
1	-.0473	-1.0177	1	-.0096	-.632
2	-2.9204	-2.2218	2	-2.1525	-4.218
3	1.3375	.7899	3	1.4807	2.206
4	1.6450	1.4879	4	1.0048	2.138
5	3.9864	2.2314	5	2.6841	4.571
6	-1.4206	-1.0426	6	-.0589	-.114
7	-9.0059	-2.7661	7	-6.2542	-5.381
8	-4.6281	-1.9652	8	-2.9125	-3.401
9	2.7446	2.4281	9	1.7194	3.707
10	-.5699	-.6354	10	-1.8401	-3.754
11	-1.4684	-1.4262	11	-1.7330	-4.100
12	-1.6820	-2.7884	12	-1.1054	-2.373
13	10.2412	3.8445	13	9.6300	12.596
14	9.4540	3.3116	14	8.4441	10.498
15	9.5587	3.6162	15	9.0199	9.141
16	-.2711	-.2422	16	-.5632	-1.155

Sınıflandırma çizelgesi

Lojistik modelin

	1	4
1	52	4
4	2	57
Toplam	54	61

Diskriminant fonksiyonunun

	1	4
1	51	7
4	3	54
Toplam	54	61

b) 2-4 karşılaştırması için

Lojistik modelin			Diskriminant fonksiyonunun		
	Katsayı kestirimleri	t değerleri		Katsayı kestirimleri	t değerleri
0	-12.4347	-2.9406	0	15.1234	
1	.0436	.4421	1	.0559	3.078
2	3.6870	1.7154	2	1.0777	1.893
3	5.0964	3.2390	3	12.4462	17.850
4	2.5444	1.1853	4	1.6699	3.789
5	-2.1315	-.6915	5	-3.5695	-4.365
6	3.3328	1.6397	6	.2656	.528
7	1.8460	.7147	7	2.5197	3.329
8	4.6295	1.5649	8	4.8805	7.026
9	1.8181	1.5671	9	2.5545	8.140
10	-2.2418	-1.1428	10	-2.8474	-5.123
11	-1.8708	-1.1982	11	-2.5250	-6.205
12	.7194	.6988	12	2.7534	6.678
13	-6.2398	-1.9907	13	-3.1714	-3.735
14	-.4544	-.0686	14	-1.3473	-1.136
15	-7.5461	-1.0603	15	-6.1774	-4.674
16	-3.5021	-1.7716	16	1.8447	-2.963

Sınıflandırma çizelgesi

Lojistik modelin

	2	4
2	71	0
4	0	61
Toplam	71	61

Diskriminant fonksiyonunun

	2	4
2	70	4
4	1	57
Toplam	71	61

c) 3-4 karşılaştırması için

Lojistik modelin			Diskriminant fonksiyonunun		
	Katsayı kestirimleri	t değerleri		Katsayı . kestirimleri	t değerleri
0	-168.4888	-.0012	0	-3.5028	
1	-.7695	-1.2325	1	-.0460	-2.619
2	37.4988	.0888	2	1.5966	1.981
3	77.0129	.0011	3	.3431	.629
4	151.7462	.0321	4	2.2464	4.610
5	-105.2861	-.0008	5	1.1313	1.304
6	78.1194	.0010	6	.2395	.423
7	-35.2871	-.0282	7	-.4327	-.853
8	69.1988	.0216	8	.9400	1.622
9	92.2411	.0355	9	3.4030	12.301
10	-83.6954	-.0007	10	-1.6415	-2.781
11	-29.2988	-.0162	11	-2.7914	-6.164
12	-.1049	-.0476	12	0.4459	1.034
13	-97.5252	-.0061	13	-.8572	-.963
14	68.8849	.0005	14	1.0104	.797
15	-72.0696	-.0008	15	-1.5190	-1.364
16	-100.8864	-.0016	16	-3.7035	-5.169

Sınıflandırma çizelgesi

Lojistik modelin

	3	4
3	64	1
4	0	60
Toplam	64	61

Diskriminant fonksiyonunun

	3	4
3	59	8
4	5	53
Toplam	64	61

Modellerin etkinlik ki-kare değerleri ise sıra ile 115.7, 162.6 ve 164.3 olarak elde edilmiştir. Bu değerler 16 serbestlik derecesinde anlamlı olup modellerin geçerliliğini göstermektedir. Ayrıca her bir modelin karşılık gelen 32 değişkenli lojistik modeller ile karşılaştıtırılması sapma ölçütü ile yapılmış ve bu değerler sıra ile $D_1=24.75$, $D_2=2.07$, $D_3=3.84$ olarak elde edilmiştir.

Ölçütlerin her biri 16 serbestlik dereceli ki-kare tablo değeri ile karşılaştırıldığında seçilen 16 değişkenin anlamlı ve yeterli olduğu görülmüştür. Çizelgeden lojistik model ve diskriminant fonksiyonlarının sınıflandırma çizelgeleri incelendiğinde, 1-4, 2-4 ve 3-4 karşılaştırmaları için, lojistik modelin doğru sınıflandırma oranları sıra ile 0.95, 1.0 ve 0.99 iken diskriminant fonksiyonları için bu oranların 0.91, 0.97 ve 0.90 olduğu görülmektedir. Böylece yanıt değişkeninin kesikli olduğu durumda varsayımlar gereği kullanılması gereken lojistik modelin ayrımsama gücünün diskriminant fonksiyonundan daha yüksek olduğu söylenebilmektedir.

Her üç lojistik model için doğru sınıflandırma oranlarının ağırlıklı ortalama değeri 0.98 olup, bu değer 16 değişkenli çoklu grup lojistik modelin doğru sınıflandırma oranı olan 0.93 değerinden daha yüksektir. Bu da Begg ve Gray'in ikili lojistik regresyon analizi yaklaşımının çoklu grup lojistik yaklaşımından daha iyi ayrımsama gücüne sahip modellere götürüldüğünü göstermektedir.

Öte yandan ikili lojistik modellerin her biri için değişkenlerin farklı etkileri olabileceği düşüncesi ile bunlara 32 değişken üzerinden ayrı ayrı değişken seçimi uygulanmış ve önemli bulunan değişkenler Çizelge 3.9'da açıklanmıştır.

Buradan, son iki lojistik model için X3, X21, X22 ve X32 değişkenleri ortak iken, diğer modeli etkileyen değişkenlerin bu modelden farklı olduğu görülmektedir. Örneğin, (1-4) karşılaştırması için iskemik kalp hastalığının özelliğinden dolayı kalp sesi ile ilgili değişkenler modelde yer almaz iken, (2-4) ve (3-4) karşılaştırmalarında romatizmal ve konjenital hastalıkların özellikleri gereği önemli rol oynamaktadırlar. Böylece her bir model için seçilen değişkenler medikal olarak da anlamlıdır.

Bundan sonraki aşamada, her bir karşılaştırma için seçilen değişkenler üzerinden gerek lojistik modeller gerekse diskriminant fonksiyonlar elde edilmiştir. Lojistik model-

Çizelge 3.9. İkili lojistik modeller için değişken seçimi uygulaması sonucu elde edilen değişkenler:

a) 1-4 karşılaştırması için

Seçilen değişken no	Değişken adı
X2	Cinsiyet
X5	Şeker
X26	Ritm problemi
X27	Hipertrofi
X28	
X29	Enfaktüs
X30	

b) 2-4 karşılaştırması için

Seçilen değişken no	Değişken adı
X3	Akut
X8	Anne-babada kalp
X21	Kalp sesisin yeri
X22	
X24	Sesin yayılımı
X32	Laboratuar bulguları

c) 3-4 karşılaştırması için

Seçilen değişken no	Değişken adı
X3	Akut
X4	Önceden geçirilen kalp hastalığı
X7	Diğer hastalıklar geçirme durumu
X10	Sigara
X21	Kalp sesinin yeri
X22	
X23	Sesin şiddeti
X25	Karaciğer durumu
X32	Laboratuar bulguları

lerin etkinlik ki-kare değerleri incelendiğinde seçilen değişkenlerin anlamlı olduğu görülmüştür. Ayrıca her üç karşılaştırma için indirgenen ve tam (32 değişkenli) modelleri kıyaslayan sapma ölçütleri sıra ile D_1=36.34, D_2=11.55, D_3=11.13 olarak elde edilmiştir. Bu değerler de ilgili serbestlik derecelerinde (24, 26, 23) seçilen değişkenlerin anlamlılığını göstermektedir. İndirgenmiş modeller,32 değişkenli tam modelden farklı olmadıklarından tercih edilmektedirler.

Bunun yanı sıra,lojistik modellerin doğru sınıflandırma oranları üç karşılaştırma için sıra ile 0.91, 0.96, 0.97 iken diskriminant fonksiyonlarında bu oranlar 0.86, 0.95 ve 0.91 olarak elde edilmiştir. Diskriminant fonksiyonundan daha iyi ayrımsama gücüne sahip lojistik modellerin ortalama doğru sınıflandırma oranı 0.95'tir. Bu oran değişken sayısındaki azalmalara rağmen yüksek bir ayrımsama gücünün göstergesidir.

Buraya kadar incelenen ikili grup lojistik regresyon analizine ilişkin temel uygulama sonuçları Çizelge 3.10'da özetlenmektedir. Çizelge 3.11'de ise çeşitli modeller arasında seçim yapmaya yönelik sapma değerleri verilmektedir. Çizelgeler incelendiğinde her bir ikili karşılaştırma için ayrı değişken seçimi uygulamanın daha parsimoni modellere götürmekte olduğu görülmektedir. Ancak modellerin etkinlik ki-kare değerleri ve doğru sınıflandırma oranları kriter alındığında,çoklu grup lojistik modelde değişken seçimi sonucu elde edilen 16 değişkenle kurulan ikili lojistik modellerin daha iyi olduğu sonucuna varılmaktadır.

Böylece,Begg ve Gray'in ikili lojistik regresyon analizi yaklaşımı ile elde edilen lojistik modeller çoklu grup lojistik modellerden daha iyi ayrımsama sağlamaktadır. Ancak 2.6.4. alt bölümünde verilen bilgiler ışığında, çalışmanın amacı verilere uyan en iyi lojistik modeli bulmak ve bu model yardımı ile ileriye yönelik atamalar yapmak olduğundan, çoklu grup lojistik modelin tercih edilmesi

Çizelge 3.10. Begg ve Gray'ın yaklaşımı ile ikili grup karşılaştırmaları için özet bilgiler

Modeldeki Değişken sayısı	H_s-H_g Grup no'ları	Log-Olabilirlik	Etkinlik ki-kare değeri	Serbestlik Derecesi (p(g-1))	Lojistik modelin Doğru Sınıflandırma Oranı %	Diskriminant fonksiyonunun doğru sınıflandırma Oranı %
31	1-4	-7.688454	136.03	31	100	99
32	2-4	-6.413352	156.99	32	100	100
32	3-4	-2.552668	162.50	32	100	99
Tüm gruplar için 32 değişken üzerinden seçilen) 16	1-4	-20.063881	115.67	16	95	91
	2-4	-5.377943	162.60	16	100	97
	3-4	-4.473141	164.27	16	99	90
(Her bir grup için 32 değişken üzerinden seçilen) 7	1-4	-25.858248	97.08	7	91	86
6	2-4	-12.189863	157.85	6	96	95
9	3-4	-8.119325	156.78	9	97	91

Çizelge 3.11. 32 değişken üzerinden seçilen çeşitli değişken kümeleri için, ikili grup karşılaştırmalarına ilişkin sapma ölçüt değerleri (sd):

	32 değişken üzerinden		
	(1-4)	(2-4)	(3-4)
32 değişken üzerinden seçilen 16 değişken	24.75 (15)	2.07 (16)	3.84 (16)
Her bir modele ayrı uygulama sonucu seçilen değişkenler 7	36.34 (24)		
6		11.55 (26)	
9			11.13 (23)

$\chi^2_{(0.05;15)}=24.99$; $\chi^2_{(0.05;16)}=26.29$; $\chi^2_{(0.05;23)}=35.17$ $\chi^2_{(0.05;24)}=36.42$;

$\chi^2_{(0.05;26)}=38.89$

gerektiği söylenebilir.

Öte yandan çoklu grup ve ikili lojistik regresyon analizlerinin yanı sıra, yanıt değişkeninin her bir düzeyini ayrı ayrı incelemek ve tüm bu inceleme sonuçlarını birleştirerek çoklu grup lojistik model sonuçları ile karşılaştırmak amacı ile her bir grubu kalan tüm grupların birleşimine karşı inceleyen dört ayrı ikili grup lojistik regresyon analizi de uygulanmıştır. Elde edilen ilk lojistik model, iskemik grubu kalan üç grup ile, ikincisi, romatizmal grubu kalan üç grup ile, üçüncüsü, konjenital grubu kalan üç grup ile, dördüncüsü ise, diğer grubu kalan üç grup ile karşılaştırmaktadır. Bu modeller için doğru sınıflandırma oranları sıra ile 0.96, 0.96, 0.90 ve 0.90 iken, bu oranın dört model üzerinden ağırlıklı ortalama değeri 0.93'dür. Bu da çoklu grup lojistik model ile aynıdır.

Bu yaklaşımda gözlemlerin dört gruptan birisine atanması, her bir modelde ayrı ayrı sınanmasından sonra gerçekleşmektedir. Özellikle yanıt değişkeninin belirli bir düzeyi üzerinde duruluyor, geri kalan düzeylerin hepsi aynı önemde değerlendiriliyorsa, bu tür ikili karşılaştırmalar çoklu grup lojistik model ile aynı ayrımsama gücünde olduğu için tercih edilebilir. Ancak gözlemlerin atanması grup sayısı çok olduğunda uzun zaman alacaktır. Ayrıca, Begg ve Gray yaklaşımında olduğu gibi, amaç katsayı kestirimi ve çıkarsama olduğunda tercih edilmeyecektir.

Çalışmada tüm bu analizlere ek olarak standartlaştırılmış veriler üzerinden (seçilen 16 değişken için) doğrusal regresyon analizi de yapılmıştır. Burada amaç, yanıt değişkeninin kesikli olduğu durumda varsayım bozulumu nedeni ile doğrusal regresyon modelinin yetersizliğini göstermektir. Verileri standartlaştırmanın nedeni ise, değişkenlerin kesikliliğinden kaynaklanan sorunları gidermek ve de çoklu grup lojistik modelin yarı-tam bölünmeden dolayı standartlaştırılmış verilerden elde edilmiş olmasından dolayı, karşılaştırmanın yapılabilmesi içindir.

Standartlaştırılmış verilerin doğrusal regresyon analiz sonucu elde edilen katsayı kestirimleri, t değerleri ve modelin F değeri,Çizelge 3.12'de verilmektedir.

Çizelge 3.12. Standartlaştırılmış veriler üzerinden doğrusal regresyon analizi sonuçları

Değişken No	Katsayı	t değerleri
Sabit	2.540	51.292
1	0.001	0.057
2	0.262	2.678
3	-0.652	-7.307
4	-0.177	-2.172
5	-0.166	-1.561
6	-0.079	-0.902
7	0.152	1.419
8	-0.010	-0.094
9	-0.221	-3.141
10	0.235	2.548
11	0.303	3.607
12	0.027	0.341
13	-1.159	-10.012
14	-0.918	-7.581
15	-0.605	-4.810
16	-0.009	-0.083

VARYANS ANALİZİ $R^2=0.712$

Kaynak	KT	SD	KO	F
Regresyon	147.253	16	9.203	15.012
Artık	142.847	233	0.613	

Çizelgede verilen katsayılar yardımı ile yanıt değişkeninin her bir düzeyi için, $i=1,..,4$ olmak üzere,

$$\bar{Z}_i=b_{1i}\bar{X}_1+...+b_{16,i}\bar{X}_{16} \quad (3.1)$$

ortalama değerleri elde edilmiştir. Burada amaç, $\bar{Z}_i$ değerlerinden bulunacak sınır değerlerini kullanarak, diskriminant analizinde olduğu gibi (2.8. alt bölümünde değinildiği gibi) gözlemleri gruplara atamaktır. Elde edilen $\bar{Z}_i$ değerleri dört grup için sıra ile 1.701, 2.017, 2.551 ve

2.569'dur.

Sınır değerler ise, i=1,2,3 için,

$$C_i = \frac{n_i \bar{Z}_i + n_{i+1} \bar{Z}_{i+1}}{n_i + n_{i+1}} \qquad (3.2)$$

eşitliğinden yararlanarak, $C_1=1.88$, $C_2=2.27$, $C_3=2.56$ şeklinde elde edilmiştir. Böylece $\hat{Y}$ kestirim değerleri için atama kuralı şöyledir; j'nci gözlem (j=1,..,250)

$\hat{y}_j \leq 1.88$ ise 1. gruba

$1.88 < \hat{y}_j \leq 2.27$ ise 2. gruba

$2.27 < \hat{y}_j \leq 2.56$ ise 3. gruba

$\hat{y}_j > 2.56$ ise 4. gruba

atanmaktadır.

Bu şekilde atanan gözlemler gerçek değerleri ile karşılaştırılmış ve doğrusal regresyon modelinin doğru sınıflandırma oranı 0.50 olarak elde edilmiştir. Böyle düşük bir oran, doğrusal regresyondaki varsayım bozulumundan dolayı beklenen bir sonuçtur. Bu durum, çalışmada lojistik regresyon analizine olan gereksinimin bir göstergesidir.

Benzer şekilde orijinal verilere de yanıt değişkeni sürekliymiş gibi doğrusal regresyon analizi uygulanmış ve aynı yolla yapılan atamalar sonucunda, modelin doğru sınıflandırma oranı 0.49 bulunmuştur. Yukarıda yapılan yorumlar burada da geçerlidir.

3.2. Öğrenci Seçme Sınavına İlişkin Veriler Üzerine Uygulama

Yükseköğretim, lise ve dengi okullar üzerine en az iki yıl eğitim veren yükseköğretim kuruluşlarındaki öğretimdir. Genel olarak üniversite adını alan bu kuruluşlardan yararlanma olanağı,kontenjan kısıtı nedeni ile sınırlıdır. Bu nedenle üniversitelerde okuma şansını elde etmek için sıralama sınavı uygulanmakta ve sınavda belirli başarıyı sağlayanlar yükseköğrenime kabul edilmektedir.

Yükseköğretim programlarına alınacak adayların seçimi ve tercihlerine göre yerleştirilmeleri "Öğrenci Seçme ve Yerleştirme Merkezi (ÖSYM)" tarafından iki basamaklı bir sınav sistemi uygulayarak yapılmaktadır. Sistemin genel adı "Öğrenci Seçme ve Yerleştirme Sınavı (ÖSYS)" dır. Öğrenci Seçme Sınavı (ÖSS) olarak adlandırılan birinci basamak sınavının amacı, ikinci basamak sınavına girebilecek adayları seçmenin yanı sıra,bu sınav sonucuna göre öğrenci alan bazı yükseköğretim programlarına öğrenci yerleştirmektir. Öğrenci Yerleştirme Sınavı (ÖYS) olarak adlandırılan ikinci basamak sınavında ise amaç, birinci basamak sınavını kazanan adayları,ikinci sınav sonuçlarına göre öğrenci alan programlara yerleştirmektir.

ÖSS'de adaylara iki bölümden oluşan test uygulanmaktadır. Sözel olarak adlandırılan birinci bölümde, adayların Türkçe'yi kullanma güçleri ile sosyal bilimlerdeki temel kavramlarla düşünme, sayısal olan ikinci bölümde ise, matematik ve fen bilimlerindeki ilişki ve kavramlarla düşünme becerileri sınanmaktadır.

ÖYS'de ise adaylara fen bilimleri, matematik, Türkçe, sosyal bilimler ve yabancı dil (İngilizce, Fransızca, Almanca) dallarında testler uygulanmaktadır.

ÖSS'yi kazanan adaylar yalnız bu sınav sonucuna göre öğrenci alan yükseköğretim programlarına girmek istedikleri taktirde, ikinci basamak sınavına katılmaksızın puanlarının elverdiği programlara yerleştirilebilmektedirler (ÖSYS

1989 kılavuzu).

Bu alt bölümde amaç, kişinin hayatında önemli bir yer alan yükseköğrenim fırsatını elde etmede etkili olabileceği düşünülen etkenleri içeren verilere,lojistik regresyon analizi uygulamaktır. Bulunacak çoklu grup lojistik model yardımı ile,adayların çeşitli bilgileri doğrultusunda, tercih ettikleri yükseköğretim programlarından hangisine girebilecekleri tahmin edilecek ve bu tahmin mümkün olduğunca az bilgi ile en doğru şekilde yapılmaya çalışılacaktır.

3.2.1. Çalışmada incelenen değişkenler

Yukarıda değinilen amaçlar doğrultusunda ÖSYM'den 1987-1988 yılında ÖSS sınavına giren adaylardan rastgele 1000 tanesine ilişkin bilgiler alınmıştır. Bu bilgiler, adayların sınavdaki başarılarını etkileyebileceği düşünülen değişkenler şeklinde olup şöyle tanımlanmaktadır:

X1: Cinsiyet. Kesikli olup, 1: Erkek, 2: Kadın şeklinde kodlanmıştır. ÖSS sınavında başarı durumunun cinsiyete göre değişebileceği düşünülerek çalışmaya alınmıştır.

X2: Doğum tarihi. Sürekli olup doğum yılının son iki rakamı ile verilmektedir. Yeni lise mezunlarının bilgilerinin daha taze olacağı düşüncesi ile başarıya etki eden etkenlerden biri olarak alınmıştır.

X3: Oturulan il. Sürekli olup, 67 ilin trafik kodları ile verilmektedir (bilgiler alındığında 67 il mevcut idi). ÖSS sınavında başarıyı lise eğitimi düzeyi önemli derecede etkilemektedir. Her ilde lise eğitimi düzeyleri farklı olduğu için oturulan ilin başarıyı etkileyeceği düşünülmüştür.

X4: Mezun olunan okul türü. Kesikli olup ÖSYM kılavuzlarında verilen bilgiler doğrultusunda 15 düzeylidir.
Bu düzeylere göre kodlama şöyledir; 1: Anadolu liseleri, 2: Modern fen liseleri, 3: Öğretmen liseleri, 4: Dört yıllık öğretmen liseleri, 5: Klasik fen liseleri, 6: İmam ha-

tip liseleri, 7: Askeri liseler, 8: Polis koleji, 9: Teknik liseler, 10: Normal lise ve dengi okullar, 11: Endüstri meslek liseleri, 12: Kız meslek liseleri, 13: Sağlık liseleri, 14: Ticaret liseleri, 15: Diğer meslek liseleri.

Aynı bölgede eğitim veren lise ve dengi okulların eğitim düzeyleri farklılık göstermektedir. Bu durumda mezun olunan okul türünün başarıyı etkileyebileceği düşüncesi ile çalışmaya alınmıştır.

X5: ÖSS sınavına giriş sayısı. Kesikli olup, 0: Yanıt yok, 1: İlk kez, 2: İkinci kez, 3: Üçüncü kez, 4: Dört ve daha çok kez şeklinde kodlanmıştır.

Sınava giriş sayısı üzerinde bir kısıt yoktur ve aday isterse her yıl şansını deneyebilmektedir. Bir kaç kez sınav denemesinin adaya tecrübe kazandıracağı ve dolayısı ile başarıyı etkileyebileceği düşüncesi ile çalışmaya alınmıştır.

X6: İkamet yerinin türü. Kesikli olup, 0: Yanıt yok, 1: Büyük kent, 2: Kent, 3: İlçe, 4: Nahiye, 5: köy şeklinde kodlanmıştır.

Adayların yaşadıkları yerin kentleşmiş olması daha iyi ve düzeyli eğitim olanaklarını da beraberinde getirmektedir. Bu olanaklar da ÖSS sınavında başarı kazanmada önemli bir etken teşkil ettiğinden çalışmaya alınmıştır.

X7: Adayın üniversiteye hazırlık kursu görüp görmediği. Kesikli olup, 1: Gördü, 0: Görmedi şeklinde kodlanmıştır.

Hazırlık kurslarının adaylara sınav pratiği ve alışkanlığı sağladığı düşüncesiyle çalışmaya alınmıştır.

X8: ÖSS puanı. Sürekli olup ikinci sınava katılabilmek için alt sınırı 105 olarak belirlenmiştir.

3.2.2. Uygulama sonuçları

ÖSYM'den alınan bilgiler arasında adayların kaçıncı tercihlerini kazandıkları bilgisi de yer almaktadır. Bu bilgiler yanıt değişkenini oluşturmaktadır. Adaylar ÖSS ve ÖYS sonuçlarına göre öğrenci alan yükseköğretim programları arasından toplam 24 tercih yapabilmektedirler. Ayrıca merkezi yerleştirme sonucu, Anadolu Üniversitesi Açık Öğretim Fakültesi "Merkezi Açık Öğretim Programlarına" kontenjan açığı kaldığı taktirde ÖSS'yi kazanamayan istekli adaylardan ÖSS puan (en az 100) sırasına göre kontenjan açığı kadar yerleştirme yapılmaktadır. Adaylar bu isteklerini ve tercihlerini sınavda cevap kağıdında belirtmektedirler. Böylece açık öğretim de 25'nci tercih olarak alınmış ve sonuçta yanıt değişkeni sıra ile kazanamadı (0) ve 1-25'nci tercihlerini kazandı şeklinde 26 düzeyli olarak kodlanmıştır.

Verilere çoklu grup lojistik regresyon analizi uygulamadan önce, yanıt değişkeninin düzey sayısının, çok olması nedeni ile indirgenmesi yoluna gidilmiştir. ÖSYM'ye danışılarak bazı düzeyler birleştirilmiş ve 1: kazanamadı, 2: 1'nci tercihini kazandı, 3: 2'nci tercihini kazandı, 4: 3 veya 4'ncü tercihini kazandı, 5: 5-8 arası tercihlerinden birini kazandı, 6: 9-13 arası tercihlerinden birini kazandı, 7: 14-18 arası tercihlerinden birini kazandı, 8: 19-24 arası tercihlerinden birini kazandı, 9: 25'nci tercihini kazandı şeklinde 9 düzeye indirgenmiştir.

Her bir yanıt düzeyi için ÖSS puanını gösteren değişkenin (X8) ortalama değerleri incelendiğinde bazılarının birbirine çok yakın olduğu görülmüştür. Bu yakın değerli düzeyler tekrar birleştirilerek yanıt değişkeni, 1: kazanamadı ya da 25'nci tercihini kazandı, 2: 1 ya da 2'nci tercihini kazandı, 3: 3-13 arası tercihlerinden birini kazandı, 4: 14-24 arası tercihlerinden birini kazandı şeklinde 4 düzeyli olarak yeniden tanımlanmıştır.

Öte yandan açıklayıcı değişkenler arasında kategorik olanlar için düzey sayısının bir eksiği kadar göstermelik değişken eklenmesi zorunluluğu nedeniyle; Mezun olunan okul türü (X4) değişkeni 15 düzeyli olduğu için 14 tane, ÖSS sınavına giriş sayısı değişkeni (X5) 5 düzeyli olduğu için 4 tane, ikamet edilen yer değişkeni (X6) 6 düzeyli olduğu için 5 tane göstermelik değişkenle ifade edilmiştir. Böylece toplam 28 açıklayıcı değişken elde edilmiştir.

Bin gözlemli bu yeni veri kümesine lojistik regresyon analizi uygulanmış ancak,X4 değişkeninin bazı düzeyleri için sıfır ya da çok az sayıda gözlem olmasından dolayı katsayı kestirimleri elde edilememiştir. Bunun için ilgili değişkenin 4, 7, 8, 10 ve 14'ncü düzeyleri tek bir kategoride toplanarak düzey sayısı 10'a indirgenmiştir. Sonuçta toplam değişken sayısı 24'e düşmüştür.

24 değişkenli verilere yeniden lojistik regresyon analizi uygulanmış, elde edilen modelin katsayı kestirimleri ve sınıflandırma çizelgesi Çizelge 3.13. a ve b'de verilmiştir.

Çizelge 3.13. 24 değişkenli ÖSS verilerine ilişkin çoklu grup lojistik modelin,

a) Katsayı kestirimleri (sutunlar: gruplar, satırlar: değişkenler)

	1	2	3
0	3.6234	-.3882	2.3277
1	.1193	.0356	.1820
2	-.0482	-.1031	-.1137
3	.0069	.0236	-.0007
4	-.2862	.1850	-.1568
5	.0962	-.7946	-.9212
6	.4571	-8.9936	-.7472
7	.4152	-1.0518	-.9652
8	.2021	-.6027	-.9000
9	1.0985	-8.4463	.5625
10	-.1528	-.6279	-.4865
11	.3750	1.3762	-1.0200
12	-.8104	-10.0227	-1.8294
13	.0062	.2482	-.0164
14	1.1600	-7.4903	.5803
15	1.2084	1.6324	1.0863
16	.9717	.7242	.9168
17	.8481	.2676	.6850
18	-.0153	.5354	-.2374
19	-.0718	.4296	-.0323
20	.0067	-.5338	.1004
21	-.0419	-.0244	.3494
22	.5773	.5982	.5060
23	-.0327	-.0819	.1349
24	-.0062	.0369	.0409

b) Sınıflandırma çizelgesi (sütunlar: gerçek gruplar, satırlar: atanan gruplar)

	1	2	3	4
1	507	28	185	152
2	0	0	0	0
3	37	10	59	19
4	1	0	1	1
Toplam	545	38	245	172

Uygulamada ÖSS puanı değişkeninin (X8) ortalama değeri, yanıt değişkeninin her bir düzeyi için sıra ile 121.95, 128.35, 127.88 ve 122.45 olarak elde edilmiştir. 1 ile 4 ve 2 ile 3 gruplarının puanlarındaki yakınlık, kazanamayanlar ile 14-24 arası tercihlerini kazananların ve ilk iki tercihi ile 3-13 arası tercihlerini kazananların puanlarının çok yakın olması şeklinde yorumlanabilmektedir. Böylece adayların tercih sıralamalarında rasyonel olmadıkları ve ilk tercihlerini kazanabilecek iken,sıralama hatası yüzünden alt sıralardaki tercihlerini kazandıkları,hatta kazanamadıkları söylenebilmektedir.

Çizelge 3.13 incelendiğinde, lojistik modelin doğru sınıflandırma oranının 0.57 olduğu görülmektedir. Yalnız kazanamayanlar grubu (1) için doğru sınıflandırma oranı yüksek iken,diğer gruplar için modelin ayrımsaması pek iyi değildir. Özellikle 14-24 arası tercihlerini kazanan grubun (4) büyük bir kısmı kazanamadı grubuna atanmıştır. Buradan,13'ncü sıradan sonraki tercihlerin rastgele olarak ve kazanamama korkusu ile yapıldığı sonucuna varılabilmektedir.

Ayrıca lojistik modelin etkinlik ki-kare değeri 137.93 olarak bulunmuştur. 72 serbestlik derecesinde önemli olan bu değer 24 değişkenli modelin anlamlılığını göstermektedir.

Çalışmada önemsiz olabilecek değişkenleri elemek ve adaylara ilişkin mümkün olduğunca az bilgi kullanarak hangi tercihlerini kazanabileceklerini doğru olarak tahmin etmek amacı ile değişken seçimi uygulanmıştır. Uygulamada X5, X7, X15, ve X24 değişkenleri önemli bulunarak seçilmiştir. Bu değişkenler sıra ile mezun olunan okul türünün iki düzeyi, sınava giriş sayısının bir düzeyi ve ÖSS puanına karşılık gelmektedir. Buna göre adayların mezun oldukları okul türünün sadece modern ya da klasik fen lisesi olması, sınava ikinci kez girip girmediği önemli olup, diğer düzeyler bir arada düşünülebilmektedir. Böylece kalan düzeyler birleştirilerek yeniden düzenlenmiş ve sonuçta toplam değişken sayısı 13'e indirgenmiştir. İndirgenen veri kümesine önce lojistik regresyon analizi uygulanmış ve modelin doğru sınıflandırma oranı 0.56 bulunmuştur. Modelin etkinlik ki-kare değeri 108.72 olup, 39 serbestlik derecesinde anlamlıdır.

Daha sonra, uygulanan değişken seçimi ile yine mezun olunan okul türü (her iki düzeyi de), sınava giriş sayısı ve ÖSS puanı olmak üzere 4 değişken önemli bulunmuştur. Seçilen her bir değişken için modelin log-olabilirlik değeri, ki-kare ölçütü, birikimli doğru sınıflandırma oranları aşağıda Çizelge 3.14'de verilmektedir.

Çizelge 3.14. 13 değişken üzerinden kurulan lojistik modele değişken seçimi uygulaması sonuçları

Log-olabilirlik	Seçilen değişken no	Ki-kare	P	Birikimli doğru sınıflandırma %
-1077.458934	13	49.93	.0000	.54
-1068.257540	5	18.40	.0004	.56
-1063.542302	4	9.43	.0241	.55
-1060.203303	6	6.68	.0829	.55

Çizelgeden de görülebileceği gibi modele anlamlı olarak alınan değişkenler ayrımsama gücünde azaltma yaratabilmektedirler. Böylece doğru sınıflandırma oranının, lojistik modelin uyum iyiliği ölçütü olarak yeterli olmadığı söylenebilmektedir.

Bundan sonraki aşamada, seçilen 4 değişken üzerinden lojistik regresyon analizi tekrar uygulanmıştır. Analiz sonucu elde edilen modelin katsayı kestirimleri ve Z değerleri ile sınıflandırma çizelgesi Çizelge 3.15. a ve b'de verilmektedir.

Çizelge 3.15. Seçilen 4 değişken üzerinden lojistik modelin,

a) Katsayı kestirimleri ve Z değerleri (sütunlar: gruplar, satırlar: değişkenler)

	Katsayı Kestirimleri		
	1	2	3
0	-9.0844	-9.0844	-9.0844
1	-.1472	-.3623	-.2756
2	9.9215	7.1334	9.8022
3	.1206	.0075	-.2405
4	.0070	.0168	.0009

	Z Değerleri		
	1	2	3
0	-.0967	-.0967	-.0967
1	-.7887	-.8990	-1.2802
2	.1056	.0759	.1044
3	.6083	.0183	-1.0339
4	1.4323	1.6788	.1552

b) Sınıflandırma çizelgesi (sütunlar gerçek gruplar, satırlar: atanan gruplar)

	1	2	3	4
1	545	38	245	171
2	0	0	0	0
3	0	0	0	0
4	0	0	0	1
Toplam	545	38	245	172

Buradan, lojistik modelin doğru sınıflandırma oranının 0.55 gibi düşük bir değer olup, 2, 3 ve 4'ncü gruplardan 1'nci gruba büyük oranda yanlış atamaların olduğu görülmektedir.

Öte yandan modelin etkinlik ki-kare değeri 14.46 olarak elde edilmiştir. Bu değer 12 serbestlik derecesinde önemli olmayıp, seçilen 4 değişkenli modelin anlamlı olmadığının bir göstergesidir. Ayrıca bu modeli 13 değişkenli lojistik model ile karşılaştıran sapma ölçütü de anlamlı bulunmuştur (D=94.26, 27 serbestlik derecesinde anlamlı). Böylece 13 değişkenli model tercih edilmektedir. Ancak bu model tercih edilse bile doğru sınıflandırma oranı 0.56 gibi yine oldukça düşük bir değerdir.

Bu sonuç, lojistik modelin ayrımsamada yetersizliği olarak yorumlanmamalıdır. Çünkü elde edilen katsayı kestirimleri istatistiksel özellikleri sağlamaktadır ve model istatistiksel olarak anlamlıdır. Ayrımsama gücünün zayıflığı tercih sıralamalarının birbirlerine yakınlığından kaynaklanmaktadır. Bu durum, ÖSS'ye giren adayların tercih sıralamalarında rasyonel olmadıklarını, bilinçsizce ve kazanamama korkusu ile yaptıkları yanlış sıralamanın başarısızlıklarına yol açtığını göstermektedir.

Çoklu grup lojistik modelin iyi ayrımsama yapamadığı bu durumda genelleştirilmiş doğrusal modeller ailesinin diğer elemanları olan alternatif modellerin sınanması gerektiği söylenebilir. Bu yoruma, Hosmer ve Lemeshow'un (1989b), lojistik modellerin yanlış sınıflandırma olasılıklarının çok büyük olduğunda alternatif modellerin sınanması gerektiği şeklindeki önerilerine dayanarak gidilmektedir. Ancak vurguladıkları gibi bu alternatiflerin seçiminde çok dikkatli olunması gerekmektedir.

Tüm bunların yanı sıra çalışmada yanıt değişkenini kazanamayanlar (kazanamadı +23+24+25'nci tercihleri kazandı) ve kazananlar (1-22 arası tercihlerini kazananlar) şeklinde iki düzeyli olarak ifade ederek, iki grup lojistik regresyon analizi de uygulanmıştır. Elde edilen lojistik modelin doğru sınıflandırma oranı 0.64 olarak bulunmuştur. Ancak burada ilk ve 22'nci tercihini kazananlar aynı gruba yerleştirildiğinden bilgi kaybı olmaktadır ve detaylı bilgiler gözden kaçmaktadır. Bu nedenle çoklu grup lojistik model uygulaması daha iyi bir yaklaşımdır.

Çalışmanın bu kısmında amaç, ÖSS verilerine uyan en iyi modeli bulmaktan çok, lojistik modelin bazen yetersiz kaldığını ve bu yetersizliğin de verinin durumundan kaynaklandığını bir örnekle göstermekti. Bu nedenle verilere, Begg ve Gray yaklaşımı, diskriminant analizi ve etkileşim terimlerini katarak yeni model seçimi uygulanmamıştır. Ayrıca aynı düşünce ile bulunan modelin artık terimleri de incelenmemiştir.

4. SONUÇ VE TARTIŞMA

Anket türü verilerde kesikli yanıt olması ve açıklayıcı değişkenlere çok sık rastlanması, amaç regresyon analizi olduğunda regresyon modelini kurmayı, ayrımsama olduğunda da diskriminant analizi ile diskriminant fonksiyonunu oluşturmayı yetersiz kılmaktadır. Böyle durumlarda alternatif olarak lojistik regresyon analizi kullanılmaktadır.

Bu çalışmada yanıt değişkeninin kesikli olduğu iki ayrı veri kümesi için lojistik regresyon modeli kurulmuş ve en iyi ayrımsamayı verecek model araştırmasına gidilmiştir.

Bu amaçla çalışmanın ikinci bölümünde, gerek ikili gerekse çoklu grup lojistik regresyon modelleri tanımlanmış, kestirim yöntemleri incelenmiş, diskriminant analizi ile kıyaslamasının yanı sıra artıklar ve teşhis ölçütleri de konu edilmiştir.

Uygulamaların yer aldığı üçüncü bölümde ise kardiyolojik verilerle ÖSS'ye giren adaylara ilişkin verilerin analizinde lojistik model denenmiştir.

Dört ayrı kalp hastalık türünün yanıt değişkeninin düzeylerini oluşturduğu kardiyolojik verilere yapılan uygulamada, en iyi lojistik modele ulaşmak hedeflenmiştir. Kesikli açıklayıcı değişkenlerin ikiden çok düzey içermesi nedeniyle, her biri için tanımlanan göstermelik değişkenler de eklenip, gerekli düzenlemeler yapılarak 32 değişkenli tam modele ulaşılmıştır. Bu modele değişken seçimi uygulanarak önemli bulunan 16 değişkenin oldukça iyi açıklama getirdiği ve ayrımsama yaptığı ayrıca, medikal olarak da son derece anlamlı olduğu görülmüştür. Yapılan artık incelemesi sonucunda rastlanan aykırı değerlerin verilerdeki yarı-tam bölünmeden kaynaklandığına, bu nedenle önemli bir aykırılığın söz konusu olmadığına karar verilmiştir.

Veri kümesindeki açıklayıcı değişkenler arasında hem sürekli hem de kesikli olanlarının bulunması nedeni ile, etkileşim terimlerinin de modelde yer alıp almayacağı sınanmak istenmiştir. Önce sürekli değişkenlerin modelde doğrusal etkisi olduğu gözlenmiş, daha sonra kesikli değişkenler arasında, gerek Spearman ilişki katsayılarından gerekse medikal bilgilerden, varolabilecek etkileşimlere karar verilmiştir. Seçilen 16 değişkene eklenen 21 etkileşim terimi ile, toplam 37 değişken üzerinden lojistik model incelenmiştir. Bu model 16 değişkenli yalın etkilerin modeli ile karşılaştırılmış ve etkileşimlerin önemli olmadığı görülmüştür.

Öte yandan etkileşimlerin eklendiği genişletilmiş modele değişken seçimi de uygulanmış ve 15 değişken önemli bulunmuştur. Bunlara ilişkin elde edilen lojistik modelin anlamlı olduğu gözlenmiş, ancak doğru sınıflandırma oranına dayanarak yalın etkilerin oluşturduğu 16 değişkenli çoklu grup lojistik modelin, her ne kadar yarı-tam bölünme durumundan dolayı sonlu kestirim vermese de, ayrımsama amacı ile kullanılabilen en iyi model olarak tercih edilmesine karar verilmiştir.

Çoklu ve ikili grup lojistik modellerin ayrımsama gücünü karşılaştırmak amacı ile, seçilen 16 değişken üzerinden Begg ve Gray yaklaşımı olan ikili lojistik modeller incelenmiştir. Bu amaçla, yanıt değişkeninin diğer grup kalp hastalıklarını içeren dördüncü düzeyi temel alınmış ve diğer üç grup bu temel grup ile karşılaştırılmıştır. Her üç karşılaştırma için, elde edilen ikili lojistik modellerin, çoklu grup lojistik modellerden daha iyi ayrımsama verdiği görülmüştür. Ancak, literatür bilgilerine dayanarak, amaç kestirim ve öngörü olduğunda çoklu grup lojistik modellerin tercih edilmesi gerektiği vurgulanmıştır.

Öte yandan ikili grup lojistik modellerin ve diskriminant fonksiyonunun ayrımsama güçleri de karşılaştırılmış, diskriminant analizi varsayımlarının bozulumu nedeni ile lojistik modellerin daha iyi ayrımsama verdiği görülmüştür.

İkili lojistik modellerin her birinde değişkenlerin etkilerinin farklı olabileceği düşüncesi ile, her bir ikili karşılaştırma için tam model olan 32 değişkenli model üzerinden değişken seçimi uygulanmıştır. Farklı değişkenlerin seçildiği her üç lojistik model de anlamlı bulunmuş, ancak çoklu grup lojistik model uygulamasında seçilen 16 değişkenin üç karşılaştırma için de en iyi modeli verdiği gözlenmiştir.

Tüm bu uygulamalara ek olarak, doğrusal regresyon için gerekli tüm koşulların sağlandığı varsayılarak, gerek standartlaştırılmış, gerekse orjinal verilere doğrusal regresyon analizi uygulanmış ve elde edilen modellerin ayrımsama güçleri çok düşük bulunmuştur. Böylece yanıt değişkeninin, kesikli olduğu durumda, lojistik regresyon analizinin uygulanmasının zorunluluğu vurgulanmıştır.

İkinci uygulama olarak, üniversite sınavına giren adayların seçebilecekleri 25 tercihin yanıt değişkeninin düzeylerini oluşturduğu ÖSS verilerine lojistik regresyon analizi uygulanmıştır. Tercihler dört grupta toplanarak yapılan çoklu grup lojistik regresyon analizinde, elde edilen modelin ayrımsama gücü oldukça düşük bulunmuştur. Uygulanan değişken seçimi sonucunda, seçilen değişkenlerin oluşturduğu model anlamlı bulunmamış, böylece 13 değişkenli tam model tercih edilmiştir. Bu modelin ayrımsama gücünün zayıflığı, adayların tercih sıralamasında rasyonel olmamalarından, kazanamama korkusu ile bilinçsizce sıralama yapmalarından kaynaklanabileceği biçiminde yorumlanmıştır ve bu veriler için, lojistik modelin uygun olmadığı, aynı dağılım ailesinden başka modellerin denenmesi gerektiği sonucuna varılmıştır.

Öte yandan bu uygulamada amaç, verilere uyan en iyi lojistik modeli bulmaktan çok, varsayımların bozulumu nedeni ile uygulanması gereken lojistik regresyon analizinin bazı durumlarda iyi sonuç vermediğini bir örnekle sergilemektir.

Özetle, lojistik modelin, somut ölçümlerin olduğu durumlarda iyi sonuçlar verirken soyut ölçümlere dayanan sosyal uygulamalarda yetersiz kalabildiği söylenebilir.

Çalışmada yapılan bazı zorunlu kısıtlamalar nedeni ile ortaya çıkan eksiklikler ve kuşkular aşağıdaki gibi özetlenebilir:

Elde edilen lojistik modellerde yarı-tam bölünme görülmüş ve yorumlar buna göre yapılmıştır. Pratikte yarı-tam bölünme ile çoklubağlantı durumu kesin olarak ayırt edilememektedir. Bu nedenle verilerde ilk olarak çoklubağlantı durumu araştırılmalı ve böyle durumlarda daha güçlü kestirimler elde etmek için başka yöntemler denenmelidir. Öte yandan Genel Bilgiler kısmında ele alınan kestirim yöntemleri teorik olarak da incelenmeli ve karşılaştırmaları yapılarak hangi yöntemin en iyi olduğuna karar verilmelidir.

Her ne kadar varsayım bozulumlarında lojistik regresyon analizi uygulanması zorunluluksa da uygulamalarda çoklu grup diskriminant analizi de incelenmeli, özellikle karesel ya da daha yüksek dereceden terimleri kapsayan diskriminant fonksiyonları denenerek, bulunacak sonuçlar lojistik modellerle karşılaştırmalıdır.

Uygulamalarda tek bir temel sınıf seçimi için incelemeler yapılmıştır. Ancak farklı temel sınıflar da seçilerek elde edilecek sonuçlar karşılaştırılmalı, böylece en iyi temel sınıfa karar verilmelidir.

Tartışmaya açılabilecek bir nokta da şöyle tanımlanabilir: Lojistik modelde açıklayıcı değişkenlerin kesikli ya da sürekliliği üzerinde bir kısıt yoktur ve model diskriminant fonksiyonu gibi ayrımsama amacıyla kullanılabilmektedir. Bu nedenle kesikli değişkenlerin düzey sayısına bağlı olarak göstermelik değişkenlerle ifade edilmesi yerine, bu değişkenlerin olduğu gibi kullanılması düşünülebilir. Nitekim çalışmada bu durum da sınanmış ve benzer sonuçlara ulaşılmıştır.

Bundan sonraki çalışmalarda, (2.24) eşitliği ile tanımlanan genel formda $h(X,\beta)$'nın X'de doğrusal olmadığı durumda lojistik modelin nasıl uygulanabileceği, katsayıların nasıl kestirileceği incelenebilir. Ayrıca yanıt değişkeminin sıralı olduğu durumda lojistik modellerin incelenmesi de üzerinde durulması gereken bir diğer konudur.

Birden fazla yanıt değişkeninin bulunduğu çok değişkenli lojistik modeller (multivariate logistic models) pratikte çok sık rastlanan bir durumdur. Bu çalışmada ele alınmayan çok değişkenli lojistik modellerin daha sonraki çalışmalarda incelenmesi düşünülmektedir.

Son olarak, lojistik modelde yer alan çok sayıdaki değişkenlerin -özdeğer ve özvektörler ile- daha az sayıya indirgenmesi ve hipotetik değişkenlerle lojistik model durumu,üzerinde durulması düşünülen diğer bir konudur.

DEĞİNİLEN BELGELER DİZİNİ

Albert, A. and Anderson, J.A., 1984, On the existence of maximum likelihood estimates in logistic regression models: Biometrika, 71, 1-10.

Albert, A., 1985, Main program for multiple group stepwise logistic discrimination: Laboratory of Clinical Chemistry,University of Liege, Belgium.

Albert, A., 1986, Program for performing multiple group logistic discrimination: Laboratory of Clinical Chemistry, University of Liege, Belgium.

Albert, A. and Lesaffre, E., 1986, Multiple group logistic discrimination: Computational Mathematics with Applications, 12A, 2, 209-224.

Aldrich, H.J. and Nelson, D.F., 1986, Linear probability logit and probit models: Sage publications, London, 95 p.

Al-Sarraf, Z. and Young, D.H., 1986, The extreme residuals in logistic regression models: Journal of Statistical Computations and Simulation, 25, 115-125.

Amemiya, T., 1981, Qualitative response models: A Survey: Journal of Economic Literature, 19, 1483-1536.

Anderson, J.A., 1972, Separate sample logistic discrimination: Biometrika, 59, 19-35.

Anderson, J.A., 1979, Multivariate logistic compounds: Biometrika, 66, 17-26.

Anderson, J.A., 1982, Logistic Discrimination, Handbook of Statistics, 21, 169-191.

Anderson, J.A., 1983, Robust inference using logistic models: Bulletin of International Statistical Institute, 48, 35-53.

Anderson, T. and Bahadur, R.R., 1962, Classification into two multivariate normal distributions with different covariance matrices: Annals of Mathematical Statistics, 33, 420-431.

Aranda-Ordaz, F.J., 1981, On two families of transformations to additivity for binary responce data: Biometrika, 68, 357-363.

Ashton, W.D., 1972, The logit transformation with special reference to its uses in Bioassay: Griffin, London, 87 p.

DEĞİNİLEN BELGELER DİZİNİ (Devam ediyor)

Bartolucci, A.A. and Frazer, M.D., 1977, Comperative set up and composite tests for selecting prognostic indicators associated with survival: Biometrical Journal, 19, 437-448.

Begg, C.B. and Gray, R., 1984, Calculation of polychotomous logistic regression parameters using individualized regressions: Biometrika, 71, 1, 11-18.

Berk, K.N., 1978, Comparing subset regression procedures: Technometrics, 20, 1-6.

Berkson, J., 1944, Application of the logistic function to bio-assay: Journal of The American Statistical Association, 39, 357-365.

Berkson, J., 1955, Maximum likelihood and minimum X^2 estimates of the logistic function: Journal of The American Statistical Association, 50, 130-162.

Berkson, J.M.D., 1957, Tables for the maximum likelihood estimate of the logistic function: Biometrics, 28-33.

Bishop, M.M., Fienberg, S.E. and Holland, P.W., 1975, Discrete multivariate analysis 'Theory and Practice': MIT Press, Cambridge, 557 P.

Brown, C.C., 1982, On a goodness of fit test for the logistic model based on score statistics: Communications İn Statistics. Theory and Method, 11, 10, 1087-1105.

Carrol, R.J., Spiegelman, C.H., Lan, K.K.G., Bailey, K.T. and Abbott, R.D., 1984, On errors-in-variables for binary regression models: Biometrika, 71, 1, 19-25.

Clunies, R.C.W. and Riffenburg, R.H., 1960, Geometry and linear discrimination: Biometrika, 47, 185-189.

Cook, R.D. and Weisberg, S., 1980, Characterisations of an emprical influence function for detecting influential cases in regression: Technometrics, 22, 495-508.

Copas, J.B., 1983, Plotting p against X: Applied Statistics, 32, 25-31.

Cox, D.R., 1970, Analysis of binary data: Chapman and Hall, London, 142 p.

DEĞİNİLEN BELGELER DİZİNİ (Devam ediyor)

D' Agostino, R.B. and Pozen, M.W., 1982, The logistic function as an aid in the detection of acute coronary disease in emercency patients. (A case study): Statistics In Medicine, 1, 41-82.

Dietz, E., 1987, Application of logistic regression and logistic discrimination in medical decision making: Biometrical Journal, 29, 6, 747-751.

Fienberg, S.E., 1979, The analysis of cross classified categorical data: Second Edition, Minnesota, 198 p.

Fienberg, S.E. and Gong, G.D., 1984, Comment: Journal of The American Statistical Association, 79, 385, Invited papers section, 72-77.

Finney, D.J., 1971, Probit Analysis: Third Edition, Cambridge, University Press, Cambridge, 318 p.

Gnanadesikan, R. et all., 1989, Discriminant analysis and clustering. Panel on discriminant analysis, classification and clustering: Statistical Science, 4, 1, 34-69.

Guerro, V.M. and Johnson, R.A., 1982, Use of the Box-Cox transformation with binary response models: Biometrika, 69, 309-314.

Hosmer, D.W. and Lemeshow, S., 1980, Goodness-of-fit tests for the multiple logistic regression model: Communications in Statistics, Theory and Method, A9(10), 1043-1069.

Hosmer, D.W. and Lemeshow, S., 1989a, Logistic regression diagnostics: A new look: Presented at the annual meeting of the American Statistical Association

Hosmer, D.W. and Lemeshow, S., 1989b, Applied logistic regression: John Wiley and Sons Inc. New York, 307 p.

Hosmer, D.W., Jovanovic, B. and Lemeshow, S., 1989, Best subsets logistic regression: Biometrics, 45, 1265-1270.

Jennings, D.E., 1986a, Judging inference adequacy in logistic regression: Journal of The American Statistical Association, 81, 471-476.

Jennings, D.E., 1986b, Outliers and residual distributions in logistic regression: Journal of The American Statistical Association, 81, 987-990.

DEĞİNİLEN BELGELER DİZİNİ (Devam ediyor)

Johnson, W., 1985, Influence measures for logistic regression: Another point of view: Biometrika, 72, 1, 59-65.

Kay, R. and Little, S., 1986, Assessing the fit of the logistic model: A case study of children with the Haemolytic Uraemic Syndrome: Applied Statistics, 35, 1, 16-30.

Kleinbaum, D.G. and Kupper, L.L. and Muller, K.E., 1988, Applied regression analysis and other multivariable methods: Second Edition. PWS-Kent Publishing, Boston, 718 p.

Knoke, J.D., 1982, Discriminant analysis with discrete and continuous variables: Biometrics, 38, 191-200.

Landwehr, J.M., Pregibon, D. and Shoemaker, A.C., 1984, Graphical methods for assessing logistic regression models: Journal of The American Statistical Association, 79, 385, 61-71.

Lemeshow, S. and Hosmer, D.W., 1982, A review of goodness of fit statistics for use in the development of logistic regression models: American Journal of Epidemiology, 115, 1, 92-106.

Lesaffre, E., 1986, Logistic discriminant analysis with applications in electrocardiography: PhD thesis, Katholieke Universiteit Leuven, Belgium, 354 p. (unpublished).

Lesaffre, E. and Albert, A., 1989a, Multiple group logistic regression diagnostics: Applied Statistics, 38, 3, 425-440.

Lesaffre, E. and Albert, A., 1989b, Partial separation in logistic discrimination: The Journal of the Royal Statistical Society, Series B (Methodology), 51, 1, 109-116.

Maddala, G.S., 1988, Introduction to econometrics: Macmillan Publishing Company. New York, 472 p.

Marks, S. and Jean, D.O., 1974, Discriminant functions when covariance matrices are unequal: Journal of The American Statistical Association, 346, 555-559.

McCullagh, P. and Nelder, J.A., 1983, Generalized linear models: Chapman and Hall, London, 261 p.

DEĞİNİLEN BELGELER DİZİNİ (Devam ediyor)

Miller, A.J., 1984, Selection of subsets of regression analysis: Journal of Royal Statistical Society, Series A, 147, 389-425.

Özden, R., 1983, Kalp hastalıkları: İstanbul Tıp Fakültesi Vakfı, Baudo basım yayın, 1160 s.

Pregibon, D., 1981, Logistic regression diagnostics: The Annals of Statistics, 9, 705-724.

Prentice, R.L., 1976, A generalization of the probit and logit methods for dose-response curves: Biometrics 32, 761-768.

Press, S.J. and Wilson, S., 1978, Choosing between logistic Regression and Discriminant Analysis: Journal of The American Statistical Association, 73, 384, 699-705.

Qu, Y., Williams, G.W., Beck, G.J. and Goormastic, M., 1987, A generalized linear model of logistic regression for clustered data: Communications in Statistics. Theory and Method, 16(12), 3447-3476.

Silvapulle, M.J., 1981, On the existance of maximum likelihood estimates for the binomial response models: Journal of Royal Statistical Society, Series B, 43, 310-313.

Smith, K.C., Savin, N.E. and Robertson, J.L., 1984, A Monte Carlo comparision of maximum likelihood and minimum chi-square sampling distributions in logit analysis: Biometrics, 40, 471-482.

Sokolow, M., MD. and Mcilroy, M.B. MD., 1981, Clinical cardiology: Lange Medical Publications, California, 763 p.

Stopher, P.R. and Meyburg, A.H., 1979, Survey sampling and multivariate analysis for social scientists and engineers, Canada, 385 p.

Stukel, T.A., 1988, Generalized Logistic Models: Journal of The American Statistical Association, 83, 402, 426-431.

Taylor, J.M.G., 1988, The cost of generalizing Logistic Regression: Journal of The American Statistical Association, 83, 404, 1078-1083.

Tritchler, D., 1984, An algoritm for exact logistic regression: Journal of The American Statistical Association, 79, 387, 709-711.

Tsiatis, A.A., 1980, A note on a goodness of fit test for the logistic regression model: Biometrika, 67, 250-251.

Walker, S.H. and Duncan, D.B., 1967, Estimation of the probability of an event as a function of several independent variables: Biometrika, 54, 167-179.

Weisberg, S., 1985, Applied linear regression: John Wiley, New York, 323 p.

EKLER

EK-I. 16 değişkenli veri kümesine uydurulan çoklu grup lojistik modelde her bir gruba ilişkin skor değerleri, gerçek gruplar ve gruplara atanma olasılıkları.

SKOR MATRİSİ

GÖZLEM NO	GRUP	SKORLAR 1	2	3
1	1	2.47	-34.37	-50.98
2	1	.71	-41.25	-34.07
3	1	12.84	-53.64	-40.68
4	1	6.14	-35.89	-18.10
5	1	5.81	-64.86	-55.03
6	1	.86	-35.75	-31.06
7	1	2.35	-34.36	-27.71
8	1	5.53	-45.19	-24.03
9	1	.92	-55.59	-24.23
10	1	1.34	-21.11	-29.17
11	1	6.00	-44.14	-46.78
12	1	3.71	-38.83	-27.05
13	1	1.04	-49.96	-66.05
14	1	3.49	-31.56	-25.72
15	1	4.88	-37.77	-28.98
16	1	1.11	-40.75	-32.51
17	1	.97	-44.96	-34.06
18	1	5.90	-37.65	-5.95
19	1	5.07	-46.57	-24.71
20	1	3.49	-21.96	-22.78
21	1	-.60	-11.69	-12.81
22	1	2.15	-16.84	-2.04
23	1	-3.56	-29.31	-33.33
24	1	4.30	-46.96	-34.87
25	1	8.88	-52.72	-43.95
26	1	6.24	-12.92	-5.85
27	1	1.18	-41.88	-10.47
28	1	3.74	-17.10	-15.30
29	1	1.60	-52.65	-41.78
30	1	4.60	-43.49	-25.04
31	1	8.67	-53.36	-44.26
32	1	2.44	-52.38	-31.57
33	1	3.91	-9.41	-35.38
34	1	5.53	-37.70	-18.99
35	1	2.29	3.85	5.50
36	1	3.48	-44.26	-62.86
37	1	.27	-8.69	-14.12
38	1	4.89	-47.10	-24.98
39	1	4.58	-8.08	-7.24
40	1	6.00	-52.06	-56.57
41	1	3.27	-25.97	-21.03
42	1	.88	-19.03	-28.56
43	1	2.93	-50.90	-30.84
44	1	1.47	-19.50	-15.14
45	1	4.02	-29.97	-24.93
46	1	2.81	-24.94	-24.91
47	1	3.81	-48.45	-35.60
48	1	1.51	-1.26	-11.21
49	1	6.07	-43.93	-46.67
50	1	8.75	-42.90	-44.87
51	1	3.68	-20.19	-26.49

52	1	.92	-38.11	-16.92
53	1	4.04	-37.02	-29.87
54	1	1.78	-42.19	-32.81
55	2	-1.05	12.95	9.84
56	2	-11.36	12.63	-5.82
57	2	-.80	13.69	10.21
58	2	1.34	28.60	21.61
59	2	-2.39	14.32	9.61
60	2	-3.99	12.06	11.18
61	2	-6.57	6.56	4.51
62	2	-5.16	5.39	4.84
63	2	-6.61	6.45	4.46
64	2	-2.71	9.58	7.80
65	2	-1.79	21.27	14.19
66	2	-4.14	16.64	7.80
67	2	-.69	27.73	20.42
68	2	-4.97	32.10	10.51
69	2	-3.38	8.19	3.54
70	2	2.78	10.48	9.92
71	2	-.80	13.69	10.21
72	2	1.78	19.19	16.88
73	2	-6.57	6.56	4.51
74	2	-8.67	36.57	13.55
75	2	-5.09	5.60	4.95
76	2	-6.41	12.03	8.32
77	2	-6.24	12.56	8.58
78	2	-5.16	5.39	4.84
79	2	-13.61	.94	-1.58
80	2	-3.99	12.06	11.18
81	2	-9.16	1.05	-2.16
82	2	-4.03	17.10	14.81
83	2	-.34	13.54	4.48
84	2	-3.96	12.16	11.23
85	2	-9.98	40.95	15.79
86	2	-3.69	17.95	15.20
87	2	-9.19	.95	-2.21
88	2	-11.14	1.83	-14.45
89	2	.90	21.74	-4.62
90	2	2.96	5.66	-15.46
91	2	3.34	32.17	27.47
92	2	-7.73	21.74	3.14
93	2	-3.93	3.93	7.38
94	2	-4.25	8.21	5.87
95	2	-5.24	39.83	16.46
96	2	-.80	13.69	10.21
97	2	-1.05	12.95	9.84
98	2	.99	48.10	20.37
99	2	-9.50	2.37	5.54
100	2	-.98	26.88	20.00
101	2	6.57	6.56	4.51
102	2	5.58	7.71	-8.38
103	2	-.77	13.80	10.26
104	2	.16	19.72	16.23
105	2	4.40	24.80	23.60
106	2	1.00	27.41	21.17
107	2	-2.29	14.64	9.77

108	2	-2.29	14.64	9.77
109	2	-2.29	14.64	9.77
110	2	.58	26.14	20.54
111	2	-5.09	5.60	4.95
112	2	-10.02	25.04	6.52
113	2	-2.18	19.48	16.88
114	2	-7.57	9.06	4.46
115	2	.79	26.77	20.86
116	2	-2.46	24.05	18.22
117	2	-2.26	19.69	13.37
118	2	-.77	13.80	10.26
119	2	-10.67	20.10	.89
120	2	.79	26.77	20.86
121	2	-.39	33.62	24.44
122	2	.79	26.77	20.86
123	2	-2.25	14.75	9.82
124	2	.12	29.55	24.89
125	2	-2.01	34.15	23.79
126	3	-41	5.69	10.50
127	3	-5.09	5.60	4.95
128	3	-3.44	.10	1.63
129	3	-5.71	1.13	8.68
130	3	.68	-6.75	3.17
131	3	-8.34	2.87	2.31
132	3	-.38	-9.93	1.59
133	3	-9.70	-3.20	1.68
134	3	.61	-6.96	3.06
135	3	.79	7.29	13.37
136	3	1.39	-5.25	7.48
137	3	-5.11	10.78	10.96
138	3	-.46	.20	6.80
139	3	-2.25	-4.74	2.34
140	3	4.89	11.94	20.53
141	3	3.30	-1.14	9.89
142	3	-3.75	-2.63	4.11
143	3	-3.96	3.82	7.33
144	3	.65	-6.86	3.11
145	3	-3.93	3.93	7.38
146	3	-8.22	-4.16	2.12
147	3	-3.96	3.82	7.33
148	3	-8.22	-4.16	2.12
149	3	-8.25	-4.26	2.07
150	3	-8.22	-4.16	2.12
151	3	-8.60	3.25	7.91
152	3	3.26	-1.25	9.83
153	3	-6.81	11.91	13.45
154	3	-5.09	5.60	4.95
155	3	2.17	6.01	13.65
156	3	-13.27	-1.74	2.25
157	3	-2.89	-4.67	.47
158	3	-5.23	5.18	4.74
159	3	3.74	2.80	5.89
160	3	-8.22	-4.16	2.12
161	3	-5.09	5.60	4.95

162	3	-6.61	6.45	4.46
163	3	-8.22	-4.16	2.12
164	3	-9.70	-3.20	1.68
165	3	-6.66	8.82	12.71
166	3	-.80	13.69	10.21
167	3	-8.11	9.88	12.32
168	3	.51	2.65	11.61
169	3	-7.12	2.30	8.35
170	3	-6.71	22.23	22.57
171	3	-3.89	4.04	7.43
172	3	-3.93	3.93	7.38
173	3	-1.39	-8.98	3.85
174	3	-.80	-5.80	2.73
175	3	-5.85	.71	8.47
176	3	.72	-6.65	3.22
177	3	-5.41	4.88	6.94
178	3	4.40	5.31	16.12
179	3	.72	-6.65	3.22
180	3	2.10	5.80	13.55
181	3	-6.63	8.93	12.76
182	3	-2.94	20.89	25.66
183	3	-8.22	-4.16	2.12
184	3	-6.63	8.93	12.76
185	3	.72	-6.65	3.22
186	3	4.85	11.84	20.48
187	3	-6.51	-1.57	.71
188	3	-8.18	-4.05	2.17
189	3	.68	-6.75	3.17
190	4	-1.43	-8.21	-7.45
191	4	-4.59	-10.08	-35.23
192	4	-7.43	-14.59	-8.79
193	4	-4.53	-17.05	-6.98
194	4	-3.12	-11.68	-9.54
195	4	-8.27	-5.25	-1.49
196	4	2.24	-38.05	-55.98
197	4	-.86	-29.77	-35.70
198	4	-6.25	-21.44	-11.37
199	4	-1.46	-8.32	-7.50
200	4	-11.62	-1.77	-11.19
201	4	-5.59	-3.38	-.83
202	4	-7.34	-17.40	-5.70
203	4	-6.25	-21.44	-12.37
204	4	-4.30	-4.83	-5.96
205	4	-4.60	-2.78	-4.63
206	4	-.33	-40.80	-31.28
207	4	-5.40	-20.51	-11.54
208	4	-6.25	-21.44	-12.37
209	4	-7.02	-6.71	-1.44
210	4	-8.34	-15.68	-10.20
211	4	-7.24	-24.41	-13.84
212	4	-7.15	-10.98	-5.28
213	4	2.76	-32.46	-1.86
214	4	-1.90	-12.25	-3.50
215	4	-7.40	-19.56	-10.45

216	4	-4.27	-49.32	-45.20
217	4	-6.25	-21.44	-12.37
218	4	.36	-35.12	-30.41
219	4	-9.84	-3.63	1.47
220	4	-7.41	-38.37	-43.81
221	4	-9.70	-7.67	-15.13
222	4	-.14	-53.14	-44.36
223	4	-7.11	-25.84	-31.45
224	4	-.95	-2.33	-11.94
225	4	2.62	-23.42	-21.61
226	4	-7.41	-11.22	-4.05
227	4	-9.44	-3.58	-3.92
228	4	-3.97	-15.98	-18.59
229	4	-5.63	-10.80	-6.54
230	4	-5.98	-11.86	-7.06
231	4	-5.73	-22.04	-28.37
232	4	-1.67	-24.36	-47.22
233	4	-8,27	-5.25	-1.49
234	4	-5.97	-16.80	-10.61
235	4	-3.32	-5.95	-.16
236	4	-1.31	-37.67	-29.21
237	4	-2.22	-39.18	-41.34
238	4	-7.92	-16.08	-9.53
239	4	-3.09	-11.57	-9.49
240	4	-5.05	-9.52	-2.30
241	4	-9.68	-10.54	-13.18
242	4	-6.25	-21.44	-12.37
243	4	-5.26	-20.09	-11.33
244	4	-5.46	-38.53	-21.47
245	4	-1.94	-12.36	-3.56
246	4	-4.18	-40.19	-23.87
247	4	-7.43	-14.59	-8.79
248	4	-8.18	-4.05	2.17
249	4	-7.89	-6.44	-6.32
250	4	-1.54	-2.60	-16.88

ATAMA VE OLASILIKLAR MATRİSİ

GÖZLEM NO	GRUP GERÇEK	GRUP ATANAN	ATANMA OLASILIKLARI 1	2	3	4
1	1	1	.92	.00	.00	.08
2	1	1	.67	.00	.00	.33
3	1	1	1.00	.00	.00	.00
4	1	1	1.00	.00	.00	.00
5	1	1	1.00	.00	.00	.00
6	1	1	.70	.00	.00	.30
7	1	1	.91	.00	.00	.09
8	1	1	1.00	.00	.00	.00
9	1	1	.71	.00	.00	.29
10	1	1	.79	.00	.00	.21
11	1	1	1.00	.00	.00	.00
12	1	1	.98	.00	.00	.02
13	1	1	.74	.00	.00	.26
14	1	1	.97	.00	.00	.03
15	1	1	.99	.00	.00	.01
16	1	1	.75	.00	.00	.25
17	1	1	.73	.00	.00	.27
18	1	1	1.00	.00	.00	.00
19	1	1	.99	.00	.00	.01
20	1	1	.97	.00	.00	.03
21	1	4	.36	.00	.00	.64
22	1	1	.88	.00	.01	.10
23	1	4	.03	.00	.00	.97
24	1	1	.99	.00	.00	.01
25	1	1	1.00	.00	.00	.00
26	1	1	1.00	.00	.00	.00
27	1	1	.77	.00	.00	.23
28	1	1	.98	.00	.00	.02
29	1	1	.83	.00	.00	.17
30	1	1	.99	.00	.00	.01
31	1	1	1.00	.00	.00	.00
32	1	1	.92	.00	.00	.08
33	1	1	.98	.00	.00	.02
34	1	1	1.00	.00	.00	.00
35	1	3	.03	.16	.81	.00
36	1	1	.97	.00	.00	.03
37	1	1	.57	.00	.00	.43
38	1	1	.99	.00	.00	.01
39	1	1	.99	.00	.00	.01
40	1	1	1.00	.00	.00	.00
41	1	1	.96	.00	.00	.04
42	1	1	.71	.00	.00	.29
43	1	1	.95	.00	.00	.05
44	1	1	.81	.00	.00	.19
45	1	1	.98	.00	.00	.02
46	1	1	.94	.00	.00	.06
47	1	1	.98	.00	.00	.02
48	1	1	.78	.05	.00	.17
49	1	1	1.00	.00	.00	.00
50	1	1	1.00	.00	.00	.00

51	1	1	.98	.00	.00	.02
52	1	1	.72	.00	.00	.28
53	1	1	.98	.00	.00	.02
54	1	1	.86	.00	.00	.14
55	2	2	.00	.96	.04	.00
56	2	2	.00	1.00	.00	.00
57	2	2	.00	.97	.03	.00
58	2	2	.00	1.00	.00	.00
59	2	2	.00	.99	.01	.00
60	2	2	.00	.71	.29	.00
61	2	2	.00	.88	.11	.00
62	2	2	.00	.63	.37	.00
63	2	2	.00	.88	.12	.00
64	2	2	.00	.86	.14	.00
65	2	2	.00	1.00	.00	.00
66	2	2	.00	1.00	.00	.00
67	2	2	.00	1.00	.00	.00
68	2	2	.00	1.00	.00	.00
69	2	2	.00	.99	.01	.00
70	2	2	.00	.64	.36	.00
71	2	2	.00	.97	.03	.00
72	2	2	.00	.91	.09	.00
73	2	2	.00	.88	.11	.00
74	2	2	.00	1.00	.00	.00
75	2	2	.00	.66	.34	.00
76	2	2	.00	.98	.02	.00
77	2	2	.00	.98	.02	.00
78	2	2	.00	.63	.37	.00
79	2	2	.00	.68	.05	.26
80	2	2	.00	.71	.29	.00
81	2	2	.00	.72	.03	.25
82	2	2	.00	.91	.09	.00
83	2	2	.00	1.00	.00	.00
84	2	2	.00	.72	.28	.00
85	2	2	.00	1.00	.00	.00
86	2	2	.00	.94	.06	.00
87	2	2	.00	.70	.03	.27
88	2	2	.00	.86	.00	.14
89	2	2	.00	1.00	.00	.00
90	2	2	.06	.93	.00	.00
91	2	2	.00	.99	.01	.00
92	2	2	.00	1.00	.00	.00
93	2	3	.00	.03	.97	.00
94	2	2	.00	.91	.09	.00
95	2	2	.00	1.00	.00	.00
96	2	2	.00	.97	.03	.00
97	2	2	.00	.96	.04	.00
98	2	2	.00	1.00	.00	.00
99	2	3	.00	.04	.96	.00
100	2	2	.00	1.00	.00	.00
101	2	2	.00	.88	.11	.00
102	2	2	.11	.89	.00	.00
103	2	2	.00	.97	.03	.00
104	2	2	.00	.97	.03	.00
105	2	2	.00	.77	.23	.00

106	2	2	.00	1.00	.00	.00
107	2	2	.00	.99	.01	.00
108	2	2	.00	.99	.01	.00
109	2	2	.00	.99	.01	.00
110	2	2	.00	1.00	.00	.00
111	2	2	.00	.66	.34	.00
112	2	2	.00	1.00	.00	.00
113	2	2	.00	.93	.07	.00
114	2	2	.00	.99	.01	.00
115	2	2	.00	1.00	.00	.00
116	2	2	.00	1.00	.00	.00
117	2	2	.00	1.00	.00	.00
118	2	2	.00	.97	.03	.00
119	2	2	.00	1.00	.00	.00
120	2	2	.00	1.00	.00	.00
121	2	2	.00	1.00	.00	.00
122	2	2	.00	1.00	.00	.00
123	2	2	.00	.99	.01	.00
124	2	2	.00	.99	.01	.00
125	2	2	.00	1.00	.00	.00
126	3	3	.00	.01	.99	.00
127	3	2	.00	.66	.34	.00
128	3	3	.00	.15	.70	.14
129	3	3	.00	.00	1.00	.00
130	3	3	.07	.00	.89	.04
131	3	2	.00	.62	.35	.03
132	3	3	.10	.00	.74	.15
133	3	3	.00	.01	.84	.16
134	3	3	.08	.00	.88	.04
135	3	3	.00	.00	1.00	.00
136	3	3	.00	.00	1.00	.00
137	3	3	.00	.45	.55	.00
138	3	3	.00	.00	1.00	.00
139	3	3	.01	.00	.90	.09
140	3	3	.00	.00	1.00	.00
141	3	3	.00	.00	1.00	.00
142	3	3	.00	.00	.98	.02
143	3	3	.00	.03	.97	.00
144	3	3	.08	.00	.89	.04
145	3	3	.00	.03	.97	.00
146	3	3	.00	.00	.89	.00
147	3	3	.00	.03	.97	.00
148	3	3	.00	.00	.89	.11
149	3	3	.00	.00	.89	.11
150	3	3	.00	.00	.89	.11
151	3	3	.00	.01	.99	.00
152	3	3	.00	.00	1.00	.00
153	3	3	.00	.18	.82	.00
154	3	2	.00	.66	.34	.00
155	3	3	.00	.00	1.00	.00
156	3	3	.00	.02	.89	.09
157	3	3	.02	.00	.60	.37
158	3	2	.00	.61	.39	.00
159	3	3	.10	.04	.86	.00
160	3	3	.00	.00	.89	.11
161	3	2	.00	.66	.34	.00

162	3	2	.00	.88	.12	.00
163	3	3	.00	.00	.89	.11
164	3	3	.00	.01	.84	.16
165	3	3	.00	.02	.98	.00
166	3	2	.00	.97	.03	.00
167	3	3	.00	.08	.92	.00
168	3	3	.00	.00	1.00	.00
169	3	3	.00	.00	1.00	.00
170	3	3	.00	.41	.59	.00
171	3	3	.00	.03	.97	.00
172	3	3	.00	.03	.97	.00
173	3	3	.01	.00	.97	.00
174	3	3	.03	.00	.91	.06
175	3	3	.00	.00	1.00	.00
176	3	3	.07	.00	.89	.04
177	3	3	.00	.11	.89	.00
178	3	3	.00	.00	1.00	.00
179	3	3	.07	.00	.89	.04
180	3	3	.00	.00	1.00	.00
181	3	3	.00	.02	.98	.00
182	3	3	.00	.01	.99	.00
183	3	3	.00	.00	.89	.11
184	3	3	.00	.02	.98	.00
185	3	3	.07	.00	.89	.04
186	3	3	.00	.00	1.00	.00
187	3	3	.00	.06	.63	.31
188	3	3	.00	.00	.90	.10
189	3	3	.07	.00	.89	.04
190	4	4	.19	.00	.00	.81
191	4	4	.01	.00	.00	.99
192	4	4	.00	.00	.00	1.00
193	4	4	.01	.00	.00	.99
194	4	4	.04	.00	.00	.96
195	4	4	.00	.00	.18	.81
196	4	1	.90	.00	.00	.10
197	4	4	.30	.00	.00	.70
198	4	4	.00	.00	.00	1.00
199	4	4	.19	.00	.00	.81
200	4	4	.00	.15	.00	.85
201	4	4	.00	.02	.30	.68
202	4	4	.00	.00	.00	1.00
203	4	4	.00	.00	.00	1.00
204	4	4	.01	.01	.00	.98
205	4	4	.01	.06	.01	.92
206	4	4	.42	.00	.00	.58
207	4	4	.00	.00	.00	1.00
208	4	4	.00	.00	.00	1.00
209	4	4	.00	.00	.19	.81
210	4	4	.00	.00	.00	1.00
211	4	4	.00	.00	.00	1.00
212	4	4	.00	.00	.01	.99
213	4	1	.93	.00	.01	.06
214	4	4	.13	.00	.03	.85
215	4	4	.00	.00	.00	1.00
216	4	4	.01	.00	.00	.99

217	4	4	.00	.00	.00	1.00
218	4	1	.59	.00	.00	.41
219	4	3	.00	.00	.81	.19
220	4	4	.00	.00	.00	1.00
221	4	4	.00	.00	.00	1.00
222	4	4	.47	.00	.00	.53
223	4	4	.00	.00	.00	1.00
224	4	4	.26	.07	.00	.67
225	4	1	.93	.00	.00	.07
226	4	4	.00	.00	.02	.98
227	4	4	.00	.03	.02	.95
228	4	4	.02	.00	.00	.98
229	4	4	.00	.00	.00	.99
230	4	4	.00	.00	.00	1.00
231	4	4	.00	.00	.00	1.00
232	4	4	.16	.00	.00	.84
233	4	4	.00	.00	.18	.81
234	4	4	.00	.00	.00	1.00
235	4	4	.02	.00	.45	.53
236	4	4	.21	.00	.00	.79
237	4	4	.10	.00	.00	.90
238	4	4	.00	.00	.00	1.00
239	4	4	.04	.00	.00	.96
240	4	4	.01	.00	.09	.90
241	4	4	.00	.00	.00	1.00
242	4	4	.00	.00	.00	1.00
243	4	4	.01	.00	.00	.99
244	4	4	.00	.00	.00	1.00
245	4	4	.12	.00	.02	.85
246	4	4	.02	.00	.00	.98
247	4	4	.00	.00	.00	1.00
248	4	3	.00	.00	.90	.10
249	4	4	.00	.00	.00	1.00
250	4	4	.17	.06	.00	.78