

**LOJİSTİK REGRESYONDA ÇOKLU AYKIRI GÖZLEMLERİN
BELİRLENMESİ VE ETKİLİLİKLERİNİN İNCELENMESİ**

Mustafa Selçuk YAVUZKANAT

**YÜKSEK LİSANS TEZİ
İSTATİSTİK**

**GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

**ŞUBAT 2011
ANKARA**

Mustafa Selçuk YAVUZKANAT tarafından hazırlanan“LOJİSTİK REGRESYONDA ÇOKLU AYKIRI GÖZLEMLERİN BELİRLENMESİ VE ETKİLİLİKLERİNİN İNCELENMESİ”adlı bu tezin Yüksek Lisans tezi olarak uygun olduğunu onaylarım.

Prof. Dr. M. Akif BAKIR

.....

Tez Danışmanı, İstatistik Anabilim Dalı

Bu çalışma, jürimiz tarafından oy birliği ile İstatistik Anabilim Dalında Yüksel Lisans tezi olarak kabul edilmiştir.

Prof. Dr. Bedriye SARAÇOĞLU

.....

Ekonometri Anabilim Dalı, G.Ü.

Prof. Dr. M. Akif BAKIR

.....

İstatistik Anabilim Dalı, G.Ü.

Yrd. Doç. Dr. Ufuk EKİZ

.....

İstatistik Anabilim Dalı, G.Ü.

Tarih: 03/02/2011

Bu tez ile G. Ü. Fen Bilimleri Enstitüsü Yönetim Kurulu Yüksek Lisans derecesini onamıştır.

Prof. Dr. Bilal TOKLU

.....

Fen Bilimleri Enstitüsü Müdürü

TEZ BİLDİRİMİ

Tez içindeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, ayrıca tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

Mustafa Selçuk YAVUZKANAT

**LOJİSTİK REGRESYONDA ÇOKLU AYKIRI GÖZLEMLERİN
BELİRLENMESİ VE ETKİLİLİKLERİNİN İNCELENMESİ
(Yüksek Lisans Tezi)**

Mustafa Selçuk YAVUZKANAT

**GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

Şubat 2011

ÖZET

Regresyon, bağımlı değişken ile bir veya daha çok bağımsız değişken arasındaki ilişkiyi incelemek amacıyla kullanılan bir analiz yöntemidir. Regresyon analizinde bağımlı değişken genellikle sürekli ve ölçülebilir alınır. Ancak bazı durumlarda bağımlı değişken kesiklidir ve sadece iki uçlu değerler alabilir. Böyle durumlarda bağımlı değişken Binom olasılık dağılımına sahip olacağından regresyon modelini bir olasılık modeli olarak kurmak gerekir. Bu bağlamda regresyon modeli lojistik dağılım fonksiyonuna dayalı olarak yazılabilir. Doğrusal regresyon modellerinde olduğu gibi, lojistik regresyon modellerinde de veri kümesi uzayının uçlarında yer alan noktalar olabilir. Aykırı gözlem olarak bilinen bu gözlemlerin belirlenmesi ve regresyon fonksiyonunun uyum iyiliği ve/veya parametre tahminleri üzerine etkisi olup olmadığının irdelenmesi gerekir. Aykırı gözlemlerin belirlenmesine yönelik bazı tek gözlem yöntemleri geliştirilmiştir. Uygulamada sıkça rastlanan yaklaşım, aykırı gözlemlerin bu türden yöntemlerle incelenmesidir. Ancak maskeleyme(masking) ve sahte aykırı gözlem(swamping) problemi söz konusu olduğunda tek gözlem yöntemleri başarısız olur. Bu nedenle bu problemlerden etkilenmeyecek bir yöntem ihtiyacı duyulmaktadır. Bu çalışmada, lojistik regresyonda çoklu aykırı gözlemlerin belirlenmesinde etkili bir şekilde kullanılabilecek olan Genelleştirilmiş Standart Pearson Artıkları(GSPR) yöntemi Türkiye 2009 Elit Bayan Voleybol liginde ve firma iflaslarına ilişkin

verilere uygulanarak yöntemin etkin biçimde çalıştığı ve maskeleme etkisini giderdiği gösterilmiştir.

Bilim Kodu : 205.1.066

Anahtar Kelimeler : Lojistik regresyon, aykırı gözlemler, maskeleme, sahte aykırı gözlem, etkili gözlem, GSPR

Sayfa Adedi : 77

Tez Yöneticisi : Prof. Dr. M. Akif BAKIR

**DETECTION OF MULTIPLE OUTLIERS IN LOGISTIC REGRESSION
AND
EXAMINATION OF EFFECTIVENESS
(M.Sc. Thesis)**

Mustafa Selçuk YAVUZKANAT

**GAZI UNIVERSITY
INSTITUTE OF SCIENCE AND TECHNOLOGY
February 2011**

ABSTRACT

Regression is a method of analysis performed to examine the relationship between a dependent variable and one or more independent variables. In regression analysis, dependent variable is generally held continuous and measurable. But in some cases, dependent variable is binary and it takes only discrete values. Since dependent variable has binominal probability distribution in such cases, regression model should be set up as a probability model. In that sense, regression model can be designed in accordance with logistic distribution function. Like the linear regression models, there may be points at the end of the data set space also in logistic regressions. These observations which are known as outlier observations need to be determined and it should be elaborated whether they have any impact on goodness of fit of the regression functions and/or parameter estimates. Some single observation methods have been developed to identify outliers. The most frequent approach in practice is using such kind of methods to examine the outliers. However, in case of masking and swamping, single observation methods fail. Therefore, a method which will not be affected by these problems is required. In this study, the method of Generalised Standard Pearson Residuals (GSPR) which can be used effectively for identification of multiple outliers in logistic regression has been applied to the data concerning Turkey 2009 Elite Women Volleyball League and firm

bankrupties and it has been proved that the method works effectively and eliminates the effect of masking.

Science Code : 205.1.066
Key Words : Logistic regression, outliers, masking, swamping,
influential observations, GSPR
Page Number : 77
Adviser : Prof. Dr. M. Akif BAKIR

TEŞEKKÜR

Çalışmam boyunca değerli yardım ve katkılarıyla beni yönlendiren Prof. Dr. M. Akif Bakır'a, destekleriyle beni hiçbir zaman yalnız bırakmayan çok değerli eşim Pınar Yavuzkanat'a, her zaman yanımda hissettiğim aileme ve bu çalışmaya katkı sağlayan herkese teşekkürü bir borç bilirim.

İÇİNDEKİLER

	Sayfa
ÖZET	iv
ABSTRACT	vi
TEŞEKKÜR.....	viii
İÇİNDEKİLER	ix
ÇİZELGELERİN LİSTESİ.....	xii
ŞEKİLLERİN LİSTESİ	xiii
SİMGELER VE KISALTMALAR.....	xiv
1. GİRİŞ.....	1
2. DOĞRUSAL REGRESYON ANALİZİ.....	6
2.1.Regresyon Denklemi Tahmin Yöntemleri	7
2.1.1.En küçük kareler yöntemi (EKK).....	7
2.1.2.En çok olabilirlik yöntemi (EOB)	8
3. DOĞRUSAL REGRESYON ANALİZİNDE AYKIRI GÖZLEMLER.....	11
3.1.Y (Bağımlı Değişken) Değerine Göre Aykırı Gözlemler	11
3.2.X (Bağımsız Değişken) Değerine Göre Aykırı Gözlemler.....	12
3.2.1.X gözlem değerlerine göre aykırı gözlemlerin hat matrisi ile belirlenmesi.....	13
3.2.2.X gözlem değerlerine göre aykırı gözlemlerin artıklara dayalı olarak belirlenmesi	15
3.3.Artık Çeşitleri	15
3.3.1.Sıradan artıklar	15
3.3.2.Standartlaştırılmış artıklar	16
3.3.3.Student artıklar	16

Sayfa

3.4.Y (Bağımlı Değişken) Değerine Göre Aykırı Gözlemlerin Belirlenmesi.....	17
3.5.Doğrusal Regresyon Analizinde Etkili Gözlemler	18
3.5.1.Difference between the fitted values (DFFITS).....	18
3.5.2.Difference in betas (DFBETAS).....	20
3.5.3.Cook uzaklığı (D_i)	21
3.5.4.Hadi'nin etkililik ölçütü (H_i)	21
3.6.Maskeleme ve Sahte Aykırı Gözlem Problemleri	22
3.7.Çoklu Aykırı Gözlemlerin Belirlenmesi	23
4. LOJİSTİK REGRESYON.....	26
4.1.İki Uçlu Verileri Modelleme	26
4.2.Logit Model.....	28
4.3.Lojistik Regresyonda Aykırı Gözlemler	32
4.3.1.Lojistik regresyonda artıklar ve kaldıraç değerleri	33
4.3.2.Lojistik regresyonda aykırı gözlem belirleme yöntemleri	35
4.3.3.Pearson artıkları (Pearson residuals).....	35
4.3.4.Standart pearson artıkları	36
4.3.5.Sapma artıkları	37
4.3.6.Delta ki-kare $(\Delta \chi^2)$	38
5. LOJİSTİK REGRESYONDA ÇOKLU AYKIRI GÖZLEMLERİN BELİRLENMESİ	39
5.1.Lojistik Regresyonda Çoklu Aykırı Gözlemlerin Belirlenmesi İçin Genelleştirilmiş Standart Pearson Artıkları (GSPR).....	40
6. GSPR YÖNTEMİNİN VOLEYBOL VERİLERİNE UYGULANMASI.....	46
6.1.Etkinlik Modelinin Değişkenleri	46

	Sayfa
6.2.Etkinlik Modelinde Çoklu Aykırı Gözlemlerin Belirlenmesi	47
7. GSPR YÖNTEMİNİN İFLAS VERİLERİNE UYGULANMASI.....	63
7.1.İflas Verilerine İlişkin Değişkenler	63
7.2.İflas Verilerinde Çoklu Aykırı Gözlemlerin Belirlenmesi	63
8. SONUÇ	72
KAYNAKLAR	74
ÖZGEÇMİŞ	77

ÇİZELGELERİN LİSTESİ

Çizelge	Sayfa
Çizelge 6.1. Etkinlik Modeli Verileri.....	48
Çizelge 6.2. Etkinlik Modeli olasılık, Pearson ve Standart Pearson artık değerleri.....	50
Çizelge 6.2. (Devam) Etkinlik Modeli olasılık, Pearson ve Standart Pearson artıkdeğerleri	51
Çizelge 6.3. Aykırı Gözlemlerin Pearson ve Standart Pearson Artık Değerleri	52
Çizelge 6.4. Şüpheli görülen gözlemlerin pearson ve standart pearson artık değerleri.....	54
Çizelge 6.5. D₀ gözlem grubuna ait Pearson, Standart Pearson ve GSPR değerleri.....	55
Çizelge 6.6. D₁ gözlem grubuna ait Pearson, Standart Pearson ve GSPR değerleri.....	57
Çizelge 6.7. Veri kümesinden aykırı gözlem kombinasyonları çıkarıldıktan sonra modellere ilişkin elde edilen değerler.....	59
Çizelge 6.7. (Devam) Veri kümesinden aykırı gözlem kombinasyonları çıkarıldıktan sonra modellere ilişkin elde edilen değerler	60
Çizelge 6.8 İlk ve son modellerden elde edilen değerler	62
Çizelge 7.1. Batan ve hayatını devam ettiren şirketlere ilişkin finansal oranlar.....	64
Çizelge 7.2. Etkinlik Modeli olasılık, Pearson ve standart Pearson artık değerleri ...	66
Çizelge 7.3. Aykırı Gözlemlerin Pearson ve Standart Pearson Artık Değerleri	67
Çizelge 7.4. Şüpheli görülen gözlemlerin pearson ve standart pearson artık değerleri.....	68
Çizelge 7.5. D₀ gözlem grubuna ait Pearson, Standart Pearson ve GSPR değerleri	69
Çizelge 7.6. D₁ gözlem grubuna ait Pearson, Standart Pearson ve GSPR değerleri .	70

ŞEKİLLERİN LİSTESİ

Şekil	Sayfa
Şekil 3.1. Regresyon Analizinde Aykırı Gözlemler.....	12
Şekil 3.2. Düşük h_{ii} Yüksek d_i^*	19
Şekil 3.3. Düşük h_{ii} Yüksek d_i^*	19
Şekil 3.4. Düşük h_{ii} Yüksek d_i^*	19
Şekil 3.5. Maskeleye Problemi	22
Şekil 3.6. Sahte Aykırı Gözlem Problemi.....	23
Şekil 4.1. İki uçlu değer alan bağımlı değişkene ilişkin dağılım grafiği	28
Şekil 4.2. Grupların olasılıklara göre grafiği	29
Şekil6.1. Etkinlik Modeli Verileri ile Tahmin Edilen Logit Model.....	49
Şekil 6.2. Pearson artıklarının gözlem numaralarına göre grafiği	52
Şekil 6.3. Standart Pearson artıklarının gözlem numaralarına göre grafiği	53
Şekil 6.4. R_0 gözlem grubu ile elde edilen logit model tahmini	55
Şekil 6.5. Kalan gözlem grubu R_1 ile elde edilen logit model tahmini.....	56
Şekil 6.6. GSPR(2) değerlerinin gözlem numaralarına göre grafiği.....	57
Şekil 6.7. Elde edilen tüm modellerin LR Chi2 Değerleri.....	61
Şekil 7.1. İflas Verileri ile Tahmin Edilen Logit Model	65
Şekil 7.2. Pearson artıklarının gözlem numaralarına göre grafiği	67
Şekil 7.3. R_0 gözlem grubu ile elde edilen logit model tahmini	69
Şekil 7.4. Kalan gözlem grubu R_1 ile elde edilen logit model tahmini.....	70
Şekil 7.5. GSPR(2) değerlerinin gözlem numaralarına göre grafiği.....	71

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış bazı simgeler ve kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Simgeler	Açıklama
D_i	Cook Uzaklığı
H_i	Hadi'nin Etkililik Ölçütü
$\sum$	Varyans-Kovaryans Matrisi
h_{ii}	i 'inci kaldıraç değeri
$\hat{\pi}_i$	Modelden tahmin edilen olasılık değeri
r_{pi}	i 'inci pearson artığı
r_{sp_i}	i 'inci standart pearson artığı
$(\Delta\chi^2)$	Delta Ki-Kare istatistiği
$\pi_i^{(-D)}$	D gözlem grubu çıkarıldıktan sonra elde edilen olasılık değeri
$\hat{\beta}^{(-D)}$	D gözlem grubu çıkarıldıktan sonra elde edilen parametre vektörü
$r_{si}^{(-D)}$	i 'inci Genelleştirilmiş Standart Pearson Artığı

Kısaltmalar	Açıklama
GSPR	Genelleştirilmiş Standart Pearson Artığı
EKK	En Küçük Kareler
EOB	En Çok Olabilirlik
SSE	Artık Kare Toplamı
MSE	Hata Kare Ortalaması

1. GİRİŞ

Regresyon bağımlı bir değişkeni, tahmin veya çıkarım amacıyla farklı bağımsız değişkenler ile ilişkilendiren, bağımlı değişken değerlerini tahmin etme veya kestirme amacıyla inceleyen bir istatistik yöntemidir. Bu terim, ilk kez İngiliz antropolog, meteorolojist, kaşif, mucit ve istatistikçi Sir Francis Galton (1822 - 1911) tarafından kullanılmıştır.

Regresyonda amaçlardan biri, bağımlı değişkenle bağımsız değişken arasındaki bağıntının ortaya çıkarılmasıdır. Bir diğer amaç ise modelin bilinmeyen parametreleri tahmin edildiğinde, bağımsız değişken(ler)in alacağı değerler için bağımlı değişkenin alacağı değeri tahmin etmektir.

İstatistiksel çalışmalarda, rassal olarak çekilen örnekleme oluşturan gözlemlerin, birbirinden bağımsız ve özdeş dağıldıkları varsayılır. Ancak, özellikle gerçek verilerle yapılan çalışmalarda, bazı gözlemlerin diğerlerine göre aşırı büyük ya da küçük olduğu durumlarla karşılaşılabilir. Bu gibi gözlemlerin varlığında regresyon çalışması yapılırken her bir gözlemin ayrıntılı olarak incelenmesi ve aykırı gözlemlerin belirlenmesi gerekir. Çünkü tek bir gözlem bile, regresyon modeli ya da katsayı tahminleri üzerinde büyük bir etkiye sahip olabilir. Bu nedenle veri kümesinde bulunabilecek bu tip gözlemlerin varlığının araştırılması gerekir.

Doğrusal regresyon modelleri için tekli aykırı gözlemlerin belirlenmesine yönelik Hat Matrisi, DFFITS, DFBETAS, Cook Uzaklığı (DFB), Hadi'nin Etkililik Ölçütü (H) gibi birçok yöntem önerilmiştir.

Literatürde bulunan klasik aykırı gözlem belirleme yöntemleri veri kümesinde tek bir aykırı gözlem bulunduğunda oldukça güçlü olan yöntemlerdir. Bu yöntemler aykırı gözlemleri tek tek ele alır ve model üzerindeki etkilerini inceler. Yani veri kümesinde birden fazla aykırı gözlem bulunduğunda etkileşim etkileri değerlendirilmemiş olur. Dolayısı ile bu yöntemlerin yanlış ya da eksik bilgi vermesi olasıdır.

Tekli aykırı gözlemler için geliştirilen yöntemler çoklu aykırı gözlem durumunda uygulandığında, hem aşırı işlem yükü getirmekte hem de maskeleye ve sahte aykırı gözlemler olarak tanımlanan problemlere karşı çalışabilirliğini kaybetmektedir. Maskeleye problemi, var olan aykırı gözlemlerin belirlenememesi;sahte aykırı gözlem ise aykırı olmayan gözlemlerin aykırı gözlem olarak belirlenmesi olarak tanımlanabilir.

Gerçek veri kümelerinde tek bir aykırı gözlem bulunması pek mümkün değildir. Hempel (1986) yaptığı çalışmada, normal veri setlerinde % 1-10 oranında aykırı gözlem bulunduğunu iddia etmiştir.Bu nedenle doğrusal regresyonda aykırı gözlem belirlenmesinde karşılaşılabilecek maskeleye ve sahte aykırı gözlem problemlerinin varlığında etkin olarak çalışabilecek çoklu aykırı gözlem belirleme yöntemlerini üretmek araştırmacıların yoğun ilgisini çekmiştir.

Gentleman ve Wilk (1975) Algoritması, Gray ve Ling (1984)Kümeleme Analizi Yöntemi, Marasinghe (1985) Algoritması, Atkinson (1986) Algoritması, Kianifard ve Shallow (1989) Algoritması, Rousseeuw ve Zomeren (1990) MVE Algoritması, Hadi (1992) Algoritması, Hadi ve Simonoff (1993) Algoritması, Atkinson (1994) Algoritması, Pena ve Yohai (1995) Algoritması, Sebert, Montgomery, Rollier (1998) Algoritması, Rousseeuw ve Driessen (1999) Fast MCD Algoritması, Billor, Hadi, Velleman (2000) BACON Algoritması, Marchette ve Jeffrey (2003) Veri Resmetme Yöntemi, Billor, Chatterjee, Hadi (2005) Algoritmasıdoğrusal regresyon için aykırı gözlem belirleme yöntemlerinden bazıları olarak verilebilir.

Aykırı gözlemlerin belirlenmesinde, son yıllarda robust (sağlam) yöntemler de oldukça yaygın bir biçimde kullanılmaktadır. Robust regresyon teknikleri, aykırı gözlemlere düşük ağırlıklar vererek ya da olağan dışı gözlemlerinparametre tahminleri üzerinde etkili olmadıklarını göstermek için onları göz ardı ederek, bu gözlemleri modele uyumlu hale getirmektedir. Bu yöntemle anormal gözlemleri son durumdaki ağırlıklarından ya da artık büyüklüklerinden belirlemek mümkündür. Bu yöntemlerin içinde en yaygın olarak kullanılanları; Rousseeuw tarafından 1984 ve 1985 yıllarında önerilen En Küçük Kareler Ortancası ile En Küçük Kırpılmış Kareler

ve Huber tarafından 1973 yılında önerilen M tahmin edicileridir. Robust yöntemler genellikle başarılı olurken olumsuz bir yönü ise hiç aykırı gözlem olmayan veri kümelerinde bazı gözlemlerin aykırı olarak belirlenmesidir.

Bağımlı değişkenin sürekli ve normal dağıldığı doğrusal regresyon modellerinin aksine bağımlı değişkenin kesikli yapıda olduğu, ikili değerler aldığı ve böylece binom olasılık dağılımına sahip olduğu durumlar olabilir. Böyle durumlar için en yaygın kullanılan regresyon modeli, lojistik dağılım fonksiyonunu kullanan lojistik regresyon modelidir. Bağımlı değişken; olma-olmama, başarılı-başarısız vb. değerler alabilir. Bu durumda bağımlı değişken her zaman 0 ya da 1 değerleri ile gösterilebilir. Lojistik regresyon modeli oluşturulurken En Küçük Kareler Yöntemi için gerekli olan varsayımlar sağlanmadığından, bu varsayımlara gerek duymayan En Çok Olabilirlik Yöntemi kullanılır. En çok olabilirlik yöntemi büyük örnek gerektirdiğinden model tahmini, gözlem sayısının çok fazla olmadığı durumlarda aykırı gözlemlerin varlığına oldukça duyarlıdır. Lojistik regresyonda doğrudan 0 ve 1 değerlerinin tahmini yerine bu değerlerin ortaya çıkabilme olasılıkları hesaplanır. Bu nedenle lojistik regresyonda aykırı gözlem, modelden tahmin edilen başarı olasılığı küçük olduğunda gözlemlenen durumu başarılı olan ya da modelden elde edilen başarı olasılığı yüksek olduğu halde gözlemlenen durumu başarısız olan gözlemler olarak tanımlanabilir.

Lojistik regresyon için de tekli aykırı gözlem belirlemeye dayalı bazı yöntemler geliştirilmiştir. Bunlardan en çok bilinenleri Pregibon Hat Matrisi, Pearson ve Standart Pearson Artıkları, Sapma Artıkları ve DFFITS yöntemleridir. Daha sonra bu ölçütler Williams (1987), Lee (1988), Thomas ve Cook (1989, 1990) ve Lee ve Zhao (1997) tarafından yapılan çalışmalarda genelleştirilmiş doğrusal modellere uygulanmıştır. Yine 2006 yılında Garcia, Pichardo ve Pardo tarafından Cressie ve Read'in geliştirdiği güç sapması kullanılarak, lojistik regresyon için bir etkililik ölçütü geliştirilmiştir.

Doğrusal regresyonda olduğu gibi lojistik regresyonda da aykırı gözlem belirleme yöntemleri veri kümesinde tek bir aykırı gözlem bulunduğu oldukça güçlü

yöntemlerdir. Veri kümesinde birden fazla aykırı gözlem bulunduğunda bu yöntemler, maskeleye ya da sahte aykırı gözlem durumunu tespiti yönelik olarak çok sayıdaki kombinasyon göz önüne alındığında dikkate değer hesaplama güçlüğüne sahiptir. Bu nedenle hesaplama etkinliğine sahip olan çoklu aykırı gözlem belirleme yöntemlerine gereksinim duyulmaktadır. Doğrusal regresyon modellerinde çoklu aykırı gözlemlerin belirlenmesine yönelik çok sayıda yöntem önerilmiş olmasına rağmen lojistik regresyon modelleri için sınırlı sayıda yöntem önerisi söz konusu olmuştur.

Bu tez çalışmasının amacı; lojistik regresyonda çoklu aykırı gözlemlerin belirlenmesi aşamasında maskeleye ve sahte aykırı gözlem problemlerinden etkilenmeyen bir yöntemin tanıtılmasıdır. Lojistik regresyonda çoklu aykırı gözlemlerin belirlenmesi için sadece birkaç yöntem üzerinde çalışılmıştır. Bunlardan bir tanesi Bedrick ve Hill tarafından 1990 yılında verilmiştir. Bu yöntem ile çoklu aykırı gözlemlerin belirlenmesi için kümeleme analizi tekniğine dayanan bir test önerilmiştir.

Lojistik regresyon analizinde çoklu aykırı gözlemleri belirlemek için etkili bir yöntem, bu alanda oldukça fazla sayıda çalışma yapmış olan Hadi ve Imon tarafından, 2008 yılında önerilmiştir. Hadi ve Imon şüpheli aykırı gözlem kümesinin veri kümesinden çıkarılması üzerine kurulan, Standart Pearson Artıklarının genelleştirilmiş bir şekli olan Genelleştirilmiş Standart Pearson Artıkları(GSPR) yöntemini geliştirmişlerdir. Bu yöntem ile çoklu aykırı gözlemlerin belirlenmesinde karşılaşılan en büyük problem olan maskeleye ve sahte aykırı gözlem problemlerinin etkileri giderilmektedir.

Çalışmanın uygulama aşamasında ise iki uygulama yapılmıştır. Cengiz Akarçeşme tarafından yazılan “Elit Bayan Voleybolunda Maç Sonucunu Açıklayan Değişkenlerin Lojistik Regresyon Yöntemi ile Belirlenmesi ve Maç Kazanmaya Yönelik Olasılık Modelinin Tahmini” isimli doktora tezinde kullanılan “Etkinlik Modeli” verileri ve Ali S. Hadi ve Samprit Chatterjee tarafından 2006 yılında yayımlanan “Regression Analyses by Example” isimli kitapta verilen iflas verileri kullanılmıştır. Bu veriler ile lojistik regresyon modeli oluşturulmuş ve tekli aykırı

gözlem belirleme yöntemleri ile aykırı gözlemler belirlenmiştir. Daha sonra verilere Genelleştirilmiş Standart Pearson Artıkları(GSPR) yöntemi uygulanmış ve ilk durumda belirlenemeyen aykırı gözlemler belirlenerek maskeleye etkisinin giderildiği gösterilmiştir. Ayrıca veri kümesinden çıkarılabilecek aykırı gözlemlerin tüm kombinasyonları oluşturulmuş, bu gruplar veri kümesinden çıkarılarak yöntemin en iyi sonucu verip vermediği incelenmiştir.

2. DOĞRUSAL REGRESYON ANALİZİ

Regresyon bağımlı bir değişkeni, tahmin veya çıkarım amacıyla farklı bağımsız değişkenler ile ilişkilendiren; bağımlı değişkenin başka açıklayıcı değişkenlerle olan ilişkisini, birincinin ortalama değerini ikinci(ler)in bilinen ya da değişmeyen değerleri cinsinden tahmin etme veya kestirme amacıyla inceleyen bir istatistik yöntemidir[Gujarati, 1999].

Regresyonda amaçlardan biri, bağımlı değişkenle bağımsız değişken arasındaki ilişkinin ortaya çıkarılmasıdır. Bir diğer amaç ise modelin bilinmeyen parametreleri tahmin edildiğinde, bağımsız değişken(ler)in farklı değerleri için bağımlı değişkenin alacağı değeri kestirmektir.

Regresyonda bir bağımlı değişken ve bir ya da daha fazla sayıda bağımsız değişken vardır. Burada Y ve X değişkenleri arasındaki ilişkinin şekline göre bir model kullanılır.

Doğrusal regresyon modeli aşağıdaki şekilde yazılabilir.

$$Y = X\beta + \varepsilon \quad (2.1)$$

Burada,

$Y : n \times 1$ boyutlu bağımlı değişken vektörü, $X : n \times p$ boyutlu bağımsız değişken matrisi, $\beta : p \times 1$ boyutlu parametre vektörü ve $\varepsilon : n \times 1$ boyutlu hata terimleri vektörü olur. Tahmin edilen parametreler $\hat{\beta} : p \times 1$ vektörü ile gösterilir[Draper and Smith, 1998]. $\hat{\beta}$ 'nın hesaplanmasında farklı yöntemler kullanılmaktadır. Aşağıda, bu yöntemlerden En Küçük Kareler ve En Çok Olabilirlik Yöntemleri verilmektedir.

2.1. Regresyon Denklemi Tahmin Yöntemleri

2.1.1. En küçük kareler yöntemi (EKK)

EKK yönteminde hata terimlerinin karelerinin toplamının en küçük yapılması istenir.

(2.1) modelinde $E(\varepsilon) = 0, V(\varepsilon) = I\sigma^2$ olduğu varsayılır. Hata terimlerinin beklenen değeri 0 olduğundan $E(Y) = X\beta$ olarak yazılabilir. Daha sonra hata kare toplamı aşağıdaki gibi yazılabilir.

$$\varepsilon^T \varepsilon = (Y - X\beta)^T (Y - X\beta)$$

$$\varepsilon^T \varepsilon = Y^T Y - \beta^T X^T Y - Y^T X \beta + \beta^T X^T X \beta$$

$$\varepsilon^T \varepsilon = Y^T Y - 2\beta^T X^T Y + \beta^T X^T X \beta \quad (2.2)$$

(2.2) eşitliğinin en küçük yapmak amacıyla eşitliğin β 'ya göre türevi alınır, sonuç sıfıra eşitlenerek β yerine $\hat{\beta}$ yazılırsa normal denklemler aşağıdaki gibi elde edilir.

$$(X^T X)\hat{\beta} = X^T Y \quad (2.3)$$

(2.3) denkleminde p normal denklembağımsız ise $X^T X$ 'in tersi vardır. Buradan $\hat{\beta}$ tahmin değerleri aşağıdaki gibi bulunur.

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (2.4)$$

[Draper ve Smith, 1998]

2.1.2. En çok olabilirlik yöntemi(EOB)

İstatistikte, tüm yığınlar üzerlerinde tanımlı tesadüfi değişkenlerin olasılık dağılımı ile tanımlanırlar. En Küçük Kareler yöntemi, olasılık dağılımları ile ilgili herhangi bir varsayım içermez. EKK, genel bir parametre tahmin yaklaşımından çok, regresyon doğrularını belirlemek için kullanılabilecek birhesaplama yöntemi olarak görülmelidir. EKK yönteminden daha güçlü kuramsal özellikler gösteren bir başka tahmin yöntemi ise En Çok Olabilirlik Yöntemidir (EOB). EOB yönteminde yığının olasılık dağılımı belirlenir veya bu yönde bir varsayımda bulunulur. Eldeki örnek verilerinin, hangi katsayılarla sahip yığından gelmiş olma olasılığının en yüksek olduğubulunur.

$Y = X\beta + \varepsilon$ modelinde, $\varepsilon \sim N(0, \sigma)$ olduğu varsayılınsın. Bunun sonucunda Y 'nin olasılık yoğunluk fonksiyonu şu şekilde yazılabilir.

$$f = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}} \exp\left[-\frac{1}{2}(X - \mu)^T \Sigma^{-1}(X - \mu)\right]} \quad (2.8)$$

Burada $X \sim N(\mu, \Sigma)$ ve $X : n \times 1$ matristir.

(2.1) ve (2.8) eşitlikleri kullanılarak olabilirlik fonksiyonu

$$L(\beta) \propto \exp\left[-\frac{1}{2}(Y - X\beta)^T \frac{1}{\sigma} I(Y - X\beta)\right] \quad (2.9)$$

şeklinde yazılabilir.

EOB yönteminde $L(\beta)$ fonksiyonunu en büyüklenmesi istenir. İşlemlerin kolaylaşması için $L(\beta)$ fonksiyonunun doğal logaritması alındığında

$$l(\beta) \propto (Y - X\beta)^T I(Y - X\beta) \propto (Y - X\beta)^T (Y - X\beta)$$

şeklinde yazılabilir.

Fonksiyonu en büyükleyen değerlerin bulunması için fonksiyonun β 'ya göre türevi alınır.

$$\begin{aligned} \frac{\partial}{\partial \beta} l(\beta) &\propto \frac{\partial}{\partial \beta} [(Y - X\beta)^T (Y - X\beta)] \\ \frac{\partial}{\partial \beta} l(\beta) &\propto -(Y - X\beta)^T X - (Y - X\beta)^T X \\ \frac{\partial}{\partial \beta} l(\beta) &\propto -2(Y - X\beta)^T X \\ \frac{\partial}{\partial \beta} l(\beta) &\propto -(Y - X\beta)^T X \\ \frac{\partial}{\partial \beta} l(\beta) &\propto -(Y^T - \beta^T X^T)X \\ \frac{\partial}{\partial \beta} l(\beta) &\propto -(Y^T X - \beta^T X^T X) \end{aligned} \tag{2.10}$$

Son olarak elde edilen türev fonksiyonu sıfıra eşitlendiğinde

$$\begin{aligned} \frac{\partial}{\partial \beta} l(\beta) &\propto -(Y^T X - \beta^T X^T X) = 0 \\ \beta^T X^T X &= Y^T X \\ \beta^T &= Y^T X (X^T X)^{-1} \\ \hat{\beta} &= (X^T X)^{-1} X^T Y \end{aligned} \tag{2.11}$$

$\hat{\beta}$ tahmin vektörü elde edilir.

Regresyon analizinde sıklıkla kullanılan En Küçük Kareler yöntemi regresyon varsayımları denilen varsayımlar sağlandığı zaman geçerli olabilmektedir. Bu varsayımların sağlanmadığı durumlarda bu yönteme başvurmak ciddi hatalara sebep olabilir[Hadi ve Catterjee, 2006].

En Küçük Kareler yöntemine ilişkin varsayımlar aşağıda verilmektedir.

1. Regresyon modeli katsayılar da doğrusaldır.
2. X değerleri yinelenen örneklemelerde değişmez.
3. $\epsilon_i \sim N(0, \sigma^2)$
4. Hata terimleri arasında ardışık bağımlılık yoktur.
5. Hata terimi ϵ_i ve açıklayıcı değişken X_i arasında ilişki yoktur.
6. Gözlem sayısı n , tahmin edilecek parametre sayısından büyük olmalıdır.
7. Değişkenler arasında tam çoklu doğrusallık yoktur[Gujarati, 1999].

Regresyon analizinde varsayımların sağlanıp sağlanmadığının araştırılması önemli bir adımdır. Varsayımların sağlanıp sağlanmadığının ya da diğer bir ifadeyle varsayımların bozulup bozulmadığının incelenmesi regresyon tanı yöntemleri ile yapılır. Regresyon tanı yöntemleri diye bilinen bu çalışmaların içerisinde önemli yer tutan adımların birisi de veri kümesi içindeki aykırı/etkili gözlemlerin araştırılmasıdır.

İstatistiksel çalışmalarda, rassal örneği oluşturan gözlemlerin, birbirinden bağımsız ve özdeş dağıldıkları varsayılır. Ancak, özellikle gerçek verilerle yapılan çalışmalarda, bazı gözlemlerin diğerlerine göre aşırı büyük ya da küçük olduğu durumlarla karşılaşılabilir. Bu gibi gözlemlerin varlığında regresyon çalışması yapılırken, her bir gözlemin ayrıntılı olarak incelenmesi ve aykırı gözlemlerin belirlenmesi gerekir. Çünkü tek bir gözlem bile, regresyon modeli ya da katsayı tahminleri üzerinde büyük bir etkiye sahip olabilir [Hadi ve Catterjee, 2006].

3. DOĞRUSAL REGRESYON ANALİZİNDE AYKIRI GÖZLEMLER

Regresyon analizinde en büyük artık değere sahip ya da veri kümesindeki diğer gözlemlere uzak olan gözlem değeri, aykırı gözlem olarak tanımlanır. Aykırı gözlemlerin varlığı model biçiminin doğru oluşturulmamasından, gerekli dönüşümlerin yapılmamış olmasından, ölçme-kayıt hatalarından ya da örneklem rassallığının sağlanmamasından kaynaklanabilir. Bu anlamda aykırı gözlemlerin regresyon katsayıları ya/ya da modelin uyum iyiliği üzerine anlamlı bir etkide bulunup bulunmadığını irdelemek önemlidir. Aykırı gözlem varlığında tahmin değerleri ya da modelin parametreleri etkilenebilir. Ayrıca tek tek etkili olmayan aykırı gözlemlerin etkileşim etkileride modeli etkileyebilir. Bu da model ya da tahmin edicilerin sağlığını etkileyerek yanlış yorumlara sebep olabilir. Bu tip gözlemleri belirlemek, model üstündeki etkileri üzerinde çalışmak ve veri kümesinden çıkarılıp ya da kalmalarına karar vermek gerekir.

Gözlemler, aykırı olarak nitelendirilirken iki açıdan ele alınabilir

- Y (Bağımlı Değişken) Değerine Göre Aykırı Gözlemler
- X (Bağımsız Değişken) Değerine Göre Aykırı Gözlemler

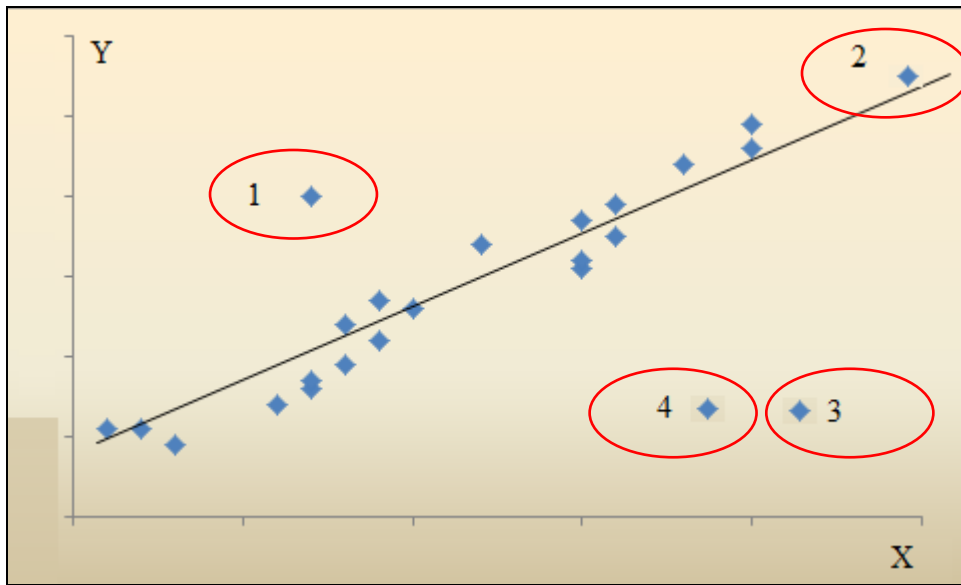
[Neder, Wasserman, Kutner, 1996].

3.1. Y (Bağımlı Değişken) Değerine Göre Aykırı Gözlemler

Büyük artık değerine sahip olan gözlemler, Y değerine göre aykırı gözlem olarak nitelendirilirler. Çünkü bu gözlemler regresyon doğrusundan, Y yönünde uzakta bulunmaktadır [Hadi ve Catterjee, 2006]. Bu tipteki aykırı gözlemler oluşturulan regresyon denklemini etkileyebilir. Regresyon doğrusu olması gereken gibi oluşturulmadığı için doğru, verilerin çoğunluğunu temsil etmeyebilir. Böylece yapılacak tahminler ve yorumlar yanlış olabilir.

3.2. X (Bağımsız Değişken) Değerine Göre Aykırı Gözlemler

Aykırı gözlemlere X değerlerinde de rastlanabilir ve bu değerler regresyon sonuçlarını etkileyebilir. Bir gözlemin X değeri, veri kümesindeki gözlem değerlerinin ortalama X değerinden uzakta ise bu gözlem X değerine göre aykırı gözlem olarak değerlendirilir. Aşağıdaki şekil incelendiğinde bu ayırım daha net görülmektedir.



Şekil 3.1. Regresyon Analizinde Aykırı Gözlemler

Şekilde görülen 1 numaralı gözlemin Y değerine göre aykırı olduğu görülmektedir. Bu noktanın X değeri gözlem aralığının ortasına yakın konumlanmasına rağmen Y değeri regresyon doğrusunun uzağına düşmektedir. 2, 3 ve 4 numaralı gözlemlerin, X değerleri diğer gözlemlerine göre daha büyük değerler aldıkları için aykırı gözlem oldukları söylenebilir. Ayrıca 3. ve 4. noktalar aynı zamanda Y değerlerine göre de aykırı olarak görünmektedir[Neder, Wasserman, Kutner, 1996].

Regresyon analizinde, aykırı gözlemlerin belirlenmesi temel bir adımdır. Bir ya da iki bağımsız değişken varlığında gözlemlerin X ve Y değerlerine göre aykırı gözlem olup olmadığı, Kutu Grafiği, Saçılım Grafiği, Artıkların Grafiği gibi grafiksel yöntemlerle değerlendirilebilir. Modelde ikiden fazla bağımsız değişken bulunduğu aykırı gözlemleri grafiksel yöntemlerle belirlemek oldukça zorlaşmaktadır. Grafiksel yöntemin dışında aykırı gözlemlerin belirlenmesinde çeşitli yöntemler geliştirilmiştir. Bunlardan en çok bilinenleri aşağıda verilmektedir[Neder, Wasserman, Kutner, 1996].

3.2.1. X gözlem değerlerine göre aykırı gözlemlerin hat matrisi ile belirlenmesi

Hat matrisi(H), kestirim değerleri ile gözlem değerleri arasında ilişki kurmaktadır. Her bir gözlemin her bir kestirim değeri üzerindeki etkisini tanımlar. H matrisinin köşegen elemanları (h_{ii}) kaldıraç (leverage) olarak adlandırılır ve kaldıraç değerleri, gözlem değerlerinin tahmin değerleri üzerinde etkisini tanımlar.

Gözlem değerlerini Y , kestirim değerlerini $\hat{Y}$ temsil etmek üzere, kestirim denklemi hat matrisi cinsinden aşağıdaki gibi yazılır.

$$\begin{aligned}\hat{Y} &= X\hat{\beta} \\ \hat{Y} &= X(X^T X)^{-1} X^T Y \\ H &= X(X^T X)^{-1} X^T\end{aligned}\tag{3.1}$$

Buradaki $X(X^T X)^{-1} X^T$ ifadesi Hat Matrisi olarak adlandırılır. $\hat{Y}$ tahmin değerleri hat matrisi ve gözlem değerleri cinsinden

$$\hat{Y} = HY\tag{3.2}$$

şeklinde yazılabilir [Johnson, 2006].

h_{ii} , hat matrisinin köşegen öğeleri ve p , tahmin edilen parametre sayısı olmak üzere;

$$\sum_{i=1}^n h_{ii} = p, \text{ ve}$$

Burada h_{ii} değerleri, kaldıraç olarak adlandırılır. i 'inci gözlemin X_i değeri ile X değerlerinin ortalaması arasındaki uzaklık olarak da değerlendirilebilir.

Artıklar ise H ve Y gözlemleri cinsinden aşağıdaki gibi ifade edilebilir.

$$e = Y - \hat{Y} = Y - HY = (I - H)Y \quad (3.3)$$

Bu durumda, artıkların varyansı,

$$\begin{aligned} V(e) &= \sigma^2(I - H) \\ V(e_i) &= \sigma^2(1 - h_{ii}) \end{aligned} \quad (3.4)$$

olarak yazılır. Büyük h_{ii} değerleri ortalamadan olan uzaklığın büyük olduğunu gösterir. i 'inci gözlem X değeri bakımından aykırı değer ve büyük h_{ii} değerine sahipse bu durum $\hat{Y}$ tahmininde büyük kaldıraç etkisi yapar.

Hat matrisi ile aykırı gözlemler belirlenirken parmak kuralı olarak genellikle

$$h_{ii} \geq \frac{2p}{n} \quad (3.5)$$

ise ilgili gözlemin X değerine göre aykırı değer olduğu kabul edilir. Burada p , modelde tahmin edilen parametre sayısıdır [Neder, Wasserman, Kutner, 1996].

3.2.2. X gözlem değerlerine göre aykırı gözlemlerin artıklara dayalı olarak belirlenmesi

Regresyon analizinde tanı yöntemlerinin basit ve etkili bir yolu artıkları incelemektir. Artıklara ilişkin grafikler, varsayımlardaki bir ya da daha fazla bozulmayı işaret edebilir. Daha da önemlisi artıkların analizi, model yapısının belirlenmesini sağlayabilir ya da bilgi içeren kayıp verilerin olduğunu gösterebilir. Bu öneriler ya da ipuçları, modelin daha iyi anlaşılmasını ve çalışma sonucunda daha iyi bir model oluşturulmasını sağlar [Hadi ve Catterjee, 2006].

3.3. Artık Çeşitleri

3.3.1. Sıradan artıklar

Regresyon analizi sonucunda elde edilen model

$\hat{Y} = X\hat{\beta}$ olsun. Sıradan artıklar, Y ve $\hat{Y}$ arasındaki fark olarak tanımlanır. Yani sıradan artıklar;

$$\begin{aligned} e &= Y - \hat{Y} \\ e_i &= Y_i - \hat{Y}_i \end{aligned} \quad (3.6)$$

olarak tanımlanır [Hadi ve Catterjee, 2006]. Bunlara ham artıklar da denir.

3.3.2. Standartlaştırılmış artıklar

Sıradan artıkların varyansı (3.4)'te $Var(e_i) = \sigma^2(1 - h_{ii})$ olarak verilmişti. Görüldüğü gibi artıklar homojen varyansa sahip değildir. Bu problemin üstesinden gelmek için, *i'tinci* artık, standart hatasına bölünerek standartlaştırılır ve standartlaştırılmış artıklar aşağıdaki gibi gösterilir.

$$z_i = \frac{e_i}{\hat{\sigma} \sqrt{1 - h_{ii}}} \quad (3.7)$$

Burada varyansın tahmini ise

$$\hat{\sigma}^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n - p - 1} \quad (3.8)$$

biçimindedir. Standartlaştırılmış artıkların ortalaması 0, standart hataları ise 1 olur[Hadi ve Catterjee, 2006].

3.3.3. Student artıklar

Sıradan artıklar, varyansları h_{ii} 'ye bağlı olduğundan ölçeğe bağlı bir dağılıma sahiptir. Bu durumda artıkların ölçekten bağımsız olan Student biçimini tanımlamak daha yararlı olabilir.

σ^2 'nin alternatif yansız bir tahmin edicisi aşağıdaki gibi verilebilir.

$$\hat{\sigma}_{(0)}^2 = \frac{SSE_{(0)}}{n - p - 2} \quad (3.9)$$

Burada $SSE_{(0)}$ i. gözlem çıkarıldıktan sonra elde edilen modelden hesaplanan artık kare toplamıdır. Yukarıda verilen $\hat{\sigma}_{(0)}^2$, σ^2 'nin yansız bir tahmin edicisidir.

$\hat{\sigma}^2$ kullanılarak elde edilen student artık değeri İçsel Student Artık olarak ifade edilir ve aşağıdaki gibi gösterilir.

$$r_i = \frac{e_i}{\hat{\sigma} \sqrt{(1 - h_{ii})}} \quad (3.10)$$

$\hat{\sigma}_{(i)}^2$ kullanılarak elde edilen student artık değeri Dışsal Student Artık olarak ifade edilir ve aşağıdaki gibi gösterilir.

$$r_i^* = \frac{e_i}{\hat{\sigma}_{(i)} \sqrt{(1 - h_{ii})}} \quad (3.11)$$

Dışsal Student artık denmesinin sebebi, $\hat{\sigma}_{(i)}$ hesaplanmasında i^{inci} gözlemin dışlanmasıdır [Cook ve Weisberg, 1982].

3.4. Y(Bağımlı Değişken) Değerine Göre Aykırı Gözlemlerin Belirlenmesi

Yüksek artık değerine sahip olan gözlemler, Y değerine göre aykırı gözlem olarak nitelendirilirler. Çünkü bu gözlemler regresyon doğrusundan, Y yönünde, uzakta bulunmaktadır. Y değerine göre aykırı gözlemler belirlenirken artıkların grafikleri kullanılır. Artık grafikleri genellikle modelde var olan bozulmalar hakkında bilgi verirler. Y değerine göre aykırı gözlemleri belirlemek amaçlı başka yöntemler de önerilmiştir. Yaygın yöntem standartlaştırılmış artıklar, z_i 'yi kullanmaktır. Standart artıklar yaklaşık olarak $r_i \sim N(0,1)$ dağılımına uyduğundan, standart artıkları3 değerinden büyük olan gözlemler Y değerine göre aykırı gözlem olarak tanımlanırlar[Hadi ve Catterjee, 2006].

3.5. Doğrusal Regresyon Analizinde Etkili Gözlemler

Regresyon analizinde aykırı gözlemler belirlendikten sonra veri kümesindeki aykırı gözlemin regresyon modeli ve/veya katsayı tahminleri üzerinde bir etkiye sahip olup olmadığı incelenmelidir. Çünkü tek bir gözlem bile model ya da katsayılar üzerinde büyük bir etkiye sahip olabilir. Yukarıda verildiği gibi en büyük artık değerine sahip ya da veri kümesindeki gözlemlere uzak olan gözlem, aykırı gözlem olarak tanımlanır. Burada önemli bir soru ortaya çıkmaktadır. Aykırı gözlemler etkili gözlemler midir? Yani gözlem, veri kümesinden çıkarıldığında model üstünde veya katsayılar üzerinde önemli değişiklikler oluşmakta mıdır?

Regresyon analizinde gözlemler aykırı olarak nitelenebilir ancak söz konusu gözlemler regresyon katsayıları ya da model üstünde anlamlı bir değişikliğe yol açmıyorsa, bu gözlemlerin etkili gözlem olduğu söylenemez. Etkili gözlemler; veri kümesinden çıkartıldığında parametre tahminlerini önemli ölçüde değiştiren gözlemlerdir. Yapılan çalışmalarda aykırı gözlemler belirlendikten sonra bunların etkili olup olmadığının belirlenmesi gerekmektedir. Etkili gözlemlerin belirlenmesi için de çeşitli yöntemler geliştirilmiştir. Kullanılan yöntemlerden 4 tanesi aşağıda verilmektedir.

3.5.1. DFFITS

DFFITS (Difference Between the Fitted Values) *i*’inci gözlemin, tahmin değeri $\hat{Y}_i$ üzerindeki etkisini gösteren bir ölçüdür. **DFFITS**, hat matrisinin köşegen elemanları olan h_{ii} ile d_i^* değerlerini bir araya getiren bir istatistiktir. olmak üzere, d_i^*

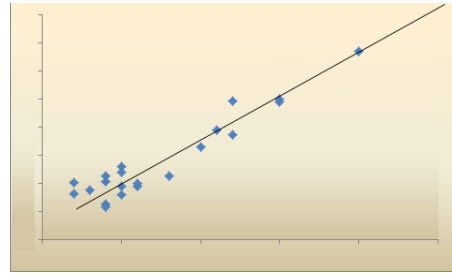
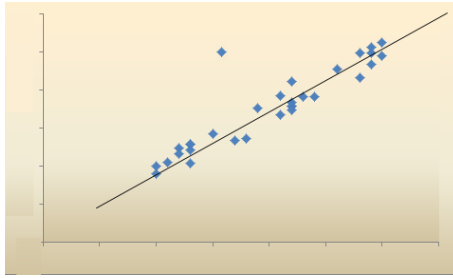
$$d_i^* = \frac{e_i}{\sqrt{MSE_i(1-h_{ii})}} \quad (3.12)$$

olarak yazılır. Burada MSE_i hata kare ortalama değeridir.

DFFITS *t'inci* gözlemin $\hat{Y}_i$ üzerindeki etkisini gösterir. Aşağıdaki gibi de elde edilebilir.

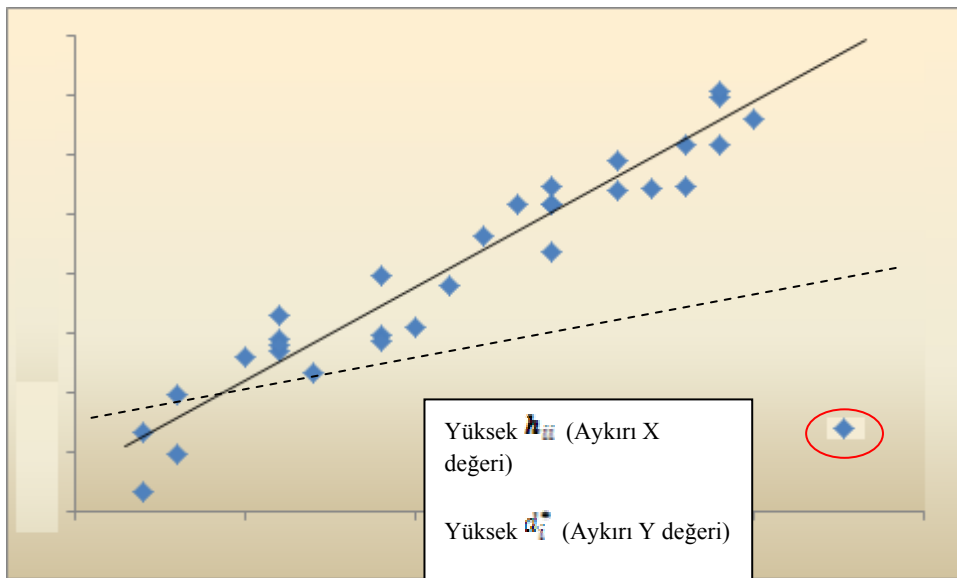
$$(DFFITS)_i = d_i^* \left(\frac{h_{ii}}{1 - h_{ii}} \right)^2 \quad (3.13)$$

Gözlemlerin tahmin değerlerini etkilemeleri için, hem yüksek kaldıraç etkisine sahip olmaları hem de Y değerine göre aykırı olarak belirlenmeleri gerekir. Aşağıdaki şekiller incelendiğinde bu durum daha açık bir şekilde anlaşılmaktadır.



Şekil 3.2. Düşük h_{ii} Yüksek d_i^*

Şekil 3.3. Yüksek h_{ii} düşük d_i^*



Şekil 3.4. Yüksek h_{ii} Yüksek d_i^*

Yukarıda görüldüğü gibi sadece hem yüksek R^2 hem de yüksek d_f^2 değerine sahip olan gözlemin bulunduğu Şekil (3.4)'te model etkilenmiştir[Habing,2004].

Gözlemlerin etkili olup olmadığına parmak hesabına dayalı olarak

$$DFFITS_{(i)} > 1 \quad (\text{Küçük ve orta büyüklükteki örnekler için})$$

$$DFFITS_{(i)} > 2\sqrt{\frac{p}{n}} \quad (\text{Büyük örnekler için})$$

ölçütleriyle karar verilir [Neder, Wasserman, Kutner, 1996].

3.5.2. DFBETAS

Bazı durumlarda model üzerindeki değil parametre tahminleri üzerindeki etki belirlenmek istenebilir. **DFBETAS** (Difference in Betas)istatistiği parametre tahminlerindeki değişimi ölçer ve tüm gözlemler ile elde edilen regresyon katsayıları ile i. gözlem çıkarıldığında elde edilen regresyon katsayılarının karşılaştırılmasına dayanır. Regresyon parametre tahminlerindeki değişimi ölçen **DFBETAS** istatistiği aşağıdaki gibi hesaplanır.

$$DFBETAS_{k(i)} = \frac{\hat{\beta}_k - \hat{\beta}_{k(i)}}{\sqrt{MSE_i c_{kk}}} \quad k = 0, 1, 2, \dots, p - 1 \quad (3.14)$$

$c_{kk} = (X'X)^{-1}$ Matrisinin köşegen elemanıdır.

MSE: Hata Kare Ortalama

Elde edilen **DFBETAS** değerleri;

$$DFBETAS > \frac{1}{2} \quad (\text{Küçük ve orta büyüklükteki örnekler için})$$

$$DFBETAS > \frac{2}{\sqrt{n}} \quad (\text{Büyük örnekler için})$$

olduğunda parmak kuralına dayalı olarak i^{inci} gözlemin etkili gözlem olduğuna işaret eder.

3.5.3. Cook uzaklığı (D_i)

En çok kullanılan etkililik ölçütlerinden birisi olan Cook Uzaklığı, tüm veri kümesi üzerinden hesaplanan regresyon katsayıları ile veri kümesinden i^{inci} gözlem çıkarıldıktan sonra elde edilen regresyon katsayıları arasındaki farkı ölçer [Hadi ve Catterjee, 2006]. Cook uzaklığı, i^{inci} gözlemin, tüm regresyon parametreleri üzerindeki etkisini gösteren bir ölçüdür. Cook uzaklığı;

$$D_i = \frac{e_i^2}{pMSE} \left[\frac{h_{ii}}{(1 - h_{ii})^2} \right] \quad (3.15)$$

şeklinde hesaplanabilir. D_i , iki faktöre bağlıdır. Artık büyüklüğü ve kaldıraç değerleri (e_i ve h_{ii}) ne kadar büyürse D_i değerinin de o kadar büyüyeceği açıktır. Cook uzaklığının 0,5'ten büyük değerleri için, ilgili gözlemin parametre tahminleri üzerinde önemli bir etkiye sahip olduğu söylenebilir.

3.5.4. Hadi'nin etkililik ölçütü (H_i)

Hadi(1992), etkili gözlemlerin bağımlı değişkene ya da bağımsız değişkene veya her ikisine göre aykırı olması üzerine kurulan bir etkililik ölçütü önermiştir. Buna göre i^{inci} gözlemin etkisi aşağıdaki gibi ölçülebilir.

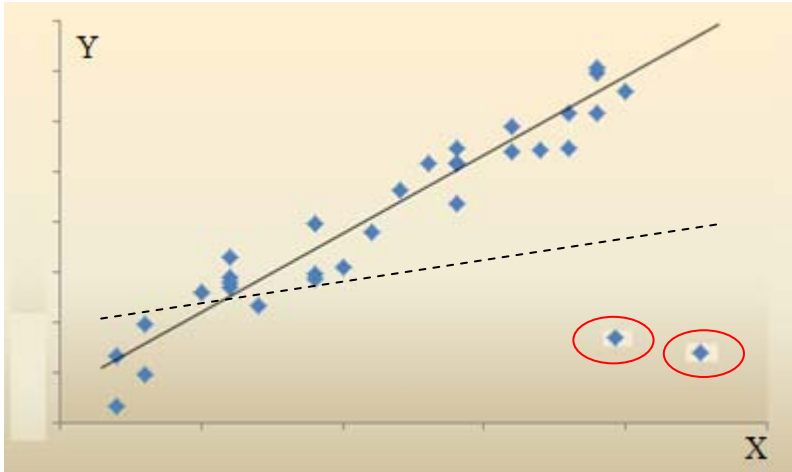
$$H_i = \frac{h_{ii}}{1 - h_{ii}} + \frac{p + 1}{1 - h_{ii}} \frac{d_i^2}{1 - d_i^2}, i = 1, 2, \dots, n \quad (3.16)$$

Burada $d_i = \frac{e_i}{\sqrt{SSE}}$ olarak ifade edilir.

Eşitsizliğin sağ tarafındaki ilk terim X değerlerine göre aykırılığı ölçer. İkinci terim, Y değerlerine göre aykırılığı ölçen, artıkların bir fonksiyonudur. Gözlem Y değerine göre ya da X değerine göre aykırı gözlem ise H_i değeri yüksek çıkar. Bu ölçüt gözlemlerin genel olarak etkililiğini ölçen bir yöntem olarak düşünülebilir. Cook Uzaklığı (D_i) ve $DFITS_i$ artık ve kaldıraç değerlerinin çarpımsal fonksiyonu iken H_i toplamsal fonksiyonudur [Hadi ve Catterjee, 2006].

3.6. Maskeleye ve Sahte Aykırı Gözlem Problemleri

Yukarıda verilen yöntemler aykırı gözlemlerin belirlenmesinde çoğu zaman iyi sonuç vermesine rağmen bazı durumlarda iyi çalışmayabilirler. Örneğin iki aykırı gözlem rassal olarak yakın değerler aldığı anda, bir aykırı gözlemi veri kümesinden çıkaran ve yeniden tahmin yapan araştırmacı, elde ettiği sonuçlarda farklılık göremeyebilir. Bunun sebebi veri kümesinde tutulan aykırı gözlemin, silinen aykırı gözlemin etkisini maskeleyesidir. Bu problem maskeleye (masking) olarak bilinir. Aşağıdaki şekil incelendiğinde bu etki daha net bir şekilde anlaşılmaktadır.

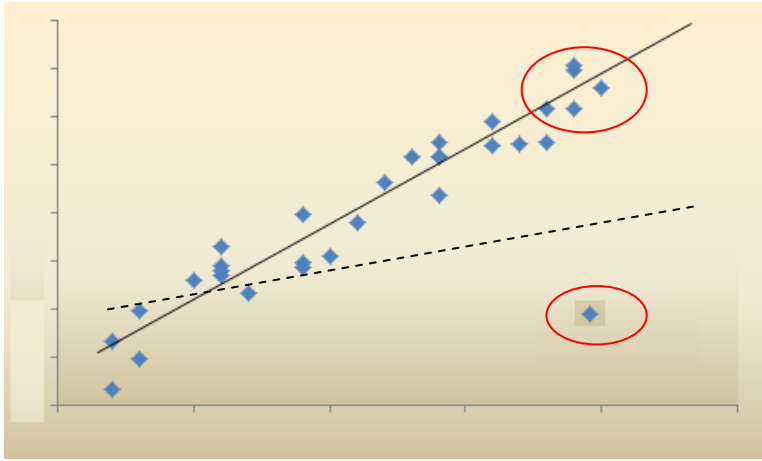


Şekil 3.5. Maskeleye Problemi

Şekil (3.5)'te görülen iki noktadan biri çıkarıldığında aykırı olan diğer gözlem halen veri kümesinde kalacağından regresyon fonksiyonu üzerindeki etkisi giderilemeyecektir. Burada kalan gözlem, çıkarılan gözlemin etkisini

maskeleymektedir. Yani maskeleye problemi var olan aykırı gözlemi belirlememizi engellemektedir[Hadi ve Catterjee, 2006].

Aykırı ve etkili gözlemlerde karşılaşılan bir diğer problem ise sahte aykırı gözlem problemidir (swamping). Bu durum aykırı olmayan bir gözlemin aykırı gözlem olarak belirlenmesidir. Aykırı gözlemler regresyon doğrusunu kendilerine doğru çekerler, bu nedenle diğer gözlemler regresyon doğrusundan uzaklaşmış olur ve aykırı gözlem gibi görünürler.



Şekil 3.6. Sahte Aykırı Gözlem Problemi

Şekil (3.6)'da ise aykırı gözlem regresyon fonksiyonunu kendisine doğru çekmiştir. Bu nedenle veri kümesinde aykırı olmayan gözlemler aykırı gibi görünecektir. Yani sahte aykırı gözlem problemi, aykırı olmayan gözlemlerin aykırı olarak belirlenmesine sebep olan bir problemdir.

3.7. Çoklu Aykırı Gözlemlerin Belirlenmesi

Klasik aykırı gözlem belirleme yöntemleri sadece tek aykırı gözlemdurumunda güçlü olan yöntemlerdir. Örnekte birden fazla aykırı gözlem bulunduğunda, bu yöntemlerin gücü büyük oranda azalmaktadır. Hempel (1986) yaptığı çalışmada, normal veri setlerinde % 1-10 oranında aykırı gözlem bulunduğunu iddia etmiştir[Rabiah, Mohd, Halim, 2003]. Gerçek veri setlerinde, doğru etki yapısının belirlenmesi için, birden

fazla aykırı gözlemin aynı anda incelenmesi gerekir. Veri kümesinde tek bir aykırı gözlem bulunması durumunda, standart tanı ölçütleri ve grafikler ile bu gözlemler başarılı bir biçimde belirlenebilir. Ancak bu yöntemler çoklu aykırı gözlemlerin belirlenmesinde, özellikle aykırı gözlemlerin birbirine yakın konumda kümелendiği durumlarda, başarısız olabilir. Tek aykırı gözlem belirleme yöntemleri çoklu aykırı gözlem durumunda kullanıldığında veri kümesindeki aykırı gözlemlerin belirlenememesi (maskeleye), aykırı olmayan gözlemlerin aykırı olarak belirlenmesi (sahte aykırı gözlemler) ya da bu problemlerin her ikisi ile de karşılaşılabilir [Wisnowski, Montgomery, Simpson, 2001]. Ayrıca bu durumda aykırı gözlemler için birleşik etki incelemeleri büyük bir hesaplama yükü oluşturacak ve çoklu gözlem etkileri iyi anlaşılamayacaktır. Son yıllarda çoklu aykırı gözlemlerin belirlenmesi büyük ilgi çekmiş ve bu alanda birçok yöntem önerilmiştir.

Hadi ve Simonoff (1993) çoklu aykırı gözlem belirleme yöntemlerini doğrudan ve dolaylı yöntemler olarak ikiye ayırmıştır. Literatürde kullanılan doğrudan yöntemlerin birçoğu, aykırı gözlemlerin sıralı olarak veri kümesinden çıkarılması (geriye doğru araştırma) ya da sıralı olarak eklenmesi (ileri doğru araştırma) üzerine kurulmuştur. Geriye doğru araştırmalarda, araştırmanın başında tüm veri kümesi değerlendirilmeye alınır ve gözlemler artık değeri büyük olan gözlemde başlanarak veri kümesinden çıkarılır. İleri doğru araştırmalarda ise benzer olarak veri kümesinin bir altkümesi alınır ve gözlemler bu kümeye sırayla eklenerek araştırma yapılır. İleri doğru araştırmalar genellikle geriye doğru araştırmalardan daha iyi sonuç verirler. Bu alanda geliştirilen yöntemlere örnek olarak Hadi ve Simonoff (1993, 1997), Swallow ve Kianifard (1996) tarafından önerilen yöntemler gösterilebilir. Ayrıca Pena ve Yoahi (1995) tarafından etki matrisinin özdeğer yapısı üzerine kurulan yöntem ve Sebert (1998) tarafından kümeleme algoritması üzerine kurulan yöntem de önerilen yöntemlerden bazılarıdır. Dolaylı yöntemlerde ise robust regresyon tahminleri kullanılır. Robust regresyon tekniklerinde aykırı gözlemler, ağırlıkları azaltılarak ya da parametre tahminleri üzerinde etkili olmadıkları gösterilmek üzere ihmal edilerek modele uyumlu hale getirilirler. Aykırı gözlemlerin, gözlemlere ait son ağırlıklara ya da artık büyüklüklerine bakılarak belirlenmesi mümkündür. Çoklu aykırı gözlemlerin

belirlenmesine yönelik birçok robust regresyon yöntemi kullanılmıştır. Bu yöntemlerin en yaygın olarak kullanılanları Rousseeuw tarafından 1984 ve 1985 yıllarında önerilen En Küçük Kareler Ortancası ile En Küçük Kırpılmış Kareler ve Huber tarafından 1973 yılında önerilen M tahmin edicileridir [Wisnowski, Montgomery, Simpson, 2001].

4. LOJİSTİK REGRESYON

Regresyon analizinde bağımlı değişken genellikle sürekli ve ölçülebilir alınır. Bağımsız değişkenler ise hem ölçülebilir hem de sınıflanabilir olabilir. Ancak bazı durumlarda bağımlı değişkenin iki uçlu değerler aldığı durumlara rastlanabilir. Bağımlı değişkenin iki uçlu değerler aldığı durumda en yaygın olarak kullanılan regresyon modeli Lojistik Regresyon modelidir[Hosmer ve Lemeshow, 2000].

4.1. İki Uçlu Verileri Modelleme

Lojistik regresyonda bağımlı değişkenler iki uçlu değer alırlar. Bu farklılık parametrik model seçimine de yansır [Hosmer ve Lemeshow]. Bağımlı değişken ilgili olduğu konu ile ilgili olma-olmama, başarılı-başarısız gibi değerler alabilir. Bu durumda bağımlı değişken her zaman 0 ya da 1 değerleri ile gösterilebilir[Hadi ve Catterjee, 2006].

Bağımlı değişkenin iki uçlu değer aldığı durumlarda p_i olayın gerçekleşmesi olasılığı olmak üzere; bağımlı değişken $Y_i = 1$ değerini p_i olasılıkla, $Y_i = 0$ değerini $1 - p_i$ olasılıkla alır.

Beklenen değer tanımından

$$E(Y_i) = \sum Y_i p_i = (1 * p_i) + (0 * (1 - p_i)) = p_i \quad (4.1)$$

olur.

(4.1) eşitliğinde görüldüğü üzere regresyondan elde edeceğimiz koşullu beklenen değer p_i 'ye eşittir. p_i , bir olasılık olduğundan, koşullu beklenen değer bir olasılık değeri vermesini ve 0 ile 1 arasında olmasını bekleriz ($0 \leq E(Y_i/X_i) \leq 1$)[Gujarati, 1999]. Dolayısıyla bağımlı değişkenin iki uçlu değer

aldığı durumlarda bağımlı değişkenin 0 ve 1 değerlerini tahmin etmek yerine bağımlı değişkenin bu değerleri alma olasılıkları modellenir[Hadi ve Catterjee, 2006].

Bağımlı değişkenin iki uçlu değerler aldığı durumlarda EKK tahmin yöntemi kullanılırsa bazı sorunlar ortaya çıkmaktadır.

1. Hata terimlerinin normal dağılıma uymaması

Bağımsız değişken iki uçlu değer aldığı anda hata terimlerinin dağılımı normal dağılıma uymaz. Bu, aşağıdaki gibi gösterilebilir.

$$Y = X\beta + \varepsilon \text{ olsun.}$$

Bu durumda

$$Y_i = 1 \rightarrow \varepsilon_i = 1 - X\beta$$

$$Y_i = 0 \rightarrow \varepsilon_i = -X\beta$$

olur. Bu durumda da artık değerlerinin Binom Dağılımına uyduğu görülmektedir [Hosmer ve Lemeshow, 2000].

2. $0 \leq E(Y_i/X_i) \leq 1$ koşulunun yerine gelmemesi

EKK yönteminin kullanılmasının yarattığı hata terimlerinin normal dağılıma uymaması ve $0 \leq E(Y_i/X_i) \leq 1$ koşulunun yerine gelmemesi sorunlarının aşılabilmesi için çeşitli yöntemler kullanılabilir. Normal dağılmama sorununu aşmak için örnek büyütülebilir veya sınırlanmış EKK yöntemine başvurulabilir, tahmin edilen olasılıkların 0-1 aralığında çıkması sağlanabilir. Ancak, yine de iki uçlu değer alan bağımlı değişkenlerde doğrusal model kullanılmasının temel sorunu, mantıksal

olarak uygun olmamasıdır. Doğrusal modelde, P_i olasılığının X_i 'ye doğrusal bağlı olarak arttığı yani; X_i 'nin değişiminde, bağımlı değişkene sağlanan katkının hep aynı kaldığı varsayılır. Ancak, P_i 'nin X_i 'ye doğrusal olmayan bir biçimde bağlı olması beklenmektedir.

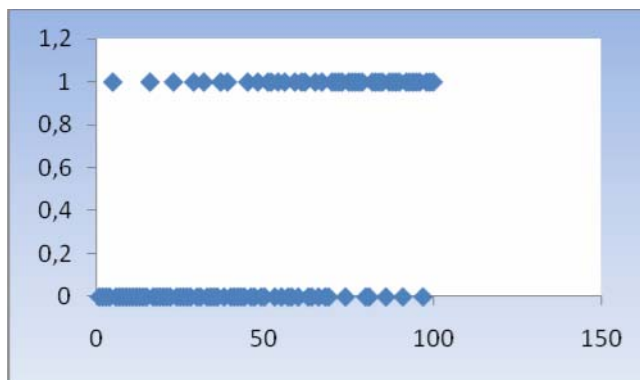
Dolayısıyla, aşağıda verilen iki özelliği taşıyan bir model bunu sağlayabilir.

- 1- X_i arttıkça P_i artmalı, ancak 0-1 aralığı dışına çıkmamalı,
- 2- P_i ile X_i arasındaki ilişki doğrusal olmamalıdır. Yani P_i 'deki değişim X_i 'deki değişime göre daha az olmalıdır.

Bu gibi problemlerde yukarıdaki özellikleri sağlayan ve kullanımı kolay olan Lojistik Dağılım Fonksiyonu kullanılmaktadır [Hosmer ve Lemeshow, 2000].

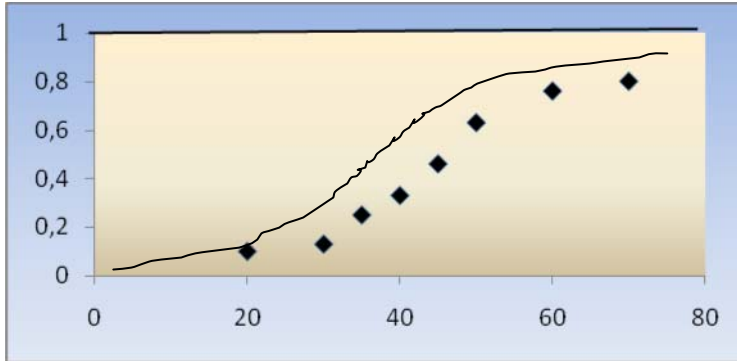
4.2. Logit Model

$((X_i, Y_i), i = 1, 2, \dots, n)$ büyüklüğünde gözlem çiftleri olsun. Burada Y_i iki uçlu değer alan bağımlı değişken olsun ve 0 - 1 değerleri ile kodlansın. Bu verilere ilişkin dağılım grafiği çizildiğinde aşağıdaki gibi bir şekil elde edilir.



Şekil 4.1. İki uçlu değer alan bağımlı değişkene ilişkin dağılım grafiği

Grafiğe bakıldığında değişkenler arasında doğrusal bir ilişki tanımlanamamaktadır. Bu durumdan kurtulmanın yollarından birisi bağımsız değişken için aralıklar oluşturarak verilerin gruplandırılmasıdır. Veriler gruplandırıldıktan sonra, grupların olasılıklara göre grafiği çizildiğinde, aşağıdaki gibi bir şekil elde edilir.



Şekil 4.2. Grupların olasılıklara göre grafiği

Grafiğe bakıldığında bağımsız değişken değerleri ile olasılıklar arasındaki ilişkinin S şeklinde olduğu yani olasılık değerlerinin 0'a ve 1'e doğru yavaş yavaş yaklaştığı görülür. İşte bu noktada doğrusal model kullanılması durumuna göre farklılıklar ortaya çıkmaktadır. İlk fark bağımlı ve bağımsız değişkenlerin arasındaki ilişkinin doğasıdır. Doğrusal regresyon analizinde bağımlı değişken, bir koşullu ortalamadır.

$$E(Y/X) = X\beta \rightarrow (\text{Doğrusal}) \quad -\infty < X < \infty \quad (4.2)$$

Lojistik regresyonda koşullu ortalama bir olasılık değeri ifade ettiği için koşullu ortalama 0 ile 1 arasında olur. Ayrıca bu ortalamanın 0 ve 1'e azalan eğimle yaklaştığı görülmektedir. $E(Y/X)$ 'teki değişme X 'teki değişime göre daha yavaş olmaktadır ve eğri bir S şeklini almaktadır. Bu eğri rasgele değişkenin birikimli dağılımını temsil etmektedir. Bu ilişkiyi tanımlamak için Lojistik Dağılım Fonksiyonu kullanılabilir.

Lojistik regresyonda $E(Y/X) = \pi(x)$ olarak alınırsa kullanılan özel model;

$$\pi(x) = \frac{e^{x\beta}}{1 + e^{x\beta}} \quad (4.3)$$

olarak tanımlanır. Burada $X^T = (X_1, X_2, \dots, X_p)$ olarak gösterilen bağımsız değişken vektörü, $\beta^T = (\beta_1, \beta_2, \dots, \beta_p)$ olarak gösterilen parametre vektörüdür.

Bu eşitlik, Lojistik Dağılım Fonksiyonu olarak adlandırılmakta olup, Lojistik Regresyon çalışmasının temelini oluşturmaktadır. Burada doğrusal olmayan bir model kullanılmaktadır. Bu nedenle tahminler elde edilirken doğrusal modelin sağladığı kolaylıklardan yararlanmak amacıyla $\pi(x)$ kullanılarak bir dönüşüm yapılmaktadır. Bu dönüşüm fonksiyonu Logit olarak adlandırılır. Logit fonksiyonu $g(x)$ ile gösterilirse

$$g(x) = \ln \left[\frac{\pi(x)}{1 - \pi(x)} \right] = X\beta \text{ (Logit)} \quad (4.4)$$

biçiminde olur.

Bu dönüşümün özelliği; $g(x)$ 'in, doğrusal regresyonun istenen birçok özelliğine sahip olmasıdır. $g(x)$ kendi parametrelerinde doğrusaldır ve X 'in tanım aralığına göre tanım aralığı alır.

Lojistik regresyon modelinin oluşturulması için $\beta^T = \beta_1, \beta_2, \dots, \beta_p$ parametre vektörünün tahmin edilmesi gerekmektedir. Doğrusal regresyonda, hesaplama kolaylığı ve geleneksel olarak parametre tahminleri için EKK yöntemi kullanılır. Lojistik Regresyonda EKK için gerekli varsayımlar sağlanmamaktadır. Bu nedenle bu tahminler için Lojistik regresyonda genel olarak kullanılan yöntem En Çok Olabilirlik (EOB) yöntemidir. Bu yöntemde başvurabilmek için ilk önce olabilirlik fonksiyonunun oluşturulması gerekmektedir. Parametrelerin en çok olabilirlik

tahmin edicileri, EOB fonksiyonunu en büyük yapan parametre değerleri olarak belirlenir [Hosmer ve Lemeshow, 2000].

$X^T = (X_1, X_2, \dots, X_p)$ olarak gösterilen bağımsız değişken vektörü, $\beta^T = (\beta_1, \beta_2, \dots, \beta_p)$ olarak gösterilen parametre vektörü olsun. Lojistik regresyon modeli logit dönüşümü yapıldıktan sonra

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \sum x_{ip} \beta_p, i = 1, 2, \dots, n \quad (4.5)$$

biçimini alır.

Her bir bağımsız değişken binom dağılımına sahip olduğu için bileşik olasılık fonksiyonu yani olabirlik fonksiyonu

$$l(\beta) = \prod \binom{n_i}{Y_i} \pi_i^{Y_i} (1-\pi_i)^{n_i-Y_i} \quad (4.6)$$

biçiminde yazılır. Bu fonksiyonun en büyüklenmesi için türevinin alınıp sıfıra eşitlenmesi gerekmektedir. Bu fonksiyonun türevinin alınması oldukça zordur ancak olabirlik fonksiyonu basitleştirilebilir. Faktöriyelli sabit terimler ihmal edilip, kalan terimlerde gerekli düzenlemeler yapıp daha sonra gerekli işlemler yapıldıktan sonra olabirlik fonksiyonu aşağıdaki hale getirilir.

$$l(\beta) = \sum Y_i (\sum x_{ip} \beta_p) - n_i \log(1 + e^{\sum x_{ip} \beta_p}) \quad (4.7)$$

Olabilirlik fonksiyonunu en büyülemek için (4.7) nolu eşitliğine aşağıdaki gibi β_k 'ya göre türevi alınır ve sıfıra eşitlenir.

$$\begin{aligned}
\frac{\partial l(\beta)}{\partial \beta_k} &= \sum Y_i x_{ik} - n_i \frac{1}{1 + e^{\sum x_{ip} \beta_p}} \frac{\partial}{\partial \beta_k} (1 + e^{\sum x_{ip} \beta_p}) \\
\frac{\partial l(\beta)}{\partial \beta_k} &= \sum Y_i x_{ip} - n_i \frac{1}{1 + e^{\sum x_{ip} \beta_p}} e^{\sum x_{ip} \beta_p} \frac{\partial}{\partial \beta_k} \sum x_{ip} \beta_p \\
\frac{\partial l(\beta)}{\partial \beta_k} &= \sum Y_i x_{ip} - n_i \frac{1}{1 + e^{\sum x_{ip} \beta_p}} e^{\sum x_{ip} \beta_p} x_{ip} \\
\frac{\partial l(\beta)}{\partial \beta_k} &= \sum Y_i x_{ip} - n_i \pi_i x_{ik}
\end{aligned} \tag{4.8}$$

β^T vektörünün tahmin edicilerinin bulunması için (4.8) nolu eşitlik sıfıra eşitlenir ve tahmin değerleri elde edilir.

Model elde edildikten sonra uyum iyiliği ölçütleriyle modelin uyumuna bakılır. Lojistik regresyonda uyum iyiliği ölçütleri olarak çoğu paket programda bulunan Perason Ki-Kare ve Hosmer-Lemeshow istatistikleri yaygın kullanım alanı bulmuşlardır. Söz konusu istatistik değerleri ve ki-kare değerinden elde edilen P olasılık değeri değerlendirilerek modelin uyumu hakkında karar verilir. Hosmer-Lemeshow istatistiğinin büyük olması modelin uyumunun iyiliğini gösterir [Hosmer ve Lemeshow, 2000].

4.3. Lojistik Regresyonda Aykırı Gözlemler

Tanı yöntemleri Regresyon Analizinin tüm dallarında kullanılmaktadır. Aykırı gözlemlerin belirlenmesi de bu tanı yöntemlerinden birisidir. Aykırı gözlemlerin varlığı, yapılacak parametre tahminlerinin ve yorumların yanlış olmasına sebep olabilir. Ayrıca EOB yöntemi, ideal veri setleri için iyi sonuçlar verirken aykırı gözlemlere karşı oldukça hassastır [Hadi ve Imon, 2008].

Lojistik regresyonda aykırı gözlemler, bağımlı değişkenin 0-1 değerlerini alma olasılıkları modellendiğinden, gerçekte başarılı sonuç alan gözlemin bu değeri alma olasılığının küçük çıkmasına ya da gerçekte başarısız sonuç alan gözlemin bu değeri alma olasılığının yüksek çıkmasına sebep olabilirler. Böylece lojistik regresyonda

aykırı gözlemler modelden elde edilen başarı olasılığı küçük olduğu halde gözlemlenen durumu başarılı olan ya da modelden elde edilen başarı olasılığı yüksek olduğu halde gözlemlenen durumu başarısız olan gözlemler olarak tanımlanabilirler.

Doğrusal regresyon modelleri için birçok aykırılık ve etkililik ölçütü geliştirilmiştir ve bunlar her bir gözlemin modelde nasıl davrandığının anlaşılmasına yardım eder. Böylece, söz konusu gözlemin diğer gözlemlerden uzakta olup olmadığı ya da tahmin modelini etkileyip etkilemediği belirlenebilir. Benzer yöntemler, lojistik regresyon modelleri için de geliştirilmiştir.

4.3.1. Lojistik regresyonda artıklar ve kaldıraç değerleri

Artık ve kaldıraç değerleri, doğrusal regresyonda olduğu gibi lojistik regresyonda da aykırı gözlemlerin belirlenmesinde oldukça önemli bir yer tutmaktadır.

Lojistik regresyon modeli

$$Y = \pi(x) + \varepsilon \quad (4.9)$$

ve

$$\pi(x) = \frac{e^{x\beta}}{1 + e^{x\beta}} \quad (4.10)$$

olsun. Burada;

$$X\beta = \beta_1 + \beta_2 X_2 + \dots + \beta_p X_p ,$$

$Y : n \times 1$ boyutlu vektör,

$Y_i = 0$ başarısızlık, $Y_i = 1$ başarı durumu,

$X : n \times k$ boyutlu matris ve her durum için $k = p + 1$,

$\beta^T = (\beta_1, \beta_2, \dots, \beta_p)$ regresyon parametreleri vektörü,

$\epsilon : n \times 1$ boyutlu rassal hata vektörü olarak tanımlanır.

EOB yöntemi ile model tahmin edildikten sonra, $\hat{\beta}$ tahmin edilen katsayıların vektörünü gösterebilir. Böylece lojistik modelin tahmin değerleri $\hat{\pi}(x_i)$, (4.10)'da verilen model kullanılarak bulunur.

Lojistik Regresyonda i' inci değişkenin tahmin değeri $\hat{Y}_i = \hat{\pi}_i$ olarak gösterilir.

Böylece i , artık

$$e_i = Y_i - \hat{\pi}_i, \quad i = 1, 2, \dots, n \quad (4.11)$$

olarak belirlenir.

Doğrusal regresyonda artık değerleri sıklıkla Hat Matrisi terimleri olarak ifade edilir ve birçok tanı yönteminin temelini oluştururlar. Pregibon(1981), ağırlıklandırılmış EKK yöntemini kullanarak tahmin değerlerine doğrusal yaklaşan ve lojistik regresyonda Hat Matrisinin yerini alan bir matris önermiştir. Bu matris

$$H = V^{\frac{1}{2}} X (X^T V X)^{-1} X^T V^{\frac{1}{2}} \quad (4.12)$$

olarak tanımlanır. Burada $V : v_i = \hat{\pi}_i(1 - \hat{\pi}_i)$ ile $n \times n$ köşegen matristir.

h_{ii} (4.11)'de tanımlanan Hat Matrisinin i' inci köşegen elemanını gösterebilir. Buradan köşegen elemanları doğrudan

$$h_{ii} = \hat{\pi}_i(1 - \hat{\pi}_i) x_i^T (X^T V X)^{-1} x_i \quad (4.13)$$

eşitliği ile hesaplanabilir.

(4.12)'deki $x_i^T = [1, x_{1i}, x_{2i}, \dots, x_{pi}]$, $1 \times k$ boyutlu vektördür.

4.3.2. Lojistik regresyonda aykırı gözlem belirleme yöntemleri

Lojistik regresyonda, artıklar kötü uyumlu faktör / değişkenin büyüklüğünü ölçer. Böylece büyük artığa sahip gözlemlerin aykırı gözlem olup olmadığından şüphelenilebilir. Daha önce (4.11)'de tanımlanan artıklar ölçekten bağımsız değildir. Bu nedenle aykırı gözlemleri belirlemede doğrudan kullanılamazlar. Doğrusal regresyonda anahtar yaklaşım hata varyansının koşullu ortalama $E(Y_i/X_i) = \pi_i$ ye bağlı olmamasıdır. Bununla beraber lojistik regresyonda Bernoulli hataları vardır ve hata varyansı, koşullu ortalamanın bir fonksiyonudur.

$$Var(Y_i/X_i) = v_i = \pi_i(1 - \pi_i) \quad (4.14)$$

Sonuçta tanı yöntemi için kullanılabilecek ölçekten bağımsız bir artık, aykırı gözlemleri belirlemek için kullanılabilir. Aşağıda lojistik regresyonda aykırı gözlem belirleme yöntemlerinden iyi bilinenleri verilmektedir[Hadi ve Imon, 2008].

4.3.3. Pearson artıkları (Pearson residuals)

Pearson Artıkları, gözlenen ve tahmin edilen sıklık arasındaki farkın standart hatasına oranı olarak tanımlanmıştır. Gözlenen ve tahmin edilen değerler arasındaki göreceli sapmayı ölçer. Pearson artıkları r_{pi} , aşağıdaki gibi hesaplanır.

$$r_{pi} = \frac{Y_i - \hat{\pi}(x_i)}{\sqrt{\hat{\pi}(x_i)(1 - \hat{\pi}(x_i))}} \quad (4.15)$$

Burada $\hat{\pi}(x_i)$ kestirim değerleridir. r_{pi} 'nin karesi t^2_{inci} gözlemin Pearson'ın Ki-Kare istatistiğine olan katkısıdır ve bu yüzden Pearson Artığı adını almaktadır. Pearson Ki-Kare istatistiği ise aşağıdaki gibi verilmektedir.

$$\chi^2 = \sum \frac{(Y_i - \hat{\pi}(x_i))^2}{\hat{\pi}(x_i)(1 - \hat{\pi}(x_i))} \quad (4.16)$$

[Hosmer ve Lemeshow, 2000]

Bir gözlemin ilgili Pearson Artık değeri, kabul edilen kritik değeri mutlak değer olarak aşıyorsa o gözleme aykırı gözlem diyebiliriz. Parmak kuralı olarak, Pearson Artık değeri mutlak değer olarak 3'ten büyükse ilgili gözlem aykırı olarak nitelenebilir. Bu normal dağılımdaki uzaklığı kuralı ile eşleşir [Hadi ve Imon, 2008].

4.3.4. Standart pearson artıkları

Pearson artıkları standartlaştırılmış artıklar değildir. Ayrıca bu artıklar, sabit varyansa da sahip değildir. Dolayısıyla bu artıkların ölçekten bağımsız hale getirilmesi gerekmektedir. Standart pearson artıkları, pearson artıklarını ölçekten bağımsız hale getirmek ve istenen özelliklere sahip olmalarını sağlamak amacıyla standartlaştırır. Pregibon Hat Matrisinin köşegen elemanları h_{ii} olmak üzere Standart Pearson Artıkları;

$$r_{sp_i} = \frac{r_{pi}}{\sqrt{1 - h_{ii}}} \quad (4.17)$$

ya da

$$r_{sp_i} = \frac{y_i - \hat{\pi}(x_i)}{\sqrt{\hat{\pi}(x_i)(1 - \hat{\pi}(x_i))(1 - h_{ii})}} \quad (4.18)$$

olarak gösterilir. $r_{sp_i} \geq 3$ ise o gözlemin aykırı gözlem olduğu söylenebilir [Hadi ve Imon, 2008].

4.3.5. Sapma artıkları

Sapmaların farkları üzerine oluşturulmuş bir artık çeşididir. Sapma artığı *İ'inci* gözlemin sapmaya yaptığı katkıyı ölçer. Lojistik regresyonda sapmaartığı

$$d_i = \text{sign}(Y_i - \hat{\pi}(x_i)) \left\{ 2 \left[Y_i \ln\left(\frac{Y_i}{\hat{\pi}(x_i)}\right) + (1 - Y_i) \ln\left(\frac{(1 - Y_i)}{(1 - \hat{\pi}(x_i))}\right) \right] \right\}^{\frac{1}{2}} \quad (4.19)$$

olarak verilmiştir. Buradaki yaklaşım gözlemler silindikten sonra sapma üzerinde meydana gelecek olan değişimin ölçülmesidir. Sapmadaki değişim

$$\Delta d_i = d_i^2 + \frac{r_{pi}^2 h_{ii}}{(1 - h_{ii})} \quad (4.20)$$

r_{pi}^2 ve d_i^2 yer değiştirirse sonuç olarak sapmadaki değişim aşağıdaki şekilde hesaplanabilir.

$$\Delta d_i = \frac{d_i^2}{(1 - h_{ii})} \quad (4.21)$$

[Hosmer ve Lemeshow, 2000]

4.3.6. Delta ki-kare($\Delta\chi^2$)

Delta ki-kare istatistiği *i'inci* gözlem çiftindeki tüm gözlemler sildikten sonra Pearson ki-kare istatistiğinde oluşacak değişmeyi ölçer.

$$\Delta\chi^2 = \frac{r_{pi}^2}{1 - h_{ii}} = r_{spi}^2 \quad (4.22)$$

Burada r_{pi} Pearson artıkları ve h_{ii} kaldıraç değerleridir [Hosmer ve Lemeshow, 2000].

Yukarıda verilen aykırı gözlem belirleme yöntemlerine ilişkin değerler belirlendikten sonra bu değerlerin olasılık ya da gözlem numaralarına göre grafikleri çizdirilerek aykırı gözlemler belirlenebilir. Ya da parmak kuralı ile belirlenen kritik değerler ile aykırı gözlemler belirlenmeye çalışılır. Ancak parmak kuralının kesin sonuç vermediğini, aykırı gözlem varlığına işaret ettiğini belirtmek gerekir. Yukarıda verilen yöntemler arasında gücü daha fazla olan yöntem pearson artıklarıdır [Hosmer, Cessie, Lemeshow, 1997].

5. LOJİSTİK REGRESYONDA ÇOKLU AYKIRI GÖZLEMLERİN BELİRLENMESİ

Aykırı gözlem belirleme yöntemleri istatistikte geniş bir uygulama alanı bulmuş ve birçok yöntem geliştirilmiştir[Barret ve Gray, 1997]. Klasik aykırı gözlem belirleme yöntemleri, veri kümesi sadece bir aykırı gözlem içerdiğinde güçlü olan yöntemlerdir. Bu yöntemlerin gücü veri kümesinde birden fazla aykırı gözlem olduğunda sert bir biçimde azalmaktadır[Hadi, 1992].

Tek aykırı gözlem belirleme ölçütleri, çoklu durumlarda uygulanabilir. Ancak çoklu durum, tekilduruma göre çok daha fazla hesap yükü getirmektedir. Çoklu aykırı gözlemlerin belirlenmesi için Doğrusal Regresyonda birçok yöntem önerilmiştir. Cook ve Weisberg(1980), Gray ve Ling(1984), Rousseeuw ve Van Zomeren(1990), Paul ve Fung (1991), Barrett ve Gray(1992), Hadi ve Siminoff(1993) bunlardan bazılarıdır[Barret ve Gray, 1997].

Lojistik regresyonda datekil aykırı gözlemleri belirlemek için birçok yöntem geliştirilmiştir. Fakat gerçek yaşamda veri setlerinin tek bir aykırı gözlem içermesi garanti değildir. Veri kümesinde birden çok aykırı gözlem bulunduğunda, bunları belirlemek için tekil durum için geliştirilen yöntemlere başvurulabileceği düşünülebilir. Ancak bu durumda bazı problemler ortaya çıkmaktadır. Aykırı gözlemler modelin uyumunu bozabilir. Bu yüzden aykırı gözlemler sahte olarak küçük artıklara sahip olabilir ve aykırı gözlem olarak görülmeyebilir. Bu problem literatürde maskeleye olarak bilinir. Maskelemenin ters etkisi ise sahte aykırı gözlem olarak bilinir. Bu durumda ise aykırı olmayan gözlemler aykırı gözlem olarak görülür. Çoklu aykırı gözlemin varlığı durumunda çoğu yöntem maskeleye ve sahte aykırı gözlem probleminden etkilenir. Bu nedenle bu problemlerden etkilenmeyecek bir yönteme ihtiyaç duyulmaktadır.

5.1. Lojistik Regresyonda Çoklu Aykırı Gözlemlerin Belirlenmesi İçin Genelleştirilmiş Standart Pearson Artıkları (GSPR)

Son yıllarda tanı yöntemleri lojistik regresyonun önemli bir parçası olmuştur [Hosmer ve Lemeshow, 2000]. Yukarıda verildiği gibi lojistik regresyonda da aykırı gözlemlerin belirlenmesi için birçok yöntem önerilmiştir. Ancak bu yöntemler tek bir aykırı gözlemin belirlenmesi için geliştirilmiş yöntemlerdir. Bu yöntemler çoklu aykırı gözlem varlığında uygulandığında güçlerini yitirmektedirler. Ayrıca bu yöntemlerle şüpheli aykırı gözlemlerin tüm kombinasyonları değerlendirilmek istendiğinde aşırı bir hesaplama yükü oluşmaktadır. Çoklu aykırı gözlem varlığında, çoğu yöntem maskeleye ve sahte aykırı gözlem problemlerinden etkilenmektedir. Yani aykırı olan gözlemler belirlenememekte ya da aykırı olmayan gözlemler aykırı gözlem olarak belirlenebilmektedir. Dolayısıyla çoklu aykırı gözlemlerin belirlenmesinde maskeleye ve sahte aykırı gözlem problemlerinden etkilenmeyen bir yöntem ihtiyacı duyulmaktadır.

Lojistik regresyonda çoklu aykırı gözlemlerin belirlenmesi için sadece birkaç yöntem üzerinde çalışılmıştır. Bunlardan bir tanesi Bedrick ve Hill tarafından 1990 yılında önerilmiştir. Bu yöntem ile çoklu aykırı gözlemlerin belirlenmesi için kümeleme analizi tekniğine dayanan bir test önerilmiştir.

Bu çalışmada, lojistik regresyonda çoklu aykırı gözlemlerin belirlenebilmesi amacıyla, Hadi ve Imon tarafından 2008 yılında önerilen, maskeleye ve sahte aykırı gözlem problemlerinden etkilenmeyen Genelleştirilmiş Standart Pearson Artıkları (GSPR) yöntemi tanıtılacaktır. Bu yöntemde artıkların grup olarak silinmiş hali ve lojistik regresyonda çoklu aykırı gözlemleri belirlemek için kullanılacak olan ağırlıklar verilmektedir.

n büyüklüğünde bir örnek olsun. Model oluşturulmadan önce n büyüklüğündeki örnekten d gözlemin çıkarıldığı varsayalım. Kalan gözlem grubu R , silinen gözlem

grubu ise D ile temsil edilsin. Böylece veri kümesinden $D(d)$ gözlem silindikten sonra, geride $R(n-d)$ gözlem kalmaktadır. Genelleştirmeyi kaybetmeden, silinen gözlemlerin X, Y, V matrislerinin son satırları olduğu varsayalım. Burada;

X : Bağımsız Değişkenler Matrisi

Y : Bağımlı Değişken Matrisi

V : Varyans-Kovaryans Matrisi olsun.

Daha açık biçimde

$$X = \begin{bmatrix} X_R \\ X_D \end{bmatrix}, Y = \begin{bmatrix} Y_R \\ Y_D \end{bmatrix}, V = \begin{bmatrix} V_R & 0 \\ 0 & V_D \end{bmatrix}$$

olarak yazılabilir.

$\hat{\beta}^{(-D)}$, silinen gözlem grubu D , veri kümesinden çıkarıldıktan sonra tahmin edilen katsayıvektörü olsun. Böylece lojistik regresyon modeli için tahmin edilen olasılık değerleri aşağıdaki biçimi alır.

$$\pi_i^{(-D)} = \frac{\exp(x_i^T \hat{\beta}^{(-D)})}{1 + \exp(x_i^T \hat{\beta}^{(-D)})}, i = 1, 2, \dots, n. \quad (5.1)$$

x_i : X matrisinin i 'inci satır vektörü

Gözlemler veri kümesinden çıkarıldıktan sonra artıklar yeniden aşağıdaki gibi tanımlanır.

$$\hat{e}_i^{(-D)} = Y_i - \hat{\pi}_i^{(-D)}, i = 1, 2, \dots, n. \quad (5.2)$$

Gözlemler veri kümesinden çıkarıldıktan sonra varyanslar ve kaldıraç değerleri de aşağıdaki gibi tanımlanır.

$$v_i^{(-D)} = \pi_i^{(-D)}(1 - \pi_i^{(-D)}) \quad (5.3)$$

$$h_{ii}^{(-D)} = \pi_i^{(-D)}(1 - \pi_i^{(-D)})x_i^T (X_R^T V_R X_R)^{-1} x_i \quad (5.4)$$

Hadi ve Simonoff (1993), doğrusal regresyonda, içsel(model tahmin edilirken kullanılan gözlemlerden elde edilen) ve dışsal(model tahmin edilirken kullanılmayan gözlemlerden elde edilen) olarak hesaplanan artıkların, benzer ölçekte ölçülemediğini göstermiş ve bu artıkları aynı ölçğe indirgemek için aşağıdaki gibi standartlaştırmışlardır.

$$t_i^* = \frac{y_i - x_i^T \tilde{\beta}^{(-D)}}{\hat{\sigma}_R \sqrt{1 - w_{ii}^{(-D)}}} \quad i \in R \quad (5.5)$$

$$t_i^* = \frac{y_i - x_i^T \tilde{\beta}^{(-D)}}{\hat{\sigma}_R \sqrt{1 + w_{ii}^{(-D)}}} \quad i \in D$$

(5.5) eşitliğinde $\hat{\sigma}_R$: dışlanmayan gözlemler üzerinden kurulan genel ölçek tahminidir.

R ve D kümeleri için (5.5)'te tanımlanan artıklar benzer ölçeklidir ve bu artık tipi regresyon tanı yöntemlerinde geniş uygulama alanı bulmuştur.

Yukarıdaki sonuçlar ve doğrusal regresyon yaklaşımına benzer bir yaklaşımkullanılarak, lojistik regresyonda çoklu aykırı gözlemlerin belirlenmesi için sunulan **i'inci** Genelleştirilmiş Standart Pearson Artığı (GSPR) aşağıdaki gibi tanımlanır.

$$r_{si}^{(-D)} = \frac{y_i - \hat{\pi}_i^{(-D)}}{\sqrt{v_i^{(-D)}(1 - h_{ii}^{(-D)})}} \quad i \in R$$

$$r_{si}^{(-D)} = \frac{y_i - \hat{\pi}_i^{(-D)}}{\sqrt{v_i^{(-D)}(1 + h_{ii}^{(-D)})}} \quad i \in D \quad (5.6)$$

Bir gözlemin ilgili GSPR değeri, mutlak değer olarak 3'ten büyükse o gözlemin aykırı gözlem olduğu söylenebilir.

GSPR terimi veri kümesindeki herhangi altküme (D) için geçerli olmasına rağmen, böyle bir grubun silinmesi, hem D hem de R kümesinin artıklarını belirleyeceği için çalışmanın en önemli kısmıdır. Çoklu aykırı gözlemlerin belirlenebilmesi için ilk başta veri kümesinden çıkarılacak olan gözlem grubu D_0 'ın tüm şüpheli aykırı gözlemleri içermesi istenir. Çünkü herhangi bir aykırı gözlem R kümesinde kalırsa, bütün GSPR kümesi başarısız olabilir. Bu anlamda şüpheli aykırı gözlemlerin belirlenmesi yöntemin başarılı olmasında oldukça önemlidir. Tüm şüpheli aykırı gözlemler belirlenirken uygun yöntemlerden birisi kullanılabilir.

D_0 kümesinin veri kümesinden çıkarılmasından sonra tüm veri kümesi için GSPR değerleri hesaplanır. Bir gözlemin aykırı gözlem olması için

$$\left| r_{si}^{(-D_0)} \right| > 3 \quad (5.7)$$

kuralı sağlanmalıdır.

Gözlemler, veri kümesinden çıkarılırken bazı gözlemler yanlışlıkla aykırı gözlem olarak belirlenebilir. GSPR değeri R ve D kümeleri için benzer ölçekte olduğundan, normal gözlemlerin aykırı olarak belirlenmesi büyük bir problem oluşturmaz. Yanlışlıkla çıkarılan gözlemler, tahmin altkümesine geri eklenir.

D_0 'ın tüm öğeleri (5.7) kuralını sağlıyorsa D_0 altkümesi çoklu aykırı gözlem olarak belirlenir. Aksi halde (5.7) kuralını sağlamayan gözlemler tekrar tahmin kümesine (R) eklenir. D kümesi yeniden gözden geçirilerek devam edilir ve tüm gözlemlerin kuralı sağladığı durum olarak ifade edilen D_m elde edilinceye kadar işleme devam edilir. Sonuçta D_m 'nin tüm öğeleri aykırı gözlem olarak adlandırılır ve çoklu aykırı gözlemler belirlenir.

Genelleştirilmiş Standart Pearson Artıkları (GSPR) yöntemine ilişkin algoritma aşağıdaki gibi özetlenebilir.

1. Tüm veriler için lojistik regresyon modeli elde edilir.
2. Elde edilen bu model üzerinden $\hat{\pi}_i$ değerleri hesaplanır.
3. Pregibon hat matrisi hesaplanır.
4. $V : v_i = \hat{\pi}_i(1 - \hat{\pi}_i)$ $n \times n$ köşegen matris elde edilir.
5. Pearson ve standart pearson artıkları hesaplanır.
6. Yukarıda verilen kaldıraç değerleri, pearson ve standart pearson artıkları, değerlendirilerek aykırı gözlem olması şüpheli görülen gözlemler veri kümesinden çıkarılır. Geriye kalan gözlem kümesi R_0 , çıkarılan gözlem kümesi ise D_0 ile gösterilir.
7. Veri kümesinde kalan gözlem kümesi R_0 kullanılarak, yeni lojistik regresyon modeli tahmin edilir.
8. Elde edilen yeni model kullanılarak tüm veri kümesi için $(R_0 + D_0)$ $\hat{\pi}_i$ değerleri tahmin edilir. Bu tahmin değerleri $\pi_i^{(-D)}$ ile gösterilir.
9. Şüpheli aykırı gözlemler veri kümesinden çıkarıldıktan sonra elde edilen artık değerleri, gözlem değerlerinden, $\pi_i^{(-D)}$ tahmin değerleri çıkarılarak bulunur.
10. $\pi_i^{(-D)}$ değerleri kullanılarak tüm veri kümesi için yeni $v_i^{(-D)}$ ve $h_i^{(-D)}$ değerleri hesaplanır.

11. i' inci genelleştirilmiş standart pearson artık değerleri (GSPR) $r_{si}^{(-D)}$ ile gösterilir. $|r_{si}^{(-D)}| > 3$ ise ilgili gözlemin aykırı gözlem olduğu söylenebilir. Şüpheli olarak ayrılan gözlem kümesi D_0 'ın tüm öğeleri bu koşulu sağlıyorsa, o grup çoklu aykırı gözlem olarak belirlenir.
12. D_0 kümesinde $|r_{si}^{(-D)}| < 3$ olan bir gözlem ya da gözlemler var ise bu gözlem ya da gözlemler R_0 kümesine eklenir. Elde edilen yeni gözlem kümeleri R_1 ve D_1 ile gösterilir ve yedinci maddeye dönerek döngü tekrar edilir. D kümesinin tüm elemanları $|r_{si}^{(-D)}| > 3$ koşulunu sağlayana kadar işlemlere devam edilir.

6. GSPR YÖNTEMİNİN VOLEYBOL VERİLERİNE UYGULANMASI

Bu kısımda Cengiz Akarçeşme (2010) tarafından yazılan “Elit Bayan Voleybolunda Maç Sonucunu Açıklayan Değişkenlerin Lojistik Regresyon Yöntemi ile Belirlenmesi ve Maç Kazanmaya Yönelik Olasılık Modelinin Tahmini” isimli doktora tezinde kullanılan “Etkinlik Modeli” verileri kullanılacaktır.

Etkinlik modelinde değişkenler maç boyunca elde edilen performans değerlerinden hesaplanarak doğrudan kullanılmıştır. Etkinlik modeline ilişkin değişkenler aşağıdaki gibi tanımlanmıştır.

6.1. Etkinlik Modelinin Değişkenleri

Sonuç:Maçın sonucunu tanımlayan ve kazanan takımın 1, kaybeden takımın 0 ile ifade edildiği sınıflandırma biçiminde tanımlı değişkendir. Etkinlik modelinin bağımlı değişkeni, “sonuç” değişkenidir.Etkinlik modeli tahmin edilirken kullanılan bağımsız değişkenler ise aşağıdaki gibi tanımlanmıştır.

LE (Libero Etkinliği): Libero oyuncusunun müsabaka boyunca servis karşılama esnasında yaptığı hata üzerinden değerlendirmeye alınmıştır. Libero oyuncusu savunma ağırlıklı bir oyuncu olduğundan ve hücum yapamadığından dolayı değerleri 0’den başlayıp hata arttıkça negatif yönlü bir seyir izlemektedir.

PE (Pasör Etkinliği): Pasörün müsabaka süresince katıldığı tüm oyun etkinliklerinde kazandığı ve kaybettiği sayıların farkını göstermektedir.

OOE (Orta Oyuncu Etkinliği): Orta oyuncuların müsabaka boyunca tüm oyun etkinliklerinde kazandığı ve kaybettiği sayıların farkını göstermektedir.

PÇE (Pasör Çapraz Etkinliği): Pasör çaprazı oyuncusunun müsabaka süresince tüm oyun boyunca kazandığı ve kaybettiği sayıların farkını ifade etmektedir.

Dört Numara Smaçör Etkinliği(N4E): Maç boyunca 4 numara smaçörlerin ikisinin birlikte tüm oyun boyunca kazandığı ve kaybettiği sayıların farkını ifade etmektedir.

6.2. Etkinlik Modelinde Çoklu Aykırı Gözlemlerin Belirlenmesi

Bu kısımda Hadi ve İmon tarafından önerilen genelleştirilmiş standart pearson artığı (GSPR) yöntemi kullanılarak “Etkinlik Modeli” verilerinde bulunan çoklu aykırı gözlemler belirlenip, veri kümesindeki maskeleye etkisi gösterilerek çoklu aykırı gözlemler incelenecektir. Yapılan çalışmada STATA 11.0, MATLAB R2008b programları kullanılmıştır.

“Elit Bayan Voleybolunda Maç Sonucunu Açıklayan Değişkenlerin Lojistik Regresyon Yöntemi ile Belirlenmesi ve Maç Kazanmaya Yönelik Olasılık Modelinin Tahmini” isimli doktora tezinin yazarı Cengiz Akarçeşme tarafından “Etkinlik Modeli” verilerinin tamamının bu tezde yayımına izin verilmediğinden verilerin bir kısmı veri yapısını göstermek için verilecektir.

Çizelge 6.1. Etkinlik Modeli Verileri

Göz. No	SONUÇ	LE	PE	OOE	PCE	N4E
1	1	-1	-2	12	10	12
2	0	-2	-7	-1	-5	3
3	1	-2	1	2	10	10
4	0	-5	1	-2	1	-4
5	0	-6	-1	9	4	-1
6	1	-2	-2	15	-1	15
7	1	0	1	3	4	12
8	0	-2	-1	-2	9	12
9	1	0	2	7	3	13
10	0	-3	-4	-2	5	-1
11	1	0	-1	2	6	22
12	0	-3	0	-5	10	-1
13	1	-3	-2	13	8	2
14	0	-2	-2	3	1	6
15	1	-3	2	9	1	6
16	0	0	3	7	4	0
17	0	-5	-1	-1	1	-5
18	1	0	-5	4	10	15
19	0	-5	-2	-1	17	12
⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮
147	1	-2	1	14	10	20
148	0	-1	1	3	-3	-4
149	0	0	1	15	-2	15
150	1	-1	-5	0	4	14
151	1	0	-2	5	-1	14
152	0	-3	-2	2	2	0
153	0	0	3	-9	-1	9
154	1	-1	1	10	23	1
155	0	0	4	3	6	-4
156	1	0	-1	-3	3	15
157	0	-1	1	-4	3	3
158	1	0	-2	10	8	-2
159	0	0	-1	-6	6	17
160	1	-1	-6	7	11	14
161	1	-2	-2	10	12	7
162	0	-5	-1	-2	-1	6
163	1	-1	2	2	-2	10
164	0	-8	0	-5	11	2

Etkinlik Modelindeki tüm gözlemler ile logit model tahmin edildiğinde aşağıdaki model elde edilmektedir.

```
logit sonuc le pe ooe pce n4e
```

```
Iteration 0: log likelihood = -113.67614
Iteration 1: log likelihood = -59.484132
Iteration 2: log likelihood = -59.119836
Iteration 3: log likelihood = -59.119066
Iteration 4: log likelihood = -59.119066
```

Logistic regression

```
Number of obs   =      164
LR chi2(5)      =     109.11
Prob > chi2     =      0.0000
Pseudo R2      =      0.4799
```

Log likelihood = -59.119066

sonuc	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
le	.3849534	.1540677	2.50	0.012	.0829862	.6869206
pe	.1521623	.078582	1.94	0.053	-.0018557	.3061802
ooe	.1524323	.0392186	3.89	0.000	.0755653	.2292992
pce	.1816252	.0455479	3.99	0.000	.092353	.2708974
n4e	.2192384	.0379209	5.78	0.000	.1449148	.2935621
_cons	-2.08733	.5678779	-3.68	0.000	-3.20035	-.9743096

Şekil6.1. Etkinlik Modeli Verileri ile Tahmin Edilen Logit Model

Genelleştirilmiş standart pearson artığı (GSPR) yönteminin ilk aşaması olarak şüpheli aykırı gözlemlerin belirlenmesi gerekmektedir. Şüpheli aykırı gözlemler belirlenirken grafik yöntemi, pearson ve standart pearson artıkları kullanılacaktır.

Modelden elde edilen  olasılık ile pearson ve standart pearson artık değerleri aşağıdaki gibi elde edilir.

Çizelge 6.2 Etkinlik Modeli olasılık, Pearson ve Standart Pearson artık değerleri

Göz. No	$\hat{\pi}$	Pearson Artıkları	Standart Pearson Artıkları	Göz. No	$\hat{\pi}$	Pearson Artıkları	Standart Pearson Artıkları
1	0,97068	0,17380	0,17474	50	0,05803	-0,24820	-0,25174
2	0,01306	-0,11503	-0,11564	51	0,99697	0,05513	0,05521
3	0,83319	0,44744	0,45233	52	0,00249	-0,04996	-0,05003
4	0,00769	-0,08803	-0,08829	53	0,75661	0,56717	0,58277
5	0,06476	-0,26314	-0,26873	54	0,37226	-0,77008	-0,78157
6	0,90308	0,32760	0,33275	55	0,68722	0,67464	0,70117
7	0,86755	0,39073	0,39524	56	0,14440	-0,41082	-0,42043
8	0,72136	-1,60899	-1,63224	57	0,84215	0,43294	0,44348
9	0,93578	0,26197	0,26426	58	0,00324	-0,05701	-0,05712
10	0,03027	-0,17668	-0,17771	59	0,53864	0,92549	0,94348
11	0,98163	0,13680	0,13742	60	0,10214	-0,33728	-0,34274
12	0,08261	-0,30008	-0,30356	61	0,06665	-0,26723	-0,27144
13	0,58093	0,84934	0,87789	62	0,83615	0,44267	0,44900
14	0,23020	-0,54684	-0,55270	63	0,99538	0,06813	0,06825
15	0,48278	1,03505	1,06483	64	0,94842	-4,28805	-4,36196
16	0,54058	-1,08474	-1,12560	65	0,71005	-1,56489	-1,61251
17	0,00532	-0,07313	-0,07329	66	0,81444	0,47732	0,48988
18	0,94616	0,23854	0,24122	67	0,04783	-0,22413	-0,22600
19	0,77724	-1,86792	-1,96543	68	0,98973	0,10187	0,10220
20	0,90705	0,32012	0,32804	69	0,30774	-0,66674	-0,68137
21	0,26060	1,68443	1,72322	70	0,45984	1,08382	1,14780
22	0,00439	-0,06640	-0,06657	71	0,48779	1,02473	1,04972
23	0,50867	-1,01749	-1,03200	72	0,06775	-0,26958	-0,27279
24	0,93940	0,25399	0,25645	73	0,41814	-0,84772	-0,87126
25	0,01817	-0,13604	-0,13674	74	0,44132	1,12514	1,15024
26	0,96515	0,19002	0,19123	75	0,69926	0,65581	0,66696
27	0,00523	-0,07251	-0,07271	76	0,83454	-2,24583	-2,28602
28	0,97592	0,15708	0,15849	77	0,00116	-0,03408	-0,03410
29	0,23188	1,82005	1,88720	78	0,99740	0,05106	0,05112
30	0,00413	-0,06440	-0,06453	79	0,22520	-0,53912	-0,55201
31	0,80255	0,49601	0,51138	80	0,73865	0,59483	0,60362
32	0,01781	-0,13466	-0,13526	81	0,90083	0,33179	0,33673
33	0,55644	0,89283	0,92038	82	0,11309	-0,35709	-0,36275
34	0,03646	-0,19452	-0,19567	83	0,24211	-0,56520	-0,57816
35	0,95699	0,21200	0,21553	84	0,47715	1,04679	1,06446
36	0,72229	-1,61272	-1,64006	85	0,76750	-1,81689	-1,93297
37	0,08338	-0,30160	-0,31138	86	0,88161	0,36645	0,37610
38	0,21884	1,88932	2,01081	87	0,95543	0,21598	0,22023
39	0,99623	0,06152	0,06162	88	0,20571	-0,50891	-0,51976
40	0,73535	-1,66691	-1,75770	89	0,68427	0,67927	0,71652
41	0,68496	0,67819	0,73253	90	0,34197	-0,72089	-0,75458
42	0,42981	-0,86822	-0,88944	91	0,45500	-0,91371	-0,94893
43	0,07073	-0,27589	-0,28032	92	0,91222	0,31020	0,31681
44	0,87019	0,38623	0,39160	93	0,92392	0,28696	0,28984
45	0,02593	-0,16316	-0,16495	94	0,04076	-0,20614	-0,20776
46	0,88331	0,36346	0,36693	95	0,97083	0,17334	0,17506
47	0,08199	-0,29885	-0,30215	96	0,02187	-0,14953	-0,15037
48	0,94304	0,24576	0,24759	97	0,88028	0,36878	0,37737
49	0,35477	1,34860	1,36939	98	0,13739	-0,39909	-0,40892

Çizelge 6.2 (Devam) Etkinlik Modeli olasılık, Pearson ve standart Pearson artık değerleri

Göz. No	$\hat{\pi}$	Pearsonres	stdpersres	Göz. No	$\hat{\pi}$	Pearsonres	stdpersres
99	0,47292	-0,94723	-0,97501	148	0,03610	-0,19353	-0,19522
100	0,98594	0,11942	0,11989	149	0,96362	-5,14661	-5,19206
101	0,28630	-0,63336	-0,65126	150	0,63709	0,75474	0,77726
102	0,95081	0,22745	0,22949	151	0,77872	0,53307	0,54392
103	0,86236	0,39951	0,40482	152	0,05323	-0,23711	-0,23880
104	0,77578	-1,86008	-2,02709	153	0,22949	-0,54575	-0,57020
105	0,90103	0,33142	0,33505	154	0,97342	0,16524	0,16713
106	0,05952	-0,25157	-0,25372	155	0,30819	-0,66745	-0,69707
107	0,74878	0,57923	0,58831	156	0,75707	0,56646	0,58037
108	0,09278	-0,31979	-0,32371	157	0,15094	-0,42163	-0,42778
109	0,06269	-0,25862	-0,26195	158	0,53673	0,92905	0,95995
110	0,96718	0,18421	0,18607	159	0,84060	-2,29642	-2,35566
111	0,10601	-0,34436	-0,34766	160	0,93986	0,25296	0,25607
112	0,30083	1,52451	1,60022	161	0,88864	0,35400	0,35854
113	0,69298	0,66561	0,68119	162	0,03438	-0,18869	-0,19029
114	0,00057	-0,02388	-0,02389	163	0,49151	1,01713	1,04130
115	0,20976	-0,51521	-0,52307	164	0,02952	-0,17441	-0,17725
116	0,06727	3,72364	3,76083				
117	0,84287	0,43177	0,43910				
118	0,33813	-0,71475	-0,73471				
119	0,70394	0,64852	0,66134				
120	0,11707	-0,36413	-0,37303				
121	0,98966	0,10222	0,10250				
122	0,06541	-0,26455	-0,26656				
123	0,02335	-0,15462	-0,15660				
124	0,45367	1,09738	1,14330				
125	0,94117	0,25001	0,25169				
126	0,00062	-0,02491	-0,02492				
127	0,92975	0,27488	0,27686				
128	0,06455	-0,26269	-0,26487				
129	0,78513	0,52314	0,53080				
130	0,03165	-0,18079	-0,18300				
131	0,61111	0,79773	0,82199				
132	0,26518	-0,60073	-0,65030				
133	0,83685	0,44154	0,45834				
134	0,00059	-0,02430	-0,02431				
135	0,88024	0,36885	0,37430				
136	0,11791	-0,36561	-0,37168				
137	0,96499	0,19047	0,19180				
138	0,00831	-0,09154	-0,09208				
139	0,09892	-0,33133	-0,33741				
140	0,95170	0,22528	0,22709				
141	0,02204	-0,15012	-0,15118				
142	0,58772	0,83755	0,85245				
143	0,50368	0,99267	1,01091				
144	0,12838	-0,38378	-0,39220				
145	0,85566	-2,43477	-2,63173				
146	0,92065	0,29358	0,29609				
147	0,99642	0,05994	0,06003				

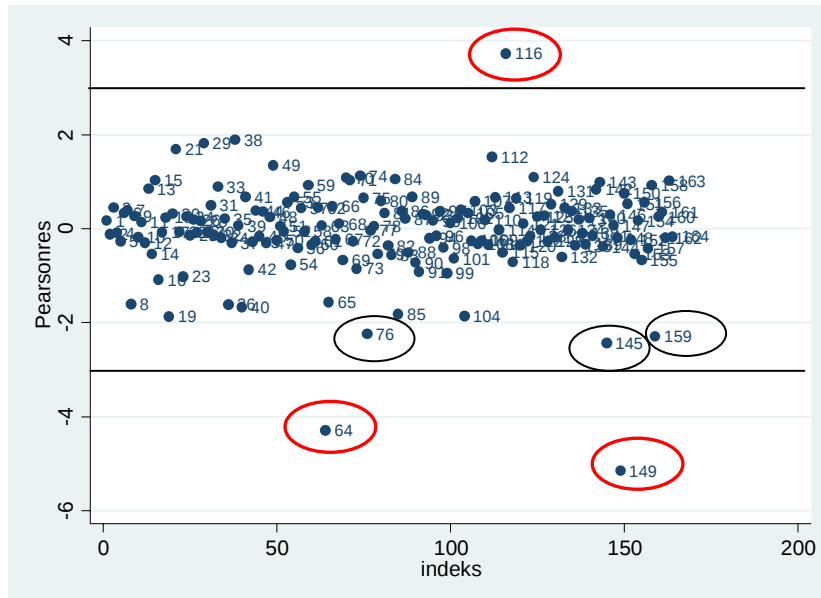
Aykırı gözlemleri belirlemede kullanılan Pearson ve Standart Pearson artıkları yöntemlerine göre, ilgili gözlemlerin Pearson ve Standart Pearson artık değerleri mutlak değer olarak 3'ten büyükse o gözlem aykırı gözlem olarak kabul edilir.

Çizelge 6.3. Aykırı Gözlemlerin Pearson ve Standart Pearson Artık Değerleri

Gözlem No	$\hat{\pi}$	Pearsonres	stdpersres
64	0,94842	-4,28805	-4,36196
116	0,06727	3,72364	3,76083
149	0,96362	-5,14661	-5,19206

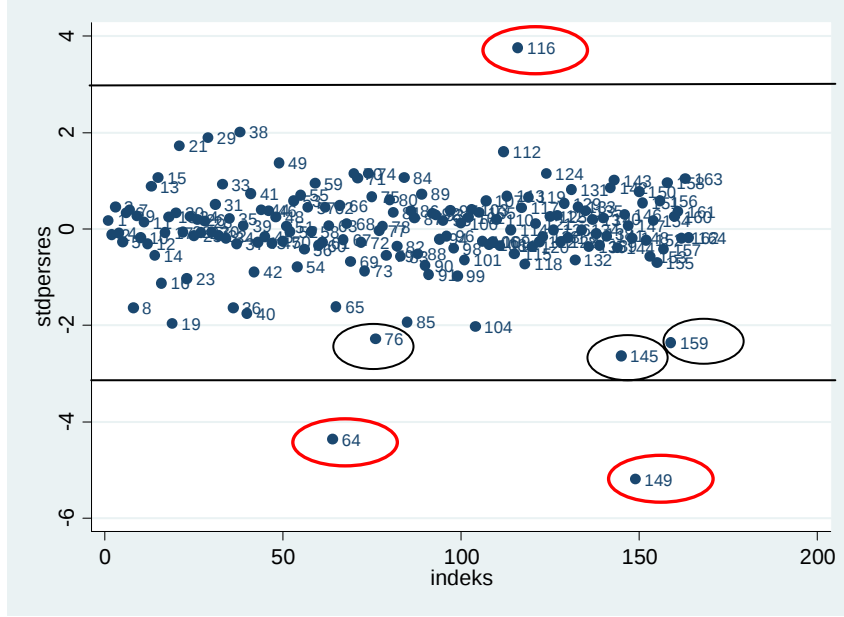
64,116 ve 149 numaralı gözlemlerin artık değerleri 3'ten büyüktür ve bu gözlemler aykırı gözlem olarak kabul edilir.

Regresyon analizinde en çok kullanılan yöntem olan grafik yöntemi de kullanılarak elde edilen sonuçların daha açık bir şekilde gösterilmesi yararlı olur. Grafik yönteminde, elde edilen Pearson ve Standart Pearson artıklarının gözlem numaralarına göre grafikleri çizilir. Grafikler aşağıda verilmektedir.



Şekil 6.2. Pearson artıklarının gözlem numaralarına göre grafiği

Grafikten de görüldüğü gibi 64, 116 ve 149 numaralı gözlemlerin Pearson artıkları sınırının dışına düşmektedir. Bu da ilgili gözlemlerin aykırı gözlem olduğuna işaret etmektedir. Aynı şekilde Standart Pearson artıklarının grafiği çizildiğinde aşağıdaki gibi elde edilir.



Şekil 6.3. Standart Pearson artıklarının gözlem numaralarına göre grafiği

64, 116 ve 149 numaralı gözlemlerin Standart Pearson artıkları da sınırının dışına düşmektedir. Pearson artık değerleri ile aynı sonuç elde edilmiş ve ilgili gözlemlerin aykırı gözlem olduğu belirlenmiştir.

Yukarıdaki grafik incelendiğinde artık değerleri değerlerinin dışına düşen 3 gözlem olmasına rağmen, artık değerleri bu değerlere yakın başka gözlemler de görülmektedir. Bu da tüm aykırı gözlemlerin belirlenip belirlenemediği sorusunu akla getirir. Elde edilen artık değerleri ve grafikler incelendiğinde 76, 145 ve 159 numaralı gözlemlerin de incelenmesi gerektiği düşünülebilir. Bu gözlemlere ait değerler aşağıda verilmektedir.

Çizelge 6.4. Şüpheli görülen gözlemlerin pearson ve standart pearson artık değerleri

Gözlem No	$\hat{\pi}$	Pearsonres	stdpersres
76	0,83454	-2,24583	-2,28602
145	0,85566	-2,43477	-2,63173
159	0,84060	-2,29642	-2,35566

Pearson ve Standart Pearson artık değerleri kritik değer olan -3'e yakındır ve grafiklerde de diğer gözlemlerden uzağa düşmektedirler. Bu durumda bu gözlemlerin aykırı gözlem olduğu halde aykırı gözlem olarak belirlenemediği yani maskeleye etkisine uğradığı düşünülebilir.

Lojistik regresyonda çoklu aykırı gözlemleri belirlemede kullanılmak üzere önerilen Genelleştirilmiş Standart Pearson Artığı yönteminin başarılı olması için şüpheli tüm aykırı gözlemlerin alınıp yonteme devam edilmesi gerektiği belirtilmişti. Etkinlik modeli verilerinden elde edilen artık değerlerine göre veri kümesinden ilk olarak çıkarılacak olan D_0 gözlem grubu 64, 76, 116, 145, 149 ve 159 numaralı gözlemlerden oluşacaktır.

Yöntemin ikinci aşamasında D_0 gözlem grubu veri kümesinden çıkarıldıktan sonra kalan gözlemler üzerinden tekrar lojistik regresyon modeli tahmin edilecektir. Geriye kalan gözlem grubu, R_0 , 158 gözlem içermektedir ve bu gözlemler ile elde edilen model aşağıda verilmektedir.

```
logit sonuc le pe ooe pce n4e
```

```
Iteration 0: log likelihood = -109.46662
Iteration 1: log likelihood = -40.725246
Iteration 2: log likelihood = -40.571336
Iteration 3: log likelihood = -40.570699
Iteration 4: log likelihood = -40.570699
```

```
Logistic regression
```

```
Number of obs   =      158
LR chi2(5)      =     137.79
Prob > chi2     =      0.0000
Pseudo R2      =      0.6294
```

```
Log likelihood = -40.570699
```

sonuc	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
le	.54054	.1987133	2.72	0.007	.151069	.9300109
pe	.2399637	.1011664	2.37	0.018	.0416812	.4382463
ooe	.2782316	.0642028	4.33	0.000	.1523965	.4040667
pce	.2172865	.0603675	3.60	0.000	.0989683	.3356047
n4e	.3251302	.0579317	5.61	0.000	.2115861	.4386743
_cons	-2.788172	.7282504	-3.83	0.000	-4.215517	-1.360828

Şekil 6.4. R_0 gözlem grubu ile elde edilen logit model tahmini

Elde edilen yeni modelden, tüm veri kümesi için $(R_0 + D_0 \text{ kümesi})$ olasılık değerleri $\hat{\pi}_i$ değerleri tahmin edilir. Bu olasılık tahmin değerleri $\pi_i^{(-D)}$ ile gösterilir. Bu olasılık değerleri bulunduktan sonra gözlem değerlerinden olasılık değerleri çıkarılarak yeni artık değerleri bulunur. t' inci Genelleştirilmiş Standart Pearson Artığı, lojistik regresyon için (5.6)'da verilen eşitlik ile elde edilir.

$|r_{si}^{(-D)}| > 3$ ise o gözlemin aykırı gözlem olduğu yorumu yapılır. Şüpheli olarak ayrılan gözlemler grubu D_0 , bu koşulu sağlıyorsa o grup çoklu aykırı gözlem olarak belirlenir. D_0 gözlem grubuna ait $r_{si}^{(-D)}$ değerlerine ait sonuçlar aşağıda verilmektedir.

Çizelge 6.5. D_0 gözlem grubuna ait Pearson, Standart Pearson ve GSPR değerleri

Göz. No	p	Pearsonres	Stdpersres	GSPR
64	0,94842	-4,28805	-4,36196	-14,55779
76	0,83454	-2,24583	-2,28602	-2,63118
116	0,06727	3,72364	3,76083	6,74056
145	0,85566	-2,43477	-2,63173	-6,59721
149	0,96362	-5,14661	-5,19206	-19,97460
159	0,84060	-2,29642	-2,35566	-2,81033

Tablodan görüldüğü gibi $|r_{si}^{(-D)}| > 3$ şartını 64, 116, 145 ve 149 numaralı gözlemler sağlamaktadır. 76 ve 159 numaralı gözlemler, $|r_{si}^{(-D)}| > 3$ şartını sağlamadığı için yöntem gereğince R_0 gözlem grubuna eklenirler.

Şu anda çoklu aykırı gözlem olarak belirlenecek grubun gözlem sayısı 4'e inmiştir. Yani döngünün ikinci adımında R_1 gözlem grubu 160 gözlem, D_1 gözlem grubu 4 gözlem içerecektir. Tüm veri kümesinden 64, 116, 145 ve 149 numaralı gözlemler çıkarıldığında ve yeni duruma GSPR adımları uygulandığında aşağıda verilen sonuçlara ulaşılır.

logit sonu le pe ooe pce n4e

Iteration 0: log likelihood = -110.89105
Iteration 1: log likelihood = -45.771808
Iteration 2: log likelihood = -45.344495
Iteration 3: log likelihood = -45.331824
Iteration 4: log likelihood = -45.331815
Iteration 5: log likelihood = -45.331815

Logistic regression

Log likelihood = -45.331815

Number of obs = 160
LR chi2(5) = 131.12
Prob > chi2 = 0.0000
Pseudo R2 = 0.5912

sonu	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]		
le	.4734877	.1811736	2.61	0.009	.1183939	.8285814	
pe	.204303	.0931315	2.19	0.028	.0217687	.3868373	
ooe	.2697762	.0595985	4.53	0.000	.1529653	.386587	
pce	.2090279	.0565411	3.70	0.000	.0982095	.3198463	
n4e	.283137	.0496602	5.70	0.000	.1858048	.3804691	
_cons	-2.705135	.6908862	-3.92	0.000	-4.059247	-1.351023	

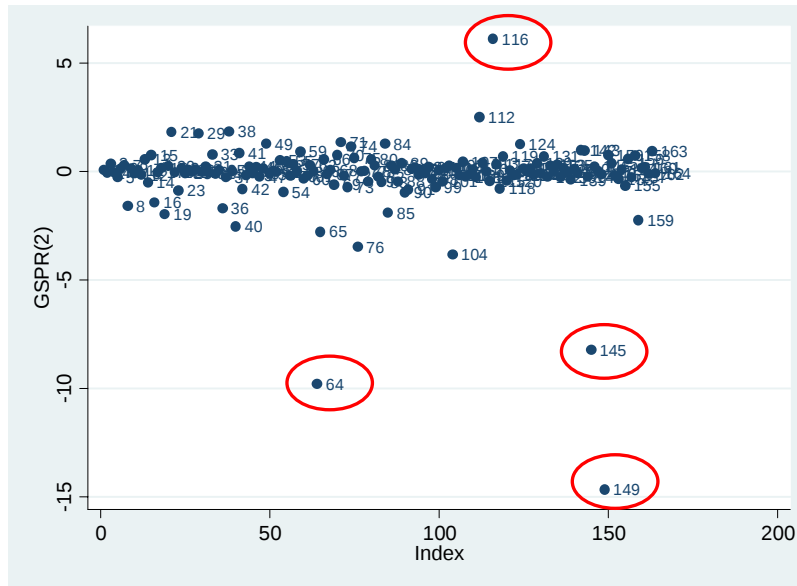
Şekil6.5. Kalan gözlem grubu R_1 ile elde edilen logit model tahmini

Elde edilen yeni modelden, tüm veri kümesi için $(R_1 + D_1 \text{ kümesi})$ olasılık değerleri $\pi_i^{(-D)}$ değerleri tahmin edilir. Bu olasılık değerleri bulunduktan sonra gözlem değerlerinden olasılık değerleri çıkarılarak yeni artık değerleri bulunur. Genelleştirilmiş Standart Pearson Artığı, (5.6) eşitliği ile elde edilir. Döngünün ikinci adımı sonucunda, D_1 gözlem grubuna ilişkin sonuçlar aşağıda verilmektedir.

Çizelge 6.6. D_1 gözlem grubuna ait Pearson, Standart Pearson ve GSPR değerleri

Göz. No	Pearsonres	stdpersres	GSPR(2)
64	-4,28805	-4,36196	16,85599
116	3,72364	3,76083	27,55073
145	-2,43477	-2,63173	72,74711
149	-5,14661	-5,19206	32,07832

Tablodan görüldüğü gibi $|r_{si}^{(-D)}| > 3$ şartını D_1 gözlem grubundaki 64, 116, 145 ve 149 numaralı gözlemlerin hepsi sağlamaktadır. Dolayısıyla döngü sona ermiş, veri kümesindeki aykırı gözlemler belirlenmiştir. Elde edilen yeni GSPR(2) değerlerinin gözlem numaralarına göre grafiği çizdirildiğinde aşağıdaki grafik elde edilir.



Şekil 6.6. GSPR(2) değerlerinin gözlem numaralarına göre grafiği

Grafikten de açık bir şekilde görüldüğü gibi 64, 116, 145, 149 numaralı gözlemler aykırı gözlem olarak belirlenmektedir. Burada dikkat edilmesi gereken nokta Pearson ve Standart Pearson artıkları kullanılarak yapılan aykırı gözlem incelemesinde 64, 116 ve 149 numaralı gözlemlerin aykırı gözlem olarak belirlenebilmiş olmasıdır. Ancak GSPR kullanılarak yapılan aykırı gözlem incelemesinde ise 64, 116, 145 ve 149 numaralı gözlemler aykırı gözlem olarak belirlenmiştir. Bu noktada ilk

durumda 145 numaralı gözlemin regresyon modeli üzerindeki etkisinin maskelendiğini söyleyebiliriz. Verilen yöntem maskeleye etkisini kaldırarak tüm aykırı gözlemleri belirleyebilmiştir.

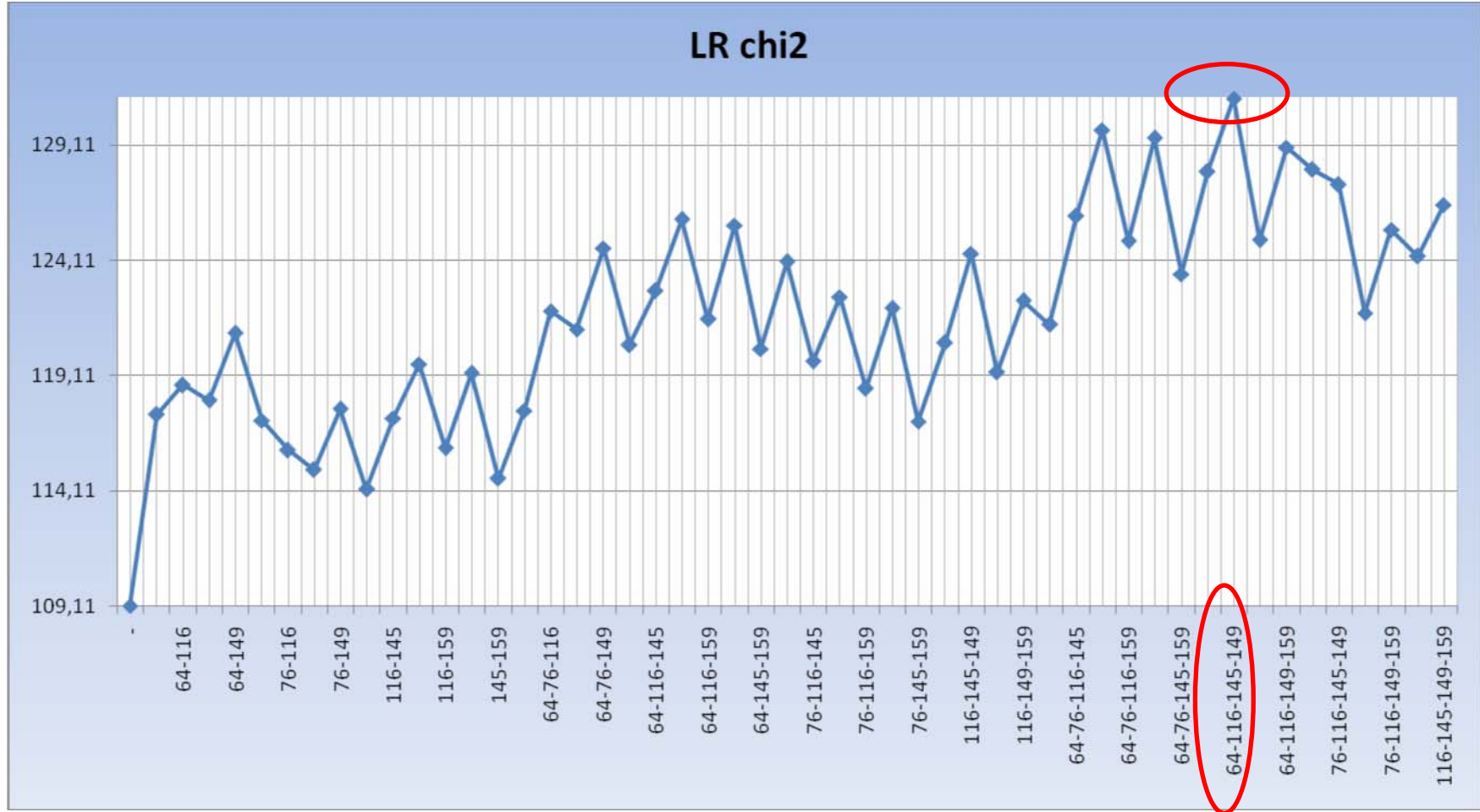
GSPR yöntemi ile çoklu aykırı gözlemler belirlenmiş ve veri kümesinden çıkarıldıktan sonra yeni model elde edilmiştir. Ancak elde edilen model en iyi model midir? Yöntem en iyi modelin elde edilmesini sağlamış mıdır? Bu durumları inceleyebilmek için ilk başta şüpheli gözlem olarak veri kümesinden çıkarılan 64, 76, 116, 145, 149 ve 159 numaralı gözlemlerin ikili, üçlü ve dörtlü alt kümeleri belirlenmiş ve bu altkümeler veri kümesinden çıkarıldıktan sonra modellere ilişkin değerler elde edilmiştir. Elde edilen değerlere ilişkin detaylı tablo aşağıda verilmektedir.

Çizelge 6.7. Veri kümesinden aykırı gözlem kombinasyonları çıkarıldıktan sonra modellere ilişkin elde edilen değerler

Veri Kümesinden Çıkarılan Gözlemler	Parametre Tahminleri					Standart Hatalar					P> z					Pearson chi2	LR chi2
	le	pe	ooe	pce	n4e	le	pe	ooe	pce	n4e	le	pe	ooe	pce	n4e		
-	0,38495	0,15216	0,15246	0,18163	0,21924	0,15407	0,07858	0,03922	0,04555	0,03792	0,012	0,053	0	0	0	145,65	109,11
64-76	0,35569	0,17563	0,17038	0,17276	0,25773	0,16047	0,08384	0,04254	0,04769	0,04383	0,027	0,036	0	0	0	156,14	117,42
64-116	0,40797	0,1836	0,1791	0,18812	0,25504	0,16591	0,08409	0,04358	0,04927	0,04373	0,014	0,029	0	0	0	144,41	118,70
64-145	0,3627	0,14185	0,20644	0,18781	0,24271	0,15912	0,08297	0,04842	0,04963	0,04174	0,023	0,087	0	0	0	162,44	118,02
64-149	0,43013	0,17985	0,18876	0,17317	0,26253	0,16789	0,08537	0,04547	0,04877	0,04485	0,01	0,035	0	0	0	120,95	120,95
64-159	0,42855	0,15215	0,15492	0,18064	0,25807	0,16632	0,08207	0,04083	0,04778	0,04398	0,01	0,064	0	0	0	146,26	117,17
76-116	0,40191	0,19677	0,17183	0,19449	0,24004	0,16249	0,08372	0,04223	0,04885	0,04155	0,013	0,019	0	0	0	150,04	115,88
76-145	0,35694	0,15557	0,19748	0,19357	0,22821	0,15546	0,08225	0,04679	0,04903	0,03979	0,022	0,059	0	0	0	167,55	115,01
76-149	0,41967	0,19199	0,18	0,1799	0,24579	0,16358	0,08453	0,04381	0,0483	0,04239	0,01	0,023	0	0	0	145,76	117,65
76-159	0,42082	0,16581	0,1487	0,18661	0,24242	0,16266	0,0815	0,03975	0,04735	0,04169	0,01	0,042	0	0	0	156,29	114,18
116-145	0,40676	0,1657	0,20917	0,20987	0,22876	0,16182	0,08302	0,04828	0,05091	0,04039	0,012	0,046	0	0	0	151,97	117,26
116-149	0,47636	0,20213	0,18998	0,19675	0,24569	0,17073	0,08516	0,04506	0,050233	0,04281	0,005	0,018	0	0	0	123,72	119,60
116-159	0,47537	0,17684	0,15776	0,20222	0,24194	0,1698	0,08229	0,04077	0,04913	0,04199	0,005	0,032	0	0	0	140,84	115,97
145-149	0,4267	0,15895	0,22003	0,19669	0,23401	0,16326	0,08389	0,05004	0,050597	0,04106	0,009	0,058	0	0	0	136,11	119,20
145-159	0,4221	0,13512	0,18131	0,19926	0,22855	0,1615	0,08076	0,04535	0,0489	0,03998	0,009	0,094	0	0	0	155,97	114,65
149-159	0,49372	0,16952	0,16451	0,1881	0,24714	0,17065	0,08277	0,04204	0,04852	0,04272	0,004	0,041	0	0	0	134,96	117,58
64-76-116	0,39129	0,20641	0,18719	0,18706	0,27045	0,16815	0,08734	0,04522	0,05045	0,0464	0,02	0,018	0	0	0	158,54	121,92
64-76-145	0,34586	0,16372	0,21539	0,18699	0,25701	0,16139	0,08608	0,05037	0,05089	0,04407	0,032	0,057	0	0	0	180,49	121,12
64-76-149	0,41668	0,2064	0,19895	0,17139	0,28057	0,17107	0,08939	0,04763	0,04996	0,04803	0,015	0,021	0	0	0	123,96	124,64
64-76-159	0,41365	0,17325	0,16171	0,17951	0,27397	0,16871	0,08503	0,04213	0,04888	0,04676	0,014	0,042	0	0	0	158,8	120,44
64-116-145	0,39825	0,17123	0,22692	0,20399	0,25575	0,16722	0,0863	0,0518	0,05274	0,04438	0,017	0,047	0	0	0	168,24	122,78
64-116-149	0,47531	0,21303	0,20853	0,18886	0,27773	0,17731	0,08929	0,04885	0,05189	0,04797	0,007	0,017	0	0	0	107,67	125,88
64-116-159	0,46957	0,1815	0,17047	0,19546	0,27064	0,17504	0,0852	0,04315	0,05061	0,04654	0,007	0,033	0	0	0	146,42	121,55
64-145-149	0,42517	0,1684	0,24205	0,18992	0,26477	0,17028	0,08843	0,0545	0,05259	0,04571	0,013	0,057	0	0	0	117,26	125,63
64-145-159	0,41598	0,13836	0,1959	0,19326	0,25551	0,16706	0,08386	0,04822	0,05059	0,04398	0,013	0,099	0	0	0	164,01	120,27
64-149-159	0,49591	0,17772	0,17993	0,18048	0,27988	0,1775	0,0867	0,04503	0,050098	0,04803	0,005	0,04	0	0	0	116,63	124,04
76-116-145	0,38986	0,18479	0,21648	0,20968	0,24005	0,16313	0,0856	0,04982	0,05203	0,04216	0,017	0,031	0	0	0	168,38	119,73

Çizelge 6.7.(Devam) Veri kümesinden aykırı gözlem kombinasyonları çıkarıldıktan sonra modellere ilişkin elde edilen değerler

Veri Kümesinden Çıkarılan Gözlemler	Parametre Tahminleri					Standart Hatalar					P> z					Pearson chi2	LR chi2
	le	pe	ooe	pce	n4e	le	pe	ooe	pce	n4e	le	pe	ooe	pce	n4e		
76-116-149	0,46136	0,22464	0,19816	0,19568	0,25941	0,17249	0,08841	0,04677	0,05137	0,04512	0,007	0,011	0	0	0	134,71	122,52
76-116-159	0,45956	0,19506	0,16332	0,20151	0,25395	0,1711	0,08471	0,04183	0,05012	0,04399	0,007	0,021	0	0	0	153,64	118,55
76-145-149	0,41166	0,18085	0,22927	0,19598	0,24705	0,1651	0,08708	0,05204	0,05182	0,04316	0,013	0,038	0	0	0	149,10	122,04
76-145-159	0,4067	0,15258	0,18723	0,19873	0,23975	0,16286	0,08302	0,04661	0,04991	0,04175	0,013	0,066	0	0	0	170,56	117,12
76-149-159	0,48038	0,19024	0,17131	0,18685	0,26107	0,1725	0,08569	0,04339	0,04956	0,04511	0,005	0,026	0	0	0	148,25	120,53
116-145-149	0,46928	0,19048	0,24267	0,21478	0,24833	0,17279	0,08765	0,05387	0,05409	0,04402	0,007	0,03	0	0	0	121,48	124,41
116-145-159	0,45988	0,16299	0,1989	0,21547	0,24015	0,1701	0,08377	0,04806	0,05189	0,04234	0,007	0,052	0	0	0	154,58	119,25
116-149-159	0,54202	0,20088	0,18134	0,20463	0,26083	0,18086	0,0863	0,04462	0,05171	0,04551	0,003	0,02	0	0	0	123,25	122,36
145-149-159	0,48242	0,15583	0,20966	0,20242	0,2463	0,17191	0,08476	0,04981	0,05163	0,04318	0,005	0,066	0	0	0	137,02	121,33
64-76-116-145	0,38192	0,19621	0,23796	0,20384	0,27169	0,16983	0,09007	0,05428	0,05427	0,0471	0,025	0,029	0	0	0	196,91	126,03
64-76-116-149	0,46375	0,2438	0,22124	0,18761	0,29815	0,18097	0,09419	0,05169	0,05333	0,05177	0,01	0,01	0	0	0	110,77	129,77
64-76-116-159	0,45531	0,2052	0,17863	0,19479	0,28805	0,17758	0,08868	0,04479	0,05192	0,04972	0,01	0,021	0	0	0	164,66	124,95
64-76-145-149	0,4124	0,1983	0,25656	0,18902	0,28379	0,17403	0,09331	0,05788	0,05417	0,04905	0,018	0,034	0	0	0	120,76	129,40
64-76-145-159	0,4012	0,16073	0,20458	0,19269	0,27127	0,16967	0,0871	0,05016	0,05195	0,04668	0,018	0,065	0	0	0	184,91	123,51
64-76-149-159	0,48627	0,20554	0,19027	0,17915	0,30055	0,18123	0,09103	0,0472	0,05144	0,05186	0,007	0,024	0	0	0	121,06	127,96
64-116-145-149	0,47349	0,2043	0,26978	0,20903	0,28314	0,18117	0,09313	0,0596	0,05654	0,04966	0,009	0,028	0	0	0	98,94	131,12
64-116-145-159	0,45681	0,16844	0,21624	0,21016	0,26944	0,17636	0,08725	0,05157	0,05392	0,04688	0,01	0,054	0	0	0	172,85	125,00
64-116-149-159	0,54907	0,21243	0,19994	0,19747	0,29688	0,18855	0,09082	0,04845	0,05361	0,05164	0,004	0,019	0	0	0	103,16	129,02
64-145-149-159	0,48813	0,16531	0,23126	0,19624	0,27994	0,18016	0,0896	0,0543	0,05383	0,04852	0,007	0,065	0	0	0	114,88	128,04
76-116-145-149	0,45481	0,21568	0,25417	0,21477	0,26311	0,17501	0,09156	0,05645	0,0556	0,04653	0,009	0,018	0	0	0	136,77	127,40
76-116-145-159	0,44439	0,18241	0,20605	0,21549	0,25234	0,17164	0,08643	0,0496	0,05311	0,04434	0,01	0,035	0	0	0	173,64	121,81
76-116-149-159	0,52928	0,22422	0,18959	0,20385	0,27617	0,18293	0,08976	0,04634	0,05297	0,04824	0,004	0,012	0	0	0	138,09	125,43
76-145-149-159	0,4694	0,17821	0,21873	0,20193	0,2605	0,17409	0,08808	0,05182	0,05295	0,04556	0,007	0,043	0	0	0	153,12	124,28
116-145-149-159	0,53131	0,18817	0,23217	0,22145	0,26174	0,1829	0,08861	0,05363	0,05541	0,04646	0,004	0,034	0	0	0	123,61	126,52



Şekil 6.7. Elde edilen tüm modellerin LR Chi2 Değerleri

Yukarıdaki tabloda verilen *LR chi2* değeri, modelin uyum iyiliğinin derecesini göstermektedir. Tüm kombinasyonlar için bu değerler hesaplanmış ve uyum iyiliği en yüksek olan model, Genelleştirilmiş Standart Pearson Artığı (GSPR) yöntemi ile elde edilen; 64, 116, 145, 149 numaralı gözlemlerin veri kümesinden çıkarıldığı model olarak belirlenmiştir.

Tüm verilerle elde edilen model ve GSPR yöntemi ile belirlenen aykırı gözlemler çıkarıldıktan sonra elde edilen modele ilişkin veriler aşağıda verilmektedir.

Çizelge 6.8 İlk ve son modellerden elde edilen değerler

Veri Kümesinden Çıkarılan Gözlemler	Parametre Tahminleri					Pearson chi2	LR chi2
	le	pe	ooe	pce	n4e		
-	0,38495	134,71	122,52	0,18163	0,21924	145,65	109,11
64-116-145-149	0,47349	0,2043	0,26978	0,20903	0,28314	98,94	131,12

Tüm veriler ile elde edilen modelden elde edilen *LR chi2* değeri 109,11 iken aykırı gözlemler çıkarıldıktan sonra elde edilen modele ilişkin *LR chi2* değeri 131,12 olarak elde edilmiştir.

Sonuç olarak modelin uyum iyiliğine önemli bir katkı sağlanmış, yöntem maskeleye etkisini gidererek ve ilk durumda belirlenemeyen aykırı gözlemi de belirleyebilmiştir.

7. GSPR YÖNTEMİNİN İFLAS VERİLERİNE UYGULANMASI

GSPR yönteminin uygulanacağı ikinci veri kümesi Ali S. Hadi ve Samprit Chatterjee tarafından 2006 yılında yayımlanan “Regression Analyses by Example” isimli kitaptan alınmıştır. Veri kümesi 2 yıllık süre sonunda batan ve iş hayatına devam eden 66 firmaya ilişkin verilerden oluşmaktadır. Çoklu aykırı gözlemlerin oluşturulması amacıyla 25, 26 ve 27 numaralı gözlemlere ilişkin veriler değiştirilmiştir.

İflas modeline ilişkin değişkenler aşağıdaki gibi tanımlanmıştır.

7.1. İflas Verilerine İlişkin Değişkenler

Y:İlgili firma iki yıl sonra iflas etmiş ise Y değişkeni 0, ilgili firma 2 yıl sonra hala iş hayatına devam ediyorsa Y değişkeni 1 değerini almaktadır.

Kullanılan bağımsız değişkenler ise aşağıdaki gibi tanımlanmıştır.

$$X_1 = \frac{\text{Alınulmuş Kazançlar}}{\text{Toplam Varlıklar}}$$

$$X_2 = \frac{\text{Faiz ve Vergilerden Önceki Kazançlar}}{\text{Toplam Varlıklar}}$$

$$X_3 = \frac{\text{Satış Hasılatı}}{\text{Toplam Varlıklar}}$$

7.2. İflas Verilerinde Çoklu Aykırı Gözlemlerin Belirlenmesi

Yapılan çalışmada STATA 11.0, MATLAB R2008b ve Excel programları kullanılmıştır. İflas verileri aşağıda verilmektedir.

Çizelge 7.1. Batan ve hayatını devam ettiren şirketlere ilişkin finansal oranlar

Göz.No	Y	X ₁	X ₂	X ₃	Göz.No	Y	X ₁	X ₂	X ₃
1	0	62,8	-89,5	1,7	34	1	43	16,4	1,3
2	0	3,3	-3,5	1,1	35	1	47	16	1,9
3	0	-120,8	-103,2	2,5	36	1	-3,3	4	2,7
4	0	-18,1	-28,8	1,1	37	1	35	20,8	1,9
5	0	-3,8	-50,6	0,9	38	1	46,7	12,6	0,9
6	0	-61,2	-56,2	1,7	39	1	20,8	12,5	2,4
7	0	-20,3	-17,4	1	40	1	33	23,6	1,5
8	0	-194,5	-25,8	0,5	41	1	26,1	10,4	2,1
9	0	20,8	-4,3	1	42	1	68,6	13,8	1,6
10	0	-106,1	-22,9	1,5	43	1	37,3	33,4	3,5
11	0	-39,4	-35,7	1,2	44	1	59	23,1	5,5
12	0	-164,1	-17,7	1,3	45	1	49,6	23,8	1,9
13	0	-308,9	-65,8	0,8	46	1	12,5	7	1,8
14	0	7,2	-22,6	2	47	1	37,3	34,1	1,5
15	0	-118,3	-34,2	1,5	48	1	35,3	4,2	0,9
16	0	-185,9	-280	6,7	49	1	49,5	25,1	2,6
17	0	-34,6	-19,4	3,4	50	1	18,1	13,5	4
18	0	-27,9	6,3	1,3	51	1	31,4	15,7	1,9
19	0	-48,2	6,8	1,6	52	1	21,5	-14,4	1
20	0	-49,2	-17,2	0,3	53	1	8,5	5,8	1,5
21	0	-19,2	-36,7	0,8	54	1	40,6	5,8	1,8
22	0	-18,1	-6,5	0,9	55	1	34,6	26,4	1,8
23	0	-98	-20,8	1,7	56	1	19,9	26,7	2,3
24	0	-129	-14,2	1,3	57	1	17,4	12,6	1,3
25	0	20,7	28,6	2,1	58	1	54,7	14,6	1,7
26	0	21,2	27,4	2,8	59	1	53,5	20,6	1,1
27	0	33,4	32,3	2,1	60	1	35,9	26,4	2
28	0	-13,1	-17,6	0,9	61	1	39,4	30,5	1,9
29	0	-38	1,6	1,2	62	1	53,1	7,1	1,9
30	0	-57,9	0,7	0,8	63	1	39,8	13,8	1,2
31	0	-8,8	-9,1	0,9	64	1	59,5	7	2
32	0	-64,7	-4	0,1	65	1	16,3	20,4	1
33	0	-11,4	4,8	0,9	66	1	21,7	-7,8	1,6

GSPR yöntemi ile çoklu aykırı gözlemlerin belirlenmesi için ilk adım olarak veri kümesindeki tüm gözlemler ile logit model tahmin edilir. Tahmin edilen modelin sonuçları aşağıda verilmektedir.

logit y x1 x2 x3

Iteration 0: log likelihood = -45.747714
Iteration 1: log likelihood = -21.529465

```

Iteration 2: log likelihood = -17.392677
Iteration 3: log likelihood = -16.248254
Iteration 4: log likelihood = -16.240439
Iteration 5: log likelihood = -16.240429
Iteration 6: log likelihood = -16.240429

```

Logistic regression

```

Number of obs   =      66
LR chi2(3)      =      59.01
Prob > chi2     =      0.0000
Pseudo R2      =      0.6450

```

Log likelihood = -16.240429

y	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
x1	.0714817	.0222926	3.21	0.001	.0277889	.1151745
x2	.0468018	.0190082	2.46	0.014	.0095465	.0840572
x3	.0940809	.641897	0.15	0.883	-1.164014	1.352176
_cons	-1.027757	1.163543	-0.88	0.377	-3.308259	1.252745

Note: 2 failures and 0 successes completely determined.

Şekil7.1. İflas Verileri ile Tahmin Edilen Logit Model

Genelleştirilmiş standart pearson artığı (GSPR) yönteminin ilk aşaması olarak şüpheli aykırı gözlemlerin belirlenmesi gerekmektedir. Şüpheli aykırı gözlemler belirlenirken pearson artıkları ve standart pearson artıkları kullanılacaktır. Modelden elde edilen  olasılık, pearson ve standart pearson artık değerleri aşağıdaki gibi elde edilir.

Çizelge 7.2 Etkinlik Modeli olasılık, Pearson ve standart Pearson artık değerleri

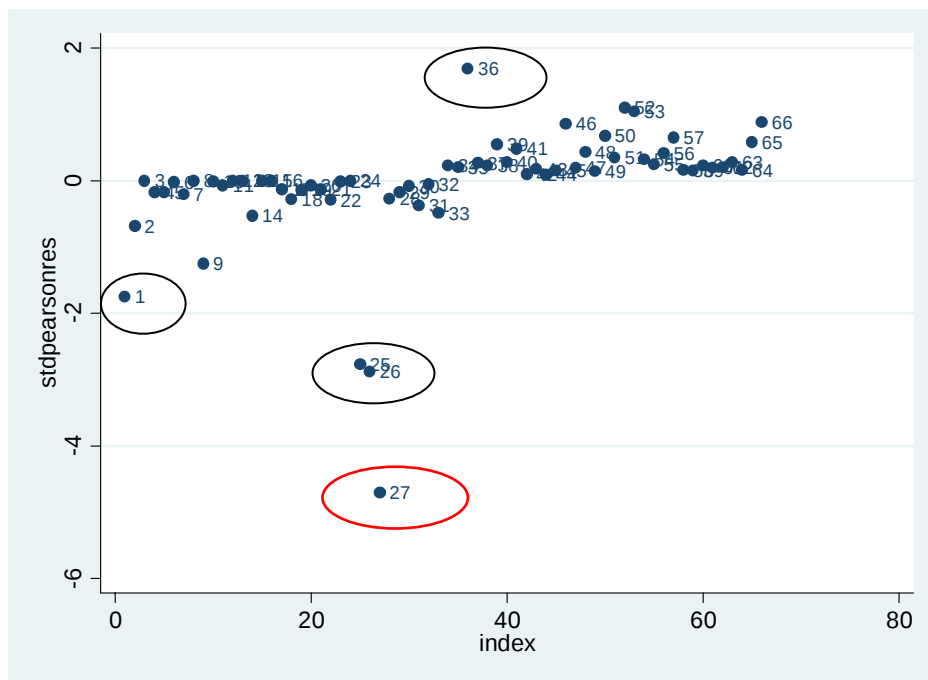
Göz. No		Pearson Artıkları	Standart Pearson Artıkları	Göz. No		Pearson Artıkları	Standart Pearson Artıkları
1	0.3618	-0.7530	-1.7500	34	0.9496	0.2304	0.2300
2	0.2990	-0.6530	-0.6800	35	0.9630	0.1960	0.2000
3	0.0000	0.0000	0.0000	36	0.3053	1.5086	1.6900
4	0.0275	-0.1681	-0.1700	37	0.9325	0.2690	0.2700
5	0.0270	-0.1667	-0.1700	38	0.9519	0.2248	0.2300
6	0.0004	-0.0195	-0.0200	39	0.7807	0.5300	0.5500
7	0.0392	-0.2020	-0.2100	40	0.9294	0.2757	0.2800
8	0.0000	0.0000	0.0000	41	0.8209	0.4671	0.4800
9	0.5871	-1.1924	-1.2500	42	0.9907	0.0967	0.1000
10	0.0001	-0.0083	-0.0100	43	0.9716	0.1711	0.1800
11	0.0045	-0.0672	-0.0700	44	0.9917	0.0913	0.0900
12	0.0000	0.0000	0.0000	45	0.9783	0.1488	0.1500
13	0.0000	0.0000	0.0000	46	0.5897	0.8341	0.8600
14	0.2006	-0.5009	-0.5300	47	0.9669	0.1849	0.1900
15	0.0000	-0.0048	0.0000	48	0.8553	0.4113	0.4300
16	0.0000	0.0000	0.0000	49	0.9807	0.1402	0.1400
17	0.0165	-0.1295	-0.1300	50	0.7815	0.5288	0.6800
18	0.0688	-0.2719	-0.2800	51	0.8938	0.3447	0.3500
19	0.0179	-0.1350	-0.1400	52	0.4823	1.0360	1.1000
20	0.0049	-0.0699	-0.0700	53	0.4981	1.0038	1.0400
21	0.0172	-0.1325	-0.1300	54	0.9101	0.3142	0.3200
22	0.0730	-0.2807	-0.2900	55	0.9453	0.2405	0.2400
23	0.0001	-0.0117	-0.0100	56	0.8654	0.3944	0.4100
24	0.0000	-0.0044	0.0000	57	0.7167	0.6287	0.6500
25	0.8795	-2.7020	-2.7700	58	0.9765	0.1552	0.1600
26	0.8843	-2.7641	-2.8800	59	0.9794	0.1448	0.1500
27	0.9556	-4.6390	-4.7100	60	0.9508	0.2274	0.2300
28	0.0628	-0.2588	-0.2700	61	0.9675	0.1832	0.1900
29	0.0278	-0.1689	-0.1700	62	0.9637	0.1941	0.2000
30	0.0063	-0.0797	-0.0800	63	0.9293	0.2758	0.2800
31	0.1194	-0.3682	-0.3800	64	0.9768	0.1541	0.1600
32	0.0029	-0.0542	-0.0500	65	0.7661	0.5526	0.5800
33	0.1775	-0.4646	-0.4900	66	0.5766	0.8569	0.8800

Aykırı gözlemleri belirlemede kullanılan pearson ve standart pearson artıkları yöntemlerine göre, ilgili gözlemlerin pearson ve standart pearson artık değerleri mutlak değer olarak 3'ten büyükse o gözlem aykırı gözlem olarak kabul edilir.

Çizelge 7.3. Aykırı Gözlemlerin Pearson ve Standart Pearson Artık Değerleri

Gözlem No	$\bar{r}$	Pearsonres	stdpersres
27	0.9556	-4.6390	-4.7100

27 numaralı gözlemin pearson ve standart pearson artıklarının mutlak değeri 3'ten büyüktür ve bu gözlem aykırı gözlem olarak kabul edilir. Standart Pearson artıklarının gözlem numaralarına göre grafikleri çizildiğinde aşağıdaki grafik elde edilir.



Şekil 7.2. Pearson artıklarının gözlem numaralarına göre grafiği

Grafikten de görüldüğü gibi 27 numaralı gözlemin artık değeri sınırının dışına düşmektedir. Bu da ilgili gözlemlerin aykırı gözlem olduğuna işaret etmektedir.

Şekil (7.2) incelendiğinde artık değerleri diğer gözlemlerden uzağa düşen başka gözlemlerin de olduğu görülmektedir. Bu da veri kümesindeki tüm aykırı gözlemlerin belirlenip belirlenemediği sorusunu akla getirmektedir. Elde edilen artık değerleri ve grafikler incelendiğinde 1, 25, 26 ve 36 numaralı gözlemlerin de incelenmesi gerektiği düşünülebilir. Bu gözlemlere ait standart pearson artık değerleri aşağıda verilmektedir.

Çizelge 7.4. Şüpheli görülen gözlemlerin pearson ve standart pearson artık değerleri

Gözlem No	p	Pearsonres	stdpersres
1	0.3618	-0.7530	-1.7500
25	0.8795	-2.7020	-2.7700
26	0.8843	-2.7641	-2.8800
36	0.3053	1.5086	1.6900

Pearson ve Standart Pearson artık değerleri diğer gözlemlerden uzağa düşmektedirler. Bu durumda bu gözlemlerin aykırı gözlem olduğu halde aykırı gözlem olarak belirlenemediği yani maskeleye etkisine uğradığı düşünülebilir. Elde edilen artık değerlerine göre veri kümesinden ilk olarak çıkarılacak olan D_0 gözlem grubu 1, 25, 26, 27 ve 36 numaralı gözlemlerden oluşur.

Yöntemin ikinci aşamasında D_0 gözlem grubu veri kümesinden çıkarıldıktan sonra kalan gözlemler üzerinden tekrar lojistik regresyon modeli tahmin edilir. Geriye kalan gözlem grubu, R_0 , 61 gözlem içermektedir ve bu gözlemler ile elde edilen model aşağıda verilmektedir.

logit y x1 x2 x3

Iteration 0: log likelihood = -42.208178
 Iteration 1: log likelihood = -8.777732
 Iteration 2: log likelihood = -5.6807351
 Iteration 3: log likelihood = -3.5720456
 Iteration 4: log likelihood = -2.9480128
 Iteration 5: log likelihood = -2.8250834
 Iteration 6: log likelihood = -2.8196053
 Iteration 7: log likelihood = -2.8195957
 Iteration 8: log likelihood = -2.8195957

Logistic regression

Number of obs = 61
 LR chi2(3) = 78.78
 Prob > chi2 = 0.0000
 Pseudo R2 = 0.9332

Log likelihood = -2.8195957

y	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
x1	.2924345	.2671246	1.09	0.274	-.2311201 .815989
x2	.1639237	.1104429	1.48	0.138	-.0525404 .3803877
x3	4.217817	4.873041	0.87	0.387	-5.333168 13.7688
_cons	-8.544247	9.755211	-0.88	0.381	-27.66411 10.57561

Note: 14 failures and 5 successes completely determined.

Şekil7.3. R_0 gözlem grubu ile elde edilen logit model tahmini

Yukarıda elde edilen yeni modelden, tüm veri kümesi için $(R_0 + D_0 \text{ kümesi}) \pi_i^{(-D)}$ değerleri tahmin edilir. Bu olasılık değerleri bulunduğundan sonra gözlem değerlerinden olasılık değerleri çıkarılarak yeni artık değerleri bulunur. Genelleştirilmiş Standart Pearson Artık değerleri, lojistik regresyon için (5.6)'da verilen eşitlik ile elde edilir. $|r_{si}^{(-D)}| > 3$ ise o gözlemin aykırı gözlem olduğu yorumu yapılır. D_0 gözlem grubuna ait $r_{si}^{(-D)}$ değerlerine ait sonuçlar aşağıda verilmektedir.

Çizelge 7.5. D_0 gözlem grubuna ait Pearson, Standart Pearson ve GSPR değerleri

Göz. No	p	Pearsonres	Stdpersres	GSPR(1)
1	0.3618	-0.7530	-1.7500	-0.8522
25	0.8795	-2.7020	-2.7700	-246.8918
26	0.8843	-2.7641	-2.8800	-1053.7389
27	0,9556	-4.6390	-4.7100	-2131.7532
36	0.3053	1.5086	1.6900	0.1807

Tablodan görüldüğü gibi $|r_{si}^{(-D)}| > 3$ şartını 25, 26 ve 27 numaralı gözlemler sağlamaktadır. 1 ve 36 numaralı gözlemler, $|r_{si}^{(-D)}| > 3$ şartını sağlamadığı için yöntem gereğince R_0 gözlem grubuna eklenirler.

Çoklu aykırı gözlem olarak belirlenecek şüpheli gözlemlerin sayısı 3'e inmiştir. Yani döngünün ikinci adımında R_1 gözlem grubu 63 gözlem, D_1 gözlem grubu 3 gözlem içerecektir. R_1 gözlem grubu ile elde edilen model aşağıda verilmektedir.

```
logit y x1 x2 x3
```

```
Iteration 0: log likelihood = -43.596817
Iteration 1: log likelihood = -9.4594996
Iteration 2: log likelihood = -9.3475201
Iteration 3: log likelihood = -5.3133059
Iteration 4: log likelihood = -3.2114955
Iteration 5: log likelihood = -3.0305113
Iteration 6: log likelihood = -3.0012046
Iteration 7: log likelihood = -3.0010371
Iteration 8: log likelihood = -3.0010371
```

Logistic regression

```
Number of obs = 63
LR chi2(3) = 81.19
Prob > chi2 = 0.0000
Pseudo R2 = 0.9312
```

Log likelihood = -3.0010371

y	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
x1	.2317493	.1509443	1.54	0.125	-.064096	.5275947
x2	.1951239	.0987957	1.98	0.048	.0014879	.3887598
x3	3.826246	3.086423	1.24	0.215	-2.223032	9.875524
_cons	-6.877849	5.470632	-1.26	0.209	-17.60009	3.844393

Note: 13 failures and 3 successes completely determined.

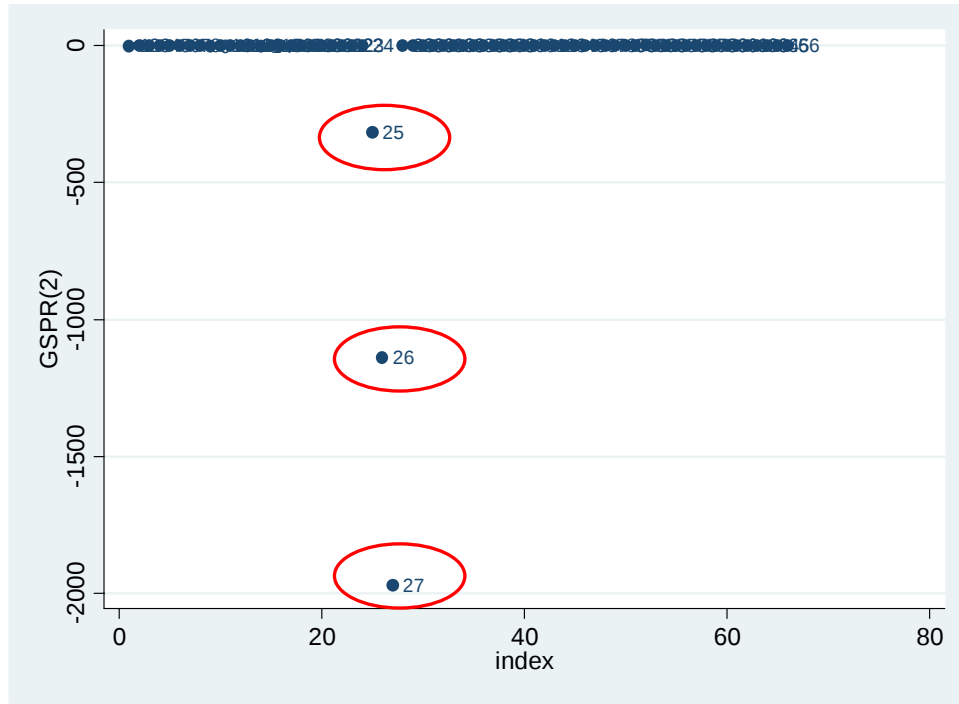
Şekil 7.4. Kalan gözlem grubu R_1 ile elde edilen logit model tahmini

Elde edilen yeni modelden, tüm veri kümesi için $(R_1 + D_1 \text{ kümesi})$ olasılık değerleri $\pi_i^{(-D)}$ değerleri tahmin edilir. Bu olasılık değerleri bulunduktan sonra gözlem değerlerinden olasılık değerleri çıkarılarak yeni artık değerleri bulunur. Genelleştirilmiş Standart Pearson Artığı, (5.6) eşitliği ile elde edilir. Döngünün ikinci adımı sonucunda, D_1 gözlem grubuna ilişkin sonuçlar aşağıda verilmektedir.

Çizelge 7.6. D_1 gözlem grubuna ait Pearson, Standart Pearson ve GSPR değerleri

Göz. No	Pearsonres	stdpersres	GSPR(2)
25	-2.7020	-2.7700	-316.6984
26	-2.7641	-2.8800	-1139.0501
27	-4.6390	-4.7100	-1969.9235

Tablodan görüldüğü gibi $|r_{si}^{(-D)}| > 3$ şartını D_1 gözlem grubundaki 25, 26 ve 27 numaralı gözlemlerin hepsi sağlamaktadır. Dolayısıyla döngü sona ermiş, veri kümesindeki aykırı gözlemler belirlenmiştir. Elde edilen yeni GSPR(2) değerlerinin gözlem numaralarına göre grafiği çizdirildiğinde aşağıdaki grafik elde edilir.



Şekil 7.5. GSPR(2) değerlerinin gözlem numaralarına göre grafiği

Grafikten de açık bir şekilde görüldüğü gibi 25, 26 ve 27 numaralı gözlemler aykırı gözlem olarak belirlenmektedir. Pearson ve standart pearson artıkları kullanarak yapılan aykırı gözlem incelemesinde sadece 27 numaralı gözlem aykırı gözlem olarak belirlenebilmiştir. Ancak GSPR kullanılarak yapılan aykırı gözlem incelemesinde ise 25, 26 ve 27 numaralı gözlemler aykırı gözlem olarak belirlenmiştir. Bu noktada ilk durumda 25 ve 26 numaralı gözlemlerin regresyon modeli üzerindeki etkisinin maskelendiğini söyleyebiliriz. Verilen yöntem maskeleye etkisini kaldırarak tüm aykırı gözlemleri belirleyebilmiştir.

8. SONUÇ

Bu çalışmada, lojistik regresyon analizinde çoklu aykırı gözlemlerin belirlenmesi amacıyla geliştirilen ve veri kümesinde birden fazla aykırı gözlemin bulunduğu durumlarda, maskeleme probleminden etkilenmeyen Genelleştirilmiş Standart Pearson Artıkları yöntemi iki gerçek olay çalışması üzerinden uygulanmıştır.

Regresyon analizinde aykırı gözlemlerin belirlenmesi önemli bir adımdır. Bu noktada klasik aykırı gözlem belirleme yöntemlerinin çoklu aykırı gözlemleri belirlemede yeterli olmadığı ortaya çıkmaktadır. Genelleştirilmiş Standart Pearson Artıkları yöntemi, çoklu aykırı gözlemlerin belirlenmesinde etkili bir yöntem olarak kullanılabilir. Bu yöntem ile klasik yöntemlerle belirlenemeyen aykırı gözlemlerin belirlenmesi sağlanmış, maskeleme ve sahte aykırı gözlem problemlerinin etkileri giderilmiş, aykırı gözlemlerin model üzerindeki etkileri belirlenebilmiştir.

Uygulama çalışmasında, Cengiz Akarçeşme tarafından yapılan “Elit Bayan Voleybolunda Maç Sonucunu Açıklayan Değişkenlerin Lojistik Regresyon Yöntemi ile Belirlenmesi ve Maç Kazanmaya Yönelik Olasılık Modelinin Tahmini” isimli doktora tezinde, voleybol maçlarının kazanılıp kaybedilme olasılığının modellenmesi amacıyla kullanılan “Etkinlik Modeli” verileri ve Ali S. Hadi ve Samprit Chatterjee tarafından 2006 yılında yayımlanan “Regression Analyses by Example” isimli kitapta yer alan iflas verileri kullanılmıştır. Veriler incelendiğinde aykırı gözlemlere rastlanmış, klasik yöntemler olarak değerlendirilen Pearson ve Standart Pearson Artıkları yöntemi ile aykırı gözlemler belirlenmiştir. Ancak bu yöntemlerin tekil aykırı gözlem durumunda güçlü olması ve veri kümesinde başka şüpheli aykırı gözlemlerin bulunması sebebiyle GSPR yöntemi uygulanmıştır. Yöntemin ilk adımında şüpheli olarak görülen gözlemler veri kümesinden çıkarılmış ve gerekli işlemler sonucunda GSPR değerleri hesaplanmıştır. Hesaplanan GSPR değerlerinden mutlak değerce 3’ten küçük olanlar veri kümesinde kalan gözlem grubuna (R_0) eklenmiş ve ikinci adıma geçilmiştir. Yine gerekli işlemlerin sonucunda

ikinci GSPR deęerleri hesaplanmış ve řüpheli gözlemler için ilgili deęerlerin 3'ten büyük olduęu görölmüřtür. Bu sonuç, veri kümesinde ilk durumda belirlenenden daha fazla sayıda aykırı gözlem olduęunu göstermektedir. İlk durumda klasik yöntemler kullanıldığında aykırı gözlem grubunun tamamı belirlenememiřtir. Yani veri kümesi maskeleye probleminden etkilenmiřtir. Ancak GSPR yöntemi kullanıldığında bu etki giderilmiş ve çoklu aykırı gözlemler belirlenebilmiřtir.

Genelleřtirilmiş Standart Pearson Artıkları(GSPR) yöntemi ile;

1. Aykırı gözlem belirlemede karřılan en önemli problemlerden olan maskeleye ve sahte aykırı gözlem etkileri giderilmektedir.
2. Veri kümesinde bulunan ve klasik yöntemlerle belirlenemeyen çoklu aykırı gözlemler başarılı bir řekilde belirlenmektedir.
3. Belirlenen aykırı gözlemler veri kümesinden çıkarıldıktan sonra elde edilen modelin en iyi model olduęu görölmektedir.

Lojistik regresyonun aykırı gözlemlere karřı oldukça hassas olduęu ve gerçek yaşamda kullanılan veri setlerinde birden fazla aykırı gözlem bulunduęu göz önüne alındığında klasik aykırı gözlem belirleme yöntemlerine alternatif olarak Genelleřtirilmiş Standart Pearson Artıkları yönteminin kullanılması çoklu aykırı gözlemleri belirlemede oldukça büyük fayda sağlayacaktır.

KAYNAKLAR

Akarçeşme C., “Elit Bayan Voleybolunda Maç Sonucunu Açıklayan Değişkenlerin Lojistik Regresyon Yöntemi ile Belirlenmesi ve Maç Kazanmaya Yönelik Olasılık Modelinin Tahmini”, *Gazi Üniversitesi Sağlık Bilimleri Enstitüsü Beden Eğitimi ve Spor Anabilimdalı Doktora Tezi*(2010)

Atkinson, A., “Masking Unmasked”, *Biometrika*, 74(3); 533- 541 (1986)

Atkinson, A., “Fast Very Robust Methods for the Detection of Multiple Outliers”, *Journal of the American Statistical Association*, 89(428); 1329-1339 (1994)

Barret B.E., Gray B.J., "Leverage, residual, and interaction diagnostics for subsets of cases in least squares regression", *Computational Statistics & Data Analysis*, 26; 39-52 (1997)

Bedrick E.J. Ve Hill J.R., "Outlier tests for logistic regression: A conditional approach", *Biometrika*, 77(4); 815-27 (1990)

Billor, N., Chatterjee, S., and Hadi, A. S., "A Re-Weighted Least Squares Method for Robust Regression Estimation," *American Journal of Mathematical and Management Sciences*, 26; 229–252 (2006)

Chatterjee S. ve Hadi Ali S., "Regression Analysis by Example", *Wiley-Interscience*, (2006)

Cook D.R., Weisberg S., "Residuals and Influence in Regression", *Chapman and Hall*, (1982)

Draper N.R., Smith H., “Applied Regression Analysis”, *Wiley Series in Probability and Statistics*, (1998)

Gentleman, J. F., Wilk, M. B., “Detecting Outliers . II. Supplementing The Direct Analysis of Residuals”, *Biometrics*, 31; 387 - 410 (1975)

Gujarati, D. N., "Temel Ekonometri", *Literatür Yayınları*, (1999)

Habing B., "More on Outlier Diagnostics", *University of South Carolina*, (2004)

Hadi A. S., "Identifying Multiple Outliers in Multivariate Data", *J.R. Statist. Soc. B*54(3); 761-771(1992)

Hadi, A. S., “Identifying Multiple Outliers in Multivariate Data”, *Journal of the Royal Statistical Society, Series B (Methodological)*, 54(3); 761-771 (1992)

Hadi A.S., Imon R.A.H.M., "Identification of Multiple Outliers in Logistic Regression", *Communications in Statistics—Theory and Methods*, 37; 1697–1709, (2008)

Hosmer D.W., Hosmer T., Cessie S. LE., Lemeshow S., "A Comparison Of Goodness-Of-Fit Tests For The Logistic Regression Model", *Statistics In Medicine*, 16; 965-980 (1997)

Hosmer D.W., Lemeshow Stanley, "Applied Logistic Regression", *Wiley-Interscience*, (2000)

Kianifard, F., Shallow, W., H., "Using Recursive Residuals, Calculated on Adaptively-Ordered Observations, to Identify Outliers in Linear Regression", *Biometrics*, 45(2); 571-585 (1989)

Lee, A.H., "Assessing partial influence in generalized linear models", *Biometrics*, 44; 71–77 (1988)

Lee, A.H., Zhao, Y., "Assessing influence on the goodness of linf in generalized linear models", *Statist. Probab. Lett.*, 31; 351–358 (1997)

Marasinghe, M., "A Multivariate Procedure for Detecting Several Outliers in Linear Regression", *Technometrics*, 27(4); 395-399 (1985)

Marchette, D.J., Solka, J.L., "Using Data Images for Outlier Detection", *Computational Statistics & Data Analysis*, 43; 541-552 (2003)

Mohammad N.M., Robiah A., Setan H, "Multiple Outliers Detection Procedures in Linear Regression", *Matematika*, 19(1); 29-45 (2003)

Muñoz G.J., Muñoz P.J.M., Pardo L., "Cressie and Read power-divergences as influence measures for logistic regression models", *Computational Statistics & Data Analysis*, 50; 3199 – 3221 (2006)

Neter J., Kutner M.H., Wasserman W., "Applied Linear Regression Models", *Irwin/mcgraw-hill*, (1996)

Pena, D., Yohai, V., "The Detection of Influential Subsets in Linear Regression by using an Influence Matrix", *Journal of the Royal Statistical Society, Series B (Methodological)*, 57(1); 145-156 (1995)

Pregibon, D., "Logistic regression diagnostics", *Ann. Statist.*, 9; 977–986 (1981).

Rousseeuw, P., J., Driessen, K., V., "A Fast Algorithm for the Minimum Covariance Determinant Estimator", *Technometrics*, 41(3); 212-223 (1999)

Rousseeuw, P., J., Zomeren, B.,C.,V., “Unmasking Multivariate Outliers and Leverage Points”, *Journal of the American Statistical Association*, 85(411); 633-639 (1990)

Sebert, M., D., Montgomery, D., C., Rollier, D., A., “A Clustering Algorithm for Identifying Multiple Outliers in Linear Regression”, *Computational Statistics & Data Analysis*, 27; 461-484 (1998)

Thomas, W., Cook, R.D., “Assessing influence on regression coefficients in generalized linear models”, *Biometrika* 76 (4); 741–749 (1989)

Thomas, W., Cook, R.D., “Assessing influence on predictions from generalized linear models”, *Technometrics*, 32 (1); 59–65 (1990)

Williams, D.A., “Generalized linear models diagnostics using the deviance and single case deletions”, *Appl. Statist.*, 36; 181–191 (1987)

Wisnowski J.W., Montgomery D.C., Simpson J.R., “A Comparative analysis of multiple outlier detection procedures in the linear regression model”, *Computational Statistics & Data Analysis*, 36; 351-382 (2001)

ÖZGEÇMİŞ

Kişisel Bilgiler

Soyadı, adı : YAVUZKANAT, Mustafa Selçuk
 Uyruğu : T.C.
 Doğum tarihi ve yeri : 31.01.1980 / Ankara
 Medeni hali : Evli
 Telefon : 0 (312) 310 00 26
 e-mail : syavuzkanat@hotmail.com

Eğitim

Derece	Eğitim Birimi	Mezuniyet tarihi
Yüksek Lisans	Gazi Üniversitesi- İstatistik Bölümü	2011
Lisans	Hacettepe Üniversitesi- İstatistik Bölümü	2002
Lise	Alparslan Lisesi	1997

İş Deneyimi

Yıl	Yer	Görev
2004-2006	T. İş Bankası	Memur
2006-2011	T.C. Kültür ve Turizm Bakanlığı- Telif Hakları ve Sinema Genel Müdürlüğü	Kültür ve Turizm Uzmanı

Yabancı Dil

İngilizce

Hobiler

Müzik, Tarih, Sinema