
GAZİ ÜNİVERSİTESİ
Gazi
1926

BINARY LOJİSTİK REGRESYON ANALİZİNDE MODEL UYUMUNDA KULLANILAN TESTLERİN PERFORMANSLARININ DEĞERLENDİRİLMESİ

Esra ARSLANOĞLU

YÜKSEK LİSANS TEZİ
İSTATİSTİK ANA BİLİM DALI

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

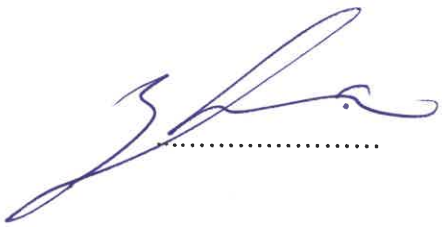
TEMMUZ 2019

Esra ARSLANOĞLU tarafından hazırlanan "BINARY LOJİSTİK REGRESYON ANALİZİNDEMODEL UYUMUNDA KULLANILAN TESTLERİN PERFORMANSLARININ DEĞERLENDİRİLMESİ" adlı tez çalışması aşağıdaki jüritarafından OY BİRLİĞİ ile Gazi Üniversitesi İstatistik Ana Bilim DalındaYÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

Danışman: Prof. Dr. Bülent ÇELİK

İstatistikAna Bilim Dalı, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum. .....



Başkan: Doç. Dr. S. Kenan KÖSE

BiyoistatistikAna BilimDalı,Ankara Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum. .....

Üye: Prof. Dr. H. Hasan ÖRKÇÜ

İstatistik Ana BilimDalı,Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum. .....

Tez SavunmaTarihi: 18/07/2019

Jüri tarafından kabul edilen bu tezin Yüksek Lisans Tezi olması için gerekli şartları yerine getirdiğini onaylıyorum.

.....

Prof. Dr. Sena YAŞYERLİ

Fen Bilimleri Enstitüsü Müdürü

ETİK BEYAN

Gazi Üniversitesi Fen Bilimleri Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmasında;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,
- Tez çalışmasında yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- Bu tezde sunduğum çalışmanın özgün olduğunu,

bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

Esra ARSLANOĞLU

18/07/2019

BINARY LOJİSTİK REGRESYON ANALİZİNDE MODEL UYUMUNDA KULLANILAN TESTLERİN PERFORMANSLARININ DEĞERLENDİRİLMESİ

(Yüksek Lisans Tezi)

Esra ARSLANOĞLU

GAZİ ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

Temmuz 2019

ÖZET

Günümüzde tıp, biyoloji, ekonomi vb. alanlarda kullanılabilen lojistik regresyon modelleri, hem bağımlı değişken ile bağımsız değişkenler (ler) arasındaki ilişkiyi tanımlayan en uygun modeli belirlemeyi sağlayan hem de bilinen bağımsız değişken (ler) yardımıyla bilinmeyen bağımlı değişken hakkında kestirim yapmamıza olanak sağlayan son derece önemli modellerdendir. Lojistik regresyon analizinin diğer regresyon analizlerinden ayırt edici özelliği ise bağımlı değişkenin niteliksel olmasıdır. Lojistik regresyon analizinde bağımlı değişkeni en iyi kestirimde bulunan bağımsız değişkenlerle model kurulduktan sonra modelin veriye uyumunu değerlendirmek amacıyla çeşitli yöntemler kullanılmaktadır. Bu çalışmada model uyumu değerlendirmesinde kullanılan bu yöntemlerin performansları değerlendirilmeye çalışılacaktır.

Bilim Kodu : 20503
Anahtar Kelimeler : Lojistik regresyon, model uyumu, hosmer ve lemeshow testi roc eğrisi analizi
Sayfa Adedi : 55
Danışman : Prof. Dr. Bülent ÇELİK

EVALUATION OF PERFORMANCE OF TESTS USED IN MODEL ALIGNMENT IN BINARY LOGISTIC REGRESSİON ANALYSIS

(M. Sc. Thesis)

Esra ARSLANOĞLU

GAZİ UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

July 2019

ABSTRACT

Logistic regression models are used in scientific fields such as medicine, biology, economy. Logistic regression models are very important models that provide both determination the optimal model for defining the relationship the dependent variable and independent variable(s) and prediction about unknown dependent variable via known independent variable(s). What distinguishes logistic regression analysis from other regression analysis is that the dependent variable is qualitative. The model in logistic regression analysis, is built by independent variable(s) that make the best prediction for the dependent variable. Then various methods are used for evaluating whether the model is fit with the data. In this thesis, the performances of these methods that use in the model fitting evaluation are investigated.

Science Code : 20503
Key Words : Logistic regression, model fitting, hosmer and lemeshow test, roc curve analysis
Page Number : 55
Supervisor : Prof. Dr. Bülent ÇELİK

TEŞEKKÜR

Teşekkür sayfası Hazırlamış olduğum tez çalışmasında benden hiçbir zaman desteğini esirgemeyen, bilgisi ve deneyimiyle yüksek lisans eğitimim boyunca yanımda olan çok değerli tez danışman hocam Prof. Dr. Bülent ÇELİK'e, İstatistik Anabilim Dalı'nda başta bölüm başkanımız Prof. Dr. İhsan ALP olmak üzere bütün öğretim üyelerine, bölüm arkadaşlarıma, yüksek lisans eğitimim boyunca tez çalışmam esnasında tüm sıkıntı ve yorgunluğumu paylaşan en zor zamanlarda bana destek olan arkadaşlarıma, her daim yanımda olan, en büyük destekçilerim, hayatımın anlamları babam Mükremin ARSLANOĞLU, annem Pembe ARSLANOĞLU, ablam Buket ÜNLÜ ve eniştem Mustafa ÜNLÜ'ye sonsuz çok teşekkür ederim.

İÇİNDEKİLER

Sayfa

Sayfa

ÇİZELGELERİN LİSTESİ

Çizelge **Sayfa**

ŞEKİLLERİN LİSTESİ

Şekil **Sayfa**

SİMGELER VE KISALTMALAR

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

Simgeler	**Açıklamalar**
x^2	Khi kare test istatistiği
p değeri	Anlamlılık düzeyi
r	Korelasyon katsayısı

Kısaltmalar	**Açıklamalar**
AO	Aritmetik ortalama
AUC	Roc eğrisi altında kalan alan
HDL	İyi kolesterol (high density lipoprotein)
LDL	Kötü kolesterol (low density lipoprotein)
ROC	Reciever operatör characteristics curve
SH	Standart hata
SS	Standart sapma

1. GİRİŞ

İstatistik bilimi, olayların gözlemlenmesi yoluyla elde edilen verilerin toplanması, verilerin özetlenmesi ile analizi ve bu verilerden elde edilen sonuçların yorumlanmasını sağlamaktadır. Bir bilim dalı olarak eldeki verileri sayısal yöntemlerle analiz ederek gelecek hakkında öngörüde bulunmayı kolaylaştırmaktadır.

Tahminde bulunmayı sağlayan önemli istatistiksel analiz türlerinden biri olan regresyon analizi, verilerden faydalanarak tahmin yapmamıza olanak sağlayan bir analizdir. Regresyon analizi bir veya birden fazla bağımsız değişken ile bir bağımlı değişken arasındaki ilişkiyi ifade etmeyi ve bağımlı değişken hakkında çıkarımda bulunmayı sağlayan bir istatistiksel analiz türüdür. Regresyon analizinin doğrusal regresyon, probit regresyon, ordinal regresyon ve lojistik regresyon gibi çeşitleri vardır.

Günümüzde en çok kullanılan regresyon analiz türleri doğrusal regresyon analizi ve lojistik regresyon analizidir. Doğrusal regresyon analizinde, bağımlı değişken sürekli değişken olmak zorunda iken bağımsız değişkenler sürekli veya kategorik olabilmektedir. Bağımlı değişkenin kategorik olduğu durumlarda ise lojistik regresyon analizi kullanılmaktadır. Lojistik regresyon bağımlı değişkenin iki veya ikiden fazla kategorisinden birine karşı diğerinin veya diğerlerinin olma olasılığından yararlanarak bağımsız değişkenlerin bağımlı değişken üzerine etkisini incelemektedir [1].

Bu tezin amacı, lojistik regresyon analizinde model uyumunun değerlendirilmesi amacıyla kullanılan testler tanıtılarak bu yöntemlerin birbirine benzer sonuç verip vermediğini incelemektir.

Bu çalışmanın birinci bölümünde lojistik regresyon analizi, varsayımları, türleri ve katsayıların önemliliği testlerine ilişkin bilgi verilecektir.

Çalışmanın ikinci bölümünde, model uyumu testleri detaylıca anlatılacak olup örneklerle konu ele alınacaktır.

Üçüncü ve son bölümünde ise lojistik regresyon analizi kullanılarak preeklampsi görülen gebeler ve preeklampsi görülmeyen sağlıklı gebeler bağımlı değişken olarak alınarak çeşitli bağımsız değişkenlerle bir model oluşturulup modelin veriye uyumlu olup

olmadığının incelenmesi konusunda pearson ki kare ve sapma istatistiği, pseudo değerleri (cox ve snell, nagelkerke, mc fadden), hosmer lemeshow testleri, sınıflandırma tabloları ve roc eğrisi altında kalan alan kullanılmıştır. Kullanılan testlerin birbirine yakın sonuçlar verip vermediğinin değerlendirilmesi amaçlanmıştır.

2. LOJİSTİK REGRESYON ANALİZİ

Regresyon modelleri, bağımsız değişken(ler)'in bağımlı değişken ile arasındaki sebep sonuç ilişkisini regresyon denklemi ile açıklanmasını sağlar. Ayrıca regresyon analizi bilinen bağımsız değişken(ler) ile bilinmeyen bağımlı değişken hakkında kestirim yapmamıza olanak sağlar.

Lojistik regresyon analizinin diğer regresyon analizlerinden ayırt edici özelliği, bağımlı değişkenin kategorik olmasıdır. Modele dahil edilen bağımsız değişken(ler) niteliksel veya niceliksel veri olabilir.

Bir olayın meydana gelip gelmeyeceğini önceden bilmek veya kestirmek oldukça zor olmakla birlikte, bu durumu tahmin etmek ve bu tahminin hangi değişkenler tarafından daha iyi yapıldığını belirlemek son derece önemlidir. Böyle bir durumu tahmin etmek için kullanılacak bağımlı değişkenin kategorik ve iki durumlu olması gerekeceği açıktır. Söz konusu olayın meydana gelip gelmeyeceğini ifade eden ve bu iki şıklı bağımlı değişkeni etkileyen faktörlerin belirlenmesi için lojistik regresyon analizine başvurulur [2, 3].

Lojistik regresyon analizi üzerine ilk çalışmalar Berkson tarafından biyolojik deneylerin analizi için kullanılmıştır [1].

Lojistik regresyon analizinin tıp, biyoloji, ekonomi, eğitim vb. alanlarda gün geçtikçe kullanımı artmaktadır. Lojistik regresyon analizin tercih edilmesindeki artışın en önemli nedenlerinden biri varsayımlarının diğer regresyon analiz türlerine oranla az olması ve yorumlamasının kolaylığıdır.

2.1. Lojistik Regresyon Analizinin Varsayımları

Lojistik regresyon analizinde lineer regresyon analizindeki varsayımlardan hiçbiri aranmaz [4]. Tek koşul bağımlı değişkenin kategorik bir değişken olmasıdır. Ancak analiz sonuçlarının daha güvenilir olması açısından bazı hususlara dikkat edilmelidir. Bu hususlar;

1. Bağımsız değişkenler arasında çoklu bağlantı yani doğrusal ilişki olmamalıdır. Birbiriyle ilişkili olan değişkenlerden bir tanesi modele dahil edilmelidir. Sürekli

değişken olan bağımsız değişkenler arasındaki ilişki korelasyon analizi ile değerlendirilmelidir.

2. Hata terimleri arasında otokorelasyon olmamalıdır, yani regresyon modelinde hata terimlerinin birbiriyle ilişkili olmaması gerekmektedir [5].
3. Lojistik Regresyon Analizi, doğrusallık varsayımını gerektirmez. Bağımsız değişken(ler)le bağımlı değişkenin arasındaki ilişki doğrusal değildir. Ancak Lojistik Regresyon denkleminin lojit dönüşümü, doğrusalaştırma işlemini barındırdığı için doğrusallık şartı arar. Bu sebeple varsayımın gerçekleşmemesi, bağımlı ve bağımsız değişken(ler) arasındaki ilişkilerin açıklanma gücünü düşürür.
4. Örneklem genişliğinin yeterli büyüklükte olması önemlidir. Değişken sayısının en az on katı kadar gözlem bulunması önerilmektedir [6].

Lojistik regresyonun giderek daha çok tercih edilir olmasının nedenleri [7,8];

1. Bağımlı değişken kategorik iken, bağımsız değişken(ler)in kesikli, sürekli veya kategorik olduğu durumlarda uygulanabilmektedir,
2. Lojistik regresyon analizi birçok bilgisayar paket programı içerisinde bulunmaktadır,
3. Bağımsız değişkenlerin olasılık fonksiyonlarının dağılımları üzerinde kısıt olmaması nedeniyle çeşitli testler uygulanabilmektedir,
4. Lojistik regresyonda tüm olasılık değerleri 0 ile 1 arasında değişmektedir.

2.2. Lojistik Regresyon Analiz Türleri

Lojistik regresyon analizinde, bağımlı değişkenin kategori durumuna göre iki kategorili (binominal) bağımlı değişkenler için iki kategorili lojistik regresyon analizi, ikiden çok sıralı olmayan kategorili bağımlı değişkenler için çok kategorili (multinominal) lojistik regresyon analizi ve ikiden çok sıralı kategorili bağımlı değişkenler için sıralı (ordinal) lojistik regresyon analizi kullanılmaktadır [6].

2.2.1. İki kategorili lojistik regresyon analizi

İki kategorili lojistik regresyon modelinde, bağımlı değişken iki kategoriden oluşmaktadır. Bağımlı değişken kodlanırken riskin olmadığı grup için 0 kodu riskin olduğu grup için 1

kodu kullanılır. Bağımlı değişkenin olasılık değeri 0 ile 1 arasında değişir. Modele dahil edilen bağımsız değişken(ler) nicel veya nitel değişken olabilir.

2.2.2. Çok kategorili (multinominal) lojistik regresyon analizi

Multinominal lojistik regresyon modelinde bağımlı değişken ikiden çok kategoriye sahip olup, kategoriler arasında bir sıralama bulunmamaktadır. Bu analiz türünde de bağımlı değişkenin olasılık değeri 0 ile 1 arasında değişir. Modele dahil edilen bağımsız değişken(ler) nicel veya nitel değişken olabilir [9].

2.2.3. Sıralı (ordinal) lojistik regresyon analizi

Ordinal lojistik regresyon modelinde, bağımlı değişken ikiden fazla kategoriye sahip olup, bu kategoriler arasında bir sıralama bulunmaktadır. Bu analiz türünde de bağımlı değişkenin olasılık değeri 0 ile 1 arasında değişir. Modele dahil edilen bağımsız değişken(ler) nicel yahut nitel değişken olabilir. Çoklu sıralayıcı lojistik regresyon analizinde, şıklar arasında kendine özgü bir sıralama işlemi olup, paralellik varsayımı vardır. Bu varsayımdan ötürü, β parametresi farklı şıklar ve farklı kesme noktaları için değişiklik göstermez. Bu sebeple bu yöntemde modeller arasındaki paralellik sınamasına gidilir. Bu durum, çoklu sıralayıcı lojistik regresyon analizinin en belirgin özelliğidir [6].

2.3. İki Kategorili Lojistik Regresyonda Lojit Model

Bağımlı değişkenin iki kategoriden oluştuğu durumlarda iki kategorili (binary) lojistik regresyon analizi uygulanmaktadır. Örnek verilecek olursa;

Diyelim ki x = (x_1, x_2,…,x_p) vektörü bağımsız değişkenleri göstersin.Bağımsız değişken sayısı p olsun.

y_i (i=1, 2,…,n) n tane veriden oluşan ikili gözlemlere sahip olan bağımlı değişkenimiz olsun. İki değer alan sonuç değişkeni y, olay meydana gelirse 1 aksi takdirde 0 değerini alsın.

X= *x* iken olayın olma olasılığı ya da Y = 1 olma olasılığı π ile gösterilsin.

X bağımsız değişkenlerine ilişkin olasılık modeli [1];

$$P\ (Y = 1\ |X = x_i) = \frac{e^{\beta_0 + \sum_{i=1}^{p} \beta_1 x_i}}{1 + e^{\beta_0 + \sum_{i=1}^{p} \beta_1 x_i}} = \pi\ (x_i) \quad (2.1)$$

$$P(Y=0 \mid X= x_i) = 1 - \pi\ (x_i) \quad (2.2)$$

x bağımsız değişkenler kategorik, kesikli veya sürekli değişken olabilir. Yukarıdaki denklemde lojit dönüşüm yaptıktan sonra aşağıdaki basit lojistik regresyon denklemini elde ederiz [2].

$$g(x_i) = \text{logit}\ (\pi\ (x_i)) = \ln\ [\pi\ (x_i)/(1- \pi\ (x_i))] = \beta_0 + \beta_1 x_i \quad (2.3)$$

Lojistik regresyon modeli, bağımsız değişkenin odds'u olarak aşağıdaki gibi belirtilebilir. Odds oranı bir olayın gerçekleşme olasılığının gerçekleşmeme olasılığına oranıdır. Odds oranının bir den büyük olması bağımsız değişkendeki bir birimlik artışın bağımlı değişken odds oranı kadar arttırması, birden küçük olması durumunda da odds oranı kadar azaltması anlamına gelir [6, 11].

$$\frac{\pi(x)}{1-\pi(x)} = e^{\beta_0 + \beta_1 x} \quad (2.4)$$

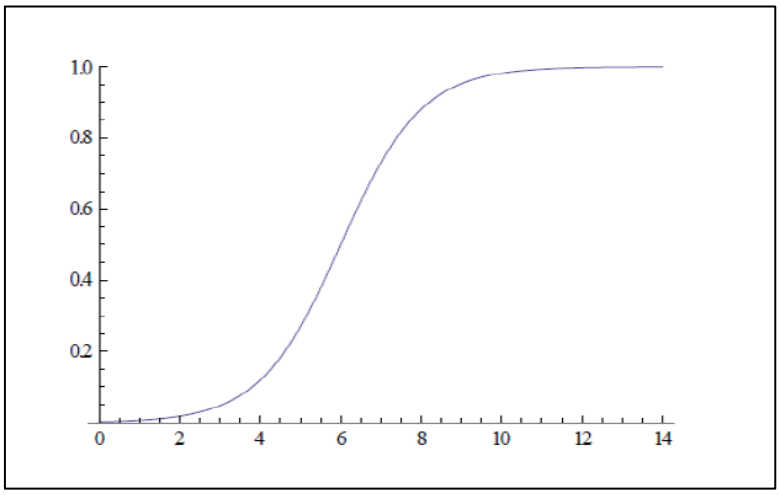
Odds'un doğal logaritması alındığında model doğrusal modele dönüşür. Bu dönüşüme lojit dönüşüm denir.

$$g(x) = \text{In}\left(\frac{\pi(x)}{1-\pi(x)}\right) = \text{In}\ e^{\beta_0 + \beta_1 x} = \beta_0 + \beta_1 x \quad (2.5)$$

Lojit dönüşüm β parametrelerinin doğrusal bir fonksiyonunu oluşturur.

Olasılık değeri 0 ile 1 arasında değişirken, x'in değerlerine bağlı olarak lojit -∞ ile +∞ arasında değer alır. Ayrıca lojistik regresyonda artıklar sıfır ortalama π (1-π) varyansla binom dağılır [7, 12].

İki durumlu lojit modelde bağımlı değişken grafiği S veya ters S şeklinde olan bir fonksiyondur. Bu fonksiyon lojistik fonksiyon olarak adlandırılır ve Şekil 4.1. deki gibi gösterilir [6,11].



Şekil 2.1. Lojistik regresyon eğrisi [11]

2.4. Regresyon Katsayılarının Kestirimi

En çok olabilirlik yöntemi (maximum likelihood method), lojistik regresyon modelinde katsayıların kestiriminde en sık kullanılan yöntemdir ve ilk kullanan Krog'dur [18]. Ayrıca minimum logit ki-kare yöntemi, ağırlıklandırılmış iteratif en küçük kareler yöntemi, diskriminant analiz yöntemi, minimum ki kare yöntemi, de katsayı kestirimlerinde kullanılan yöntemlerden bazılarıdır [13].

Örneklem genişliği büyüdükçe, en çok olabilirlik yöntemi ile ağırlıklandırılmış tekrarlayan en küçük kareler yöntemi kestiricilerinin dağılım özellikleri birbirine benzemektedir. Bağımsız değişkenlerin normal dağılım gösterdiği durumlarda diskriminant fonksiyon kestirim yöntemi kullanılır. Bağımsız değişkenler arasında kesikli değişkenlere sıkça rastlandığından normallik varsayımı çoğu zaman sağlanamamakta ve dolayısıyla katsayılar olduğundan daha büyük kestirilmektedir. Regresyon katsayılarının kestiriminde birçok yöntem olmasına karşın en etkin, yansız ve tutarlı kestirimler, en çok olabilirlik yöntemi ile yapılmaktadır [6, 14].

2.4.1. En çok olabilirlik yöntemi

En çok olabilirlik yöntemi, modeldeki katsayıların kestirimi için kullanılan yöntemlerden en sık kullanılanıdır. En çok olabilirlik tekniği, gözlenen bağımlı değişken verilerini elde etme olasılığını maksimum yapmaya dayanan, bu sayede denkleme ait parametreleri tahmin etmeye yarayan bir tekniktir [9].

En çok olabilirlik tahmin edicileri ve yeterli istatistikler arasında yakın bir ilişki vardır. Şöyle ki y parametresi için tek yeterli bir istatistik varsa, y 'nin en çok olabilirlik tahmin edicisi o yeterli istatistiğin bir fonksiyonudur [9].

N birimden oluşan x_i bağımsız ve y iki kategorili bağımlı değişken olduğunu varsayalım. En çok olabilirlik yönteminin uygulanabilmesi için öncelikle olabilirlik fonksiyonu oluşturulur. Olabilirlik fonksiyon bilinmeyen parametrelerin bir fonksiyonu olarak gözlenen değerlerin olasılığını (olabilirlik) ifade eder. Parametrelerin en çok olabilirlik tahmin edicileri, bu fonksiyonun maksimum değerleridir. Bu sayede gözlenen değerlere en yakın tahminler elde edilir [2].

Örneklem büyüklüğü arttıkça en çok olabilirlik kestiricisi tutarlıdır, yeterlidir ve asimptotik olarak normal dağılır [6].

En çok olabilirlik yöntemi kullanıldığında değişken sayısının en az 10 katı kadar gözlem sayısının olması ve bağımsız değişkenler arasında çoklu bağlantı sorununun olması durumunda da büyük örneklem üzerinde çalışılması önerilmektedir [6, 15].

Bağımlı değişkenin (y) 0 ve 1 olarak kodlandığını kabul ettiğimizde $\pi(x)$, verilen x değeri için y'nin 1'e eşit olma koşullu olasılığını, 1- $\pi(x)$'de y'nin 0'a eşit olma koşullu olasılığını vermektedir.

$$\pi(x) = P(y=1/x) \tag{2.6}$$

$$1 - \pi(x) = P(y=0/x) \tag{2.7}$$

$$P(x_i) = \pi(x)^{yi} (1 - \pi(x))^{1-yi} \tag{2.8}$$

Olabilirlik fonksiyonu bu terimlerin çarpılmasıyla elde edilirken gözlemlerin birbirinden bağımsız oldukları varsayılmaktadır.

$$\beta = \prod(x_i) \tag{2.9}$$

Matematiksel olarak eşitliğin logaritmasıyla çalışmak daha kolay olacağından log-olabilirlik fonksiyonu aşağıdaki gibi elde edilir [16].

$$L(\beta) = \text{In}\,(l(\beta)) = \sum[yi\ ln\big(\pi(\text{xi})\big) + (1 - yi)\ln(1 - \pi(\text{xi}))] \tag{2.10}$$

Bu logaritmik fonksiyon ile β parametre tahminleri elde edilir. L(β)'yı maksimum yapan β değeri için, L(β)'nin β_0 ve β_1'e göre türevini alır ve 0'a eşitleriz. Böylelikle aşağıdaki eşitlikleri elde ederiz [17, 18].

$$\sum_{i=1}^{n} |yi - \pi(\text{xi})| = 0 \tag{2.11}$$

$$\sum_{i=1}^{n} xi|yi - \pi(\text{xi})| = 0 \tag{2.12}$$

Bu eşitliklere olabilirlik eşitlikleri denir. Yukarıdaki denklemlerden elde edilen β'nın değeri en çok olabilirlik tahmini olarak adlandırılır ve $\hat{\beta}$ olarak gösterilir [3, 17].

En çok olabilirlik tahmin yönteminin en çok kullanılan yöntem olmasının sebebi, büyük örneklemde yapılan tahminlerin yüksek olmasıdır. En çok olabilirlik tahmin edicileri yeterli, asimtotik etkin, asimtotik yansız ve asimtotik normaldir. Tahminlerin örnekleme dağılımının, büyük örneklemlerde normale yakın olması sayesinde güven aralıkları oluşturulması konusunda normal ve χ^2 dağılımlarının kullanılmasına sağlar. [19]

2.4.2. Yeniden ağırlıklandırılmış iteratif en küçük kareler yöntemi

Bu yöntem, hata terimlerine ilişkin varyansların eşit olmadığı zamanlarda, değişen varyanslılık durumu söz konusu olduğunda doğru tahminde bulunulmasını sağlar.

Gruplandırılmış verilerde j grubun her birinde n_j denemeden r_j başarı elde edilsin. Başarı oranı,

$P_j = \frac{r_j}{n_j}$ olsun. (2.13)

$\text{var}\left(r_j/n_j\right) = p_j = (1 - p_j/n_j)$ (2.14)

Her binom dağılımlı gözlem için varyans değişmektedir. Bu durumda lojit(r_j/n_j)'nin açıklayıcı değişkenler üzerinde, $w_j = n_j / p_j (1 - p_j)$ ağırlığı ile ağırlıklandırılmış regresyonu uygulanmalıdır. Fakat w_j ağırlık değerleri de Pj'nin bir fonksiyonu olduğu için en küçük kareler yöntemi iteratif olarak uygulanarak ağırlık değerleri her adımda yeniden elde edilmek suretiyle çözüme ulaşılır [3, 16].

2.4.3. Minimum lojit ki-kare yöntemi

Bu teknikte yeniden ağırlıklandırılmış iteratif en küçük kareler kestirim yönteminin özel bir biçimidir [24]. Berkson'un geliştirdiği bu yöntemde, 2xj çapraz tablolarından elde edilen beklenen logit değerler ile gözlenen logit değerler arasındaki farktan yararlanılmaktadır. Bu teknik lojistik regresyon analizinde tekrarlı veriler olması durumunda kullanılır [17, 20].

Yeniden ağırlıklandırılmış iteratif en küçük kareler kestirim yönteminin özel bir şeklidir. Yeniden ağırlıklandırılmış iteratif en küçük kareler yönteminde verilen $\hat{\pi}$ olasılığı üzerinden yapılan lojit dönüşümü, bu teknikte bağımlı değişkeni vermektedir. Yöntem lojit değeri olan bağımlı değişkenin ağırlıklandırılmış regresyonundan en küçük kareler kestirimlerini bulmayı sağlar. Buradan bulunan ağırlıklı en küçük kareler kestirimleri minimum lojit ki-kare kestirimleri olmaktadır [17, 21].

2.5. Katsayıların Önemliliğinin Test Edilmesi

Model katsayısı kestiriminden sonra kurulan modeldeki bağımsız değişkenlerin modeldeki varlığının öneminin test edilmesi gerekmektedir. Modele dahil edilen bir bağımsız değişkenin, bağımlı değişkeni tahmin etme açısından modele dahil edilmediği duruma göre daha iyi kestirimde bulunulma durumu incelenir. [6, 15, 22]

Katsayıların önemi olabilirlik oran testi, wald testi ve skor testi yöntemleriyle incelenmektedir [2].

2.5.1. Olabilirlik oran testi

Olabilirlik fonksiyonunun kullanıldığı bu test iki değerin birbiriyle oranlanmasına dayanmaktadır. Üzerinde durulan modelin olabilirliği, yalnızca gerekli ve mümkün değişkenleri içeren modeli ifade eder. Doymuş modelin olabilirliği ise modelde yer alan tüm değişkenlerin sayısı kadar parametre içeren model olup doymuş model olarak da adlandırılır [20].

$$D = -2\text{Ln}\left[\frac{\text{Üzerinde durulan (şuanki) modelin olabilirliği}}{\text{Doymuş modelin olabilirliği}}\right] \text{[7]} \qquad (2.15)$$

Şuan ki model yalnızca gerekli ve mümkün değişkenleri içeren modeli ifade eder. Doymuş model ise modelde yer alan tüm değişkenlerin sayısı kadar parametre içeren modeldir.

Parantezin içindeki ifade olabilirlik oranıdır. Dağılımı bilinen bir değer elde etmek için (-2ln) katı alınır ve bu eşitliği olabilirlik oran testi denilmektedir [17].

Elde edilen test değerinin küçük olması, modele eklenen değişkenlerin lojitin kestiriminde önemli bir gelişme/katkı sağlamadığını ve değişkenlerin modelde bulunmasına gerek olmadığını göstermektedir. Bu test işlemi modeli uydurmak için gözlem sayısı (n) yeterince büyük olduğunda geçerli olmaktadır [6].

2.5.2. Wald testi

Wald istatistiği, eğim parametresi β'nın anlamlılığına ilişkin bir ölçüdür. Her bir değişken için ayrı hesaplanır ve değişkenlerin modele olan katkısını gösterir. Wald testi, β_j eğim parametresinin en çok olabilirlik kestirimi ile bu kestirimin standart hatasının oranından oluşur. Wald testi, eğim parametresinin en çok olabilirlik kestiriminin, onun standart hatasına oranı ile yapılır. Elde edilen oran, $\beta_j=0$ hipotezi altında standart normal dağılımı gösterir[6, 23].

$$W = \frac{\beta j}{SE\ (\beta j)} \qquad (2.16)$$

Eğim parametresini gösteren $\beta_1=0$ hipotezi için W istatistiği standart normal dağılım göstermektedir. Normal rassal bir değişkenin karesinin alınması 1 serbestlik dereceli ki-

kare rassal değişkenine eşit olacağından Wald istatistiği (2.17)'deki gibi de ifade edilebilir [6].

$$W^2=\{\frac{\beta j}{SE\ (\beta j)}\}^2 \sim x^2 \quad (2.17)$$

Wald test istatistiği ile modeldeki bağımsız değişkenlerin her birinin ayrı ayrı anlamlı olup olmadıkları belirlenmektedir [2].

2.5.3. Skor (score) testi

Katsayıların önemliliğini test eden diğer bir yöntem ise skor testidir. Skor test istatistiği standart normal dağılıma uymaktadır. Skor testi, log olabilirliğin türevlerinin dağılım teorisine bağlıdır. Skor testi aslında matris hesapları gerektiren çok değişkenli bir testtir. [1, 2]

$$ST = \sum_{i=1}^{n} x_i(y_i - y) \,/\, \sqrt{y\ (1-y) \sum_{i=1}^{n}(x_i - x_{ort})^2} \quad (2.18)$$

Bazı paket programlarında adımsal yöntemlerde modele alınacak ya da modelden çıkarılacak bağımsız değişkenleri belirlemek için kullanılmaktadır. [17]

2.6. Çoklu Lojistik Regresyon Analizi

Model seçimi yapılırken çalışmanın amacına doğrultusunda hareket edilmelidir. Çoğuı çalışmada, değişkenlerden hangilerinin bağımlı değişkeni etkilediği belirlemek istenilmektedir. Burada amaç hangi, değişken ya da değişkenlerin bağımlı değişkeni etkilediğini bulmaktır [6].

X'=(x_1, x_2, ..., x_k) vektörü ile gösterilen k tane bağımsız değişken olsun. Yanıt değişkeni için P(Y=1/x) = π (x) yazılsın. Çok değişkenli lojistik regresyon modelinin lojiti denklemi aşağıdaki gibi olmaktadır; [16]

$$\pi\ (x) =. P(Y=1|x) = \frac{e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k)}}{1+\ e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k)}} \quad (2.19)$$

Lojistik regresyon modeli bağımsız değişkenin oddss'u türünden aşağıdaki gibi belirtilir [6];

$$\frac{\pi\,(x)}{1-\pi\,(x)} = e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k)} \tag{2.20}$$

Odds'un doğal logaritması alınarak lojit dönüşüm yapılır[6].;

$$\text{Lojit } \pi\,(x) = \ln\left(\frac{\pi\,(x)}{1-\pi\,(x)}\right) \tag{2.21}$$

Odds'un doğal logaritması alındığında model doğrusal modele dönüşmektedir [6].

$$g\,(x) = \ln\left(\frac{\pi\,(x)}{1-\pi\,(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \tag{2.22}$$

Bağımsız değişkenler arasında kategorik değişken var ise bu değişkenler sürekli değişken gibi modele dahil edilmeyip, değişkenlerin farklı düzeylerini göstermek için göstermelik (dummy, kukla) değişken kullanılır [2].

Çoklu lojistik regresyon analizinde, en az değişkenle en iyi uyuma sahip model elde edilmek istendiğinden, adımsal yöntemler kullanılabildiği gibi bağımlı değişken ile bağımsız değişken arasındaki farklılıklar ki kare testi, iki ortalama arasındaki farklın önemlilik testi veya mann whitney u testi ile incelenebilir veya öncelikle tek değişkenli lojistik regresyon analizi uygulanıp p değeri 0,25'in altında olanlar çoklu lojistik regresyon analizine dahil edilebilir [6].

2.6.1. Adımsal yöntemler

Çoklu lojistik regresyon analizde hangi değişkenlerin modele dahil edileceği, farklı modeller denenerek en uygun modelin seçileceği aşağıda verilen çeşitli yöntemlerde bulunmaktadır.

Enter Yöntemi: Tüm değişkenlerin tek aşamada modele dahil edildiği yöntemdir.

Forward (Wald): Değişkenler modele alınırken skor istatistiğine göre, çıkarılırken de Wald istatistiğine göre karar verilmektedir. İleriye doğru adımsal bir seçim yöntemidir.

Forward (Conditional): Değişkenler modele alınırken skor istatistiğine göre, çıkarılırken de koşullu parametre tahminlerine dayanan olabilirlik oranına göre karar verilmektedir. İleriye doğru adımsal bir seçim yöntemdir.

Forward (LR): Değişkenler modele alınırken skor istatistiğine göre, çıkarılırken de kısmi maksimum olabilirlik tahminlerine dayanan olabilirlik oranına göre karar verilmektedir. İleriye doğru adımsal bir seçim yöntemdir.

Backward (Wald): Önce tüm değişkenler modele dahil edilir. Çıkarılacak değişkenlere Wald istatistiği karar verilmektedir. Geriye doğru adımsal bir seçim yöntemdir.

Backward (Conditional): Önce tüm değişkenler modele dahil edilir. Çıkarılacak değişkenler koşullu parametre tahminlerine dayanan olabilirlik oranına göre karar verilmektedir. Geriye doğru adımsal bir seçim yöntemdir [6].

Backward (LR): Önce tüm değişkenler modele dahil edilir. Çıkarılacak değişkenler kısmi maksimum olabilirlik tahminlerine dayanan olabilirlik oranına göre karar verilmektedir. Geriye doğru adımsal bir seçim yöntemdir [21].

3. LOJİSTİK REGRESYONDA MODEL UYUM İYİLİĞİNİN DEĞERLENDİRİLMESİ

Lojistik regresyon analizinde bağımlı değişkeni en iyi kestirimde bulunan bağımsız değişkenlerle model kurulduktan sonra modelin veriye uyumunu değerlendirmek amacıyla çeşitli yöntemler kullanılmaktadır.

Modelin uyumunu değerlendirmeye başlamadan önce model oluşturma aşamasındaki bağımsız değişken seçiminin yeterli-tatmin edici olduğu varsayımını kabul ederiz. Bu varsayımlar;

1. Modele girmesi gereken değişkenlerin dahil edilmiş olması
2. Değişkenlerin modele doğru fonksiyonel formda dahil edilmiş olmalarıdır.

Bu varsayımlar sağlandığında, modelin sonuç değişkenini ne kadar efektif olarak tanımladığını öğrenmek isteriz ki bu da modelin uyumu iyiliği ile değerlendirilir.

Gözlenen değerleri vektör formunda y olarak belirttiğimizde;

$$y_i = y_1, y_2, y_3, \ldots, y_n \tag{3.1}$$

Model tarafından tahmin edilen değerleri ŷ[2] ,

$$\hat{y_i} = \widehat{y_1}, \widehat{y_2}, \widehat{y_3}, \ldots, \widehat{y_n} \tag{3.2}$$

Modelin uygun olduğu sonucunu elde etmek için;

1. 1.y_i ve $\hat{y_i}$ arasındaki özet ölçümlerin uzaklıkları küçük olmalı,
2. Her bir çiftin (y_i,$\hat{y_i}$) = 1,2,3,…,n bu özet ölçümlere katkısı, sistematik olmamalı aynı zamanda modelin hata yapısına göre küçük olmalıdır [2].

Bu nedenle, model uyumunun eksiksiz değerlendirilmesi, y ve ŷ arasındaki özet ölçümlerin uzaklıklarının hesaplanmasını ve de bu ölçümlerin komponentlerinin bireysel değerlendirilmesini içerir. Model uyumunun değerlendirilmesinde logical adımlar [2];

1. Tüm uyum ölçümlerin hesaplanması ve değerlendirilmesi
2. Özet istatistiklerin bireysel bileşimlerinin incelenmesi (sıklıkla grafiksel olarak) y ve ŷ bileşenleri arasındaki uzaklığın ya da farklılığın diğer ölçümlerinin incelenmesi

Model uyumunun değerlendirilirken, bütün olarak modelin uyumunun değerlendirilmesi yanı sıra gözlemlerin değişkenlerin model uyumu açısından tek tek incelenmesi de önemlidir.

Ayrıca uyum iyiliği kovaryantın hepsiyle değil, modeldeki kovaryantlarla belirlenir. Örneğin $x'=(x_1,x_2,...,x_k)$ bağımsız değişken vektörü ve J gözlenen x 'in farklı değerlerinin sayısını göstersin. mj denek sayısını göstersin, j=1,2,..,,J [2].

1. Aynı kovaryant değerine sahip denekler analize dahil edilmiş ise, kovaryant sayısı (J) < toplam denek sayısı (n) olur, (j m değeri büyük olma eğiliminde).
2. Bağımsız değişkenlerin değerleri her bir denek için farklı ölçüm değerine sahipse J=n dir [2].

Özellikle modelde sürekli değişkenler yer alıyorsa J≈n olması beklenilir [16].

Uyum iyiliği testinde kullanılan farklı istatistikler vardır. Bunlar aşağıda açıklanmıştır ve bu istatistiklerde J=n olduğu varsayılmıştır [2].

3.1. Pearson Ki-Kare ve Sapma İstatistikleri

Gözlenen değerler ve model kullanılarak elde edilen (beklenen) değerler karşılaştırılarak modelin uyumu değerlendirilebilir. Aynı zaman da gözlenen ve beklenen değerler arasındaki farklar (yani artıklar) kullanılarak uyum değerlendirilir [6]. Pearson artıkları ve sapma artıkları kullanılır $(y_i - \hat{y}_i)$. Her iki istatistikten elde edilen küçük p değerleri model uyumunun kötü olduğunu gösterir.

Pearson artıklarının toplamı pearson ki-kare istatistiğinin, sapma artıklarının toplamı da sapma istatistiğini vermektedir [6].

Lojistik regresyon analizinin varsayımlarından bir tanesi de hataların bağımsızlığıdır. Lojistik regresyonda, gözlenen varyansın beklenen varyanstan büyük olması aşırı yayılım olarak tanımlanmaktadır. Aşırı yayılım, standart hataların küçülme eğiliminde olmasından ve regresyon modelindeki yordayıcılara ilişkin test istatistiklerinin güven aralıklarını küçültmesine sebebiyet vermesinden dolayı soruna sebep olur. Hataların bağımsızlığı varsayımı için aşırı yayılımın tespiti; uyum iyiliği değerlerindeki Pearson veya Sapma (daha küçük bir dağılım parametresi üretecek olan tercih edilir) ki kare istatistiğinin kendi serbestlik derecesine oranının hesaplanmasıyla bulunur. Bu oran dağılım parametresi (ϕ) olarak adlandırılır.

Gözlenen ve beklenen değer arasında ki lojistik regresyon farkını ölçmek için kullanılacak birkaç yol vardır. Tahmin değerler her bir bağımsız değişken deseni için hesaplanır.

Beklenen değer(tahmini) $\hat{y}_j$ olarak gösterilsin.

$$\widehat{y_j} = m_j\pi_j = m_j\left(\frac{e^{\hat{g}(x_j)}}{1+e^{\hat{g}(x_j)}}\right) \tag{3.3}$$

Buradaki $\hat{g}$(xj) kestirilen lojittir. mj ise kovaryant değerleri birbirinden farklı olan denek sayısıdır.

$$\hat{g}(x_j) = \hat{\beta}_0 + \hat{\beta}_1\hat{x}_{j1} + \hat{\beta}_2\hat{x}_{j2} + \cdots + \hat{\beta}_p\hat{x}_{jp} \tag{3.4}$$

Gözlenen ve kestirilen değerler arasındaki farkı, pearson artığı, deviance artığı ve lineer regresyonda kullanılan artıkla (standartlaştırılmamış) dikkate alırız.

Pearson Artığı;

$$r\,(y_j, \widehat{\pi_j}) = \frac{y_j - m_j\hat{\pi}_j}{\sqrt{m_j\hat{\pi}_j\,(1-\hat{\pi}_j)}} \tag{3.5}$$

Bu artıklara dayalı j-p-1 serbestlik dereceli pearson ki-kare istatistiği;

$$x^2 = \sum_{j=1}^{j}\left[r\,(y_j,\ \hat{\pi}_j)\right]^2 \tag{3.6}$$

Sapma Artığı;

$$d\,(y_j, \hat{\pi}_j) = \pm 2\left(\left[y_j In\left(\frac{y_j}{m_j \hat{\pi}_j}\right) + (m_j - y_j) In\left(\frac{m_j - y_j}{m_j(1-\hat{\pi}_j)}\right)\right]\right)^{1/2} \quad (3.7)$$

$y_j = 0$ olduğunda sapma artığı $d\,(y_j, \hat{\pi}_j)$ = -2 $\sqrt{m_j \left| In\,(1 - \hat{\pi}_j)\right|}$ (3.8)

$y_j = m_j$ olduğunda ise sapma artığı; $d\,(y_j, \hat{\pi}_j)$ = -2 $\sqrt{m_j \left| In\,(\hat{\pi}_j)\right|}$ (3.9)

Bu artıklara dayalı j-p-1 serbestlik dereceli sapma istatistiği;

$$D = \sum_{j=1}^{j} d\,(y_j, \hat{\pi}_j)^2 \quad (3.10)$$

H_0 = Model verileri itibariyle uygundur.
H_1 = Model verileri itibariyle uygun değildir.

Modelin doğru kurulduğu varsayımı altında, Mc Cullagh ve Nelder (1983), $x^2_{p\ ve}$ D istatistiklerinin dağılımının J – (p+1) serbestlik dereceli ki-kare dağılımı olduğunu göstermiştir. Sapma istatistiği aynı zamanda J parametreli doymuş modelin, p+1 parametreyle kurulan modele karşı olabilirlik oran test istatistiğidir [24, 25]

Burada J, bağımsız değişken deseni sayısı (Covariate patterns) ve p bağımsız değişken sayısıdır.

3.2. Hosmer-Lemeshow Testleri

Halteman, Hosmer ve Lemoshow, Shillington, Tsiatis özellikleri hem teorik olarak hem de bilgisayar benzetim teknikleriyle incelenmeye müsait olan ki-kare benzeri uyum iyiliği istatistiklerinin geliştirilmesi üzerine çalışmalar yapmışlardır [2, 24]

Hosmer ve Lemeshow ki-kare uyum iyiliği testi olarak da anılan test, lojistik regresyon modelinin uyumunu her bir bağımsız değişken için değil bir bütün olarak değerlendirmektedir. Özellikle bağımsız değişkenlerin sürekli değişkenler olduğu durumda veya küçük örneklemlerle çalışıldığı durumda, geleneksel ki-kare testinden çok daha

güçlüdür. Bu anlamda da geleneksel ki-kare yöntemi ile hesaplanan Omnibus Testi'nin daha güçlü bir alternatifidir. Bu teste ilişkin sonucun anlamlı olmaması (p>0,05), model-veri uyumunun yeterli düzeyde olduğunu gösterir. Gözlenen ve model tarafından kestirilen değerler arasında anlamlı fark yoktur; yani model tahminleri, gözlenen durumdan farklı değildir anlamına gelmektedir [2].

Kurulan modelin uyum iyiliği testi Hosmer-Lemeshow'un hem olasılıkların persentillerine göre hem de sabit kesim noktası yöntemine göre hesaplanmaktadır [2].

Tahmini olasılık değerlerine dayanarak gruplamayı önerir. Lojistik regresyon analizinden kestirilen olasılıklar kullanılır.

Kestirilen olasılıklar küçükten büyüğe dizilir ve n tane gruba ayrılır. (n genellikle 10 alınır). n tane sütun n tane tahmini değere karşı gelir. 1.sütun en küçük değeri n. sütun ise en büyük değere karşılık gelir.

2 çeşit gruplama stratejisi önerilir;

$\hat{C}_g^*$; Sabit denek sayılı göre gruplama

$\hat{H}_g^*$; Sabit olasılıklara göre gruplama

3.2.1. Hosmer and Lemeshow'un $\hat{C}_g^*$ istatistiği

Pearson ki kare testinden elde edilen gözlemlenen ve beklenen değerler hesaplanır:

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_{1k} - \hat{e}_{1k})^2}{\hat{e}_{1k}} - \frac{(o_{0k} - \hat{e}_{0k})^2}{\hat{e}_{0k}} \quad (3.11)$$

Bu formülü sağlaması için sağlanan şartlar şu şekilde:

$$o_{1k} = \sum_{j=1}^{c_k} y_j$$

$$o_{0k} = \sum_{j=1}^{c_k} (m_j - y_j)$$

$$\hat{e}_{1k} = \sum_{j=1}^{c_k} m_j \hat{\pi}_j$$

$$\hat{e}_{1k} = \sum_{j=1}^{c_k} m_j(1 - \hat{\pi}_j) \quad (3.12)$$

C_k ,k. Gruptaki bağımsız değişken sayısı.

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_{1k} - n_k' \bar{\pi}_k)^2}{n_k' \bar{\pi}_k(1 - \bar{\pi}_k)} \quad (3.13)$$

$\bar{\pi}_k$ ise k. grupta ki ortalama tahmini olasılığıdır.

$$\bar{\pi}_k = \frac{1}{n'_k} \sum_{j=1}^{c_k} m_j \hat{\pi}_j \quad (3.14)$$

Hosmer-Lemeshow «Ĉ» istatistiği, t-2 serbestlik dereceli ki-kare dağılımı göstermektedir.

$\hat{C}_g^*$ istatistiği bütün denekleri kullanırken X^2 istatistiğinde sadece olayın gerçekleşmiş olduğu denekler kullanır [22].

Hosmer ve lemeshow bağımsız değişken sayısının bir fazlasının grup sayısından az olduğunda (p+1)<g, $\hat{C}_g^*$ istatistiğinin (g-2) serbestlik derecesiyle ki-kare dağılımına sahip olacağını göstermişlerdir. Bundan dolayı eğer p+1<10 ise 8 (10-2=8) serbestlik derecesiyle ki-kare dağılır [4].

Onlu risk metodu en sık kullanılan yöntemlerdendir. Bu metotta 10 grup kullanılırken ilk grupta en küçük olasılığa sahip n/10 denek, en son grupta ise en büyük olasılığa sahip n/10 denek bulunur [24].

Uyum iyiliğine karar vermeden önce verilerdeki kayıp ve uç değerler incelenmeli ve gereken düzenlemeler yapılmalıdır. Bu durum, ilgili analizin bir varsayımı olarak düşünülmese de, kayıp ve uç değerlerin istatistiksel testlerin sonuçlarını bozabileceği düşünüldüğünden analiz öncesinde dikkate alınması gerekmektedir. Tüm kategorik değişken çiftleri için tüm hücrelerde beklenen frekans 1'den büyük olmalı ve beklenen frekansın 5'ten küçük olduğu gözenek sayısı %20'yi geçmemelidir [2].

Uyumu değerlendirme de avantajları;

1. Tek değer hesaplanması
2. Kolay yorumlanabilir olmasıdır.

Uyumu değerlendirmede dezavantajı;

Gruplama, az sayıda bireysel noktalar nedeniyle uyumdan önemli sapmaların gözden kaçmasına neden olabilir.

Çizelge 3.1. Sabit denek sayılı sınıflamaya göre oluşturulan risk grupları için gözlenen ve beklenen frekanslar

Risk grupları	Kesim noktası	Durum=1		Durum=0		Toplam
		Gözlemlenen değer	Beklenen değer	Gözlemlenen değer	Beklenen değer	
1	c1	g_{11}	b_{11}	g_{01}	b_{01}	(n/10)
2	c2	g_{12}	b_{12}	g_{02}	b_{02}	(n/10)
3	c3	g_{13}	b_{13}	g_{03}	b_{03}	(n/10)
4	c4	g_{14}	b_{14}	g_{04}	b_{04}	(n/10)
5	c5	g_{15}	b_{15}	g_{05}	b_{05}	(n/10)
6	c6	g_{16}	b_{16}	g_{06}	b_{06}	(n/10)
7	c7	g_{17}	b_{17}	g_{07}	b_{07}	(n/10)
8	c8	g_{18}	b_{18}	g_{08}	b_{08}	(n/10)
9	c9	g_{19}	b_{19}	g_{09}	b_{09}	(n/10)
10	c10	g_{110}	b_{110}	$g_{01}0$	b_{010}	(n/10)

Örnekte gözlem sayısı n'dir. Tablodan görüldüğü üzere n kişiyi, kestirilen olasılık değerlerine göre küçükten büyüğe doğru sıraladıktan sonra n/10'ar brimlik 10 gruba ayırdık. Örneğin ilk 50 birimden oluşan grubun kestirilen olasılık kesim noktası c1'dir.

Örneğin 6. Risk grubunu ele alırsak, bu gruptaki deneklerin olasılıkları $0.208< \pi_j \leq 0.249$ arasındadır. Bu gruptaki n kişinin g_{16}'sı gerçekte durum=1 grubunda g_{06} (n/10- g_{16}) durum=0 grubundadır. b_{16}'sı beklenen değeri durum =1 iken b_{06} 'sı (n/10- b_{16}) durum=0 'dır. Uyum iyiliğine karar vermek için, Hosmer-Lemeshow «Ĉ» istatistiği hesaplanır.

Bu «Ĉ» istatistik değerine karşılık gelen p değeri hesaplanır. Sıfır hipotezi istatistik model uyumunun iyi olduğu şeklinde kurulmuştur. Bu sebeple p değerinin 0,05'ten

büyük olması istenir. Bu örnekte p>0,05 olması durumunda modelin veri setine uyduğu söylenebilir [2].

3.2.2. Hosmer ve lemeshow'un $\hat{H}_g^*$ istatistiği

Hosmer ve Lemoshow'un önerdiği diğer bir alternatif yöntem ise $\hat{H}_g^*$ istatistiğidir. $\hat{C}_g^*$ istatistiği gibi hesaplanır. Tek farkı olasılıklar sabit aralıklı n gruba ayrılır. Bu metotta k/10, k=1,2,…,9 değerleriyle tanımlanan kesim noktaları kullanılarak 10 grup oluşturulur ve gruplar kestirilen olasılıkları ard arda gelen kesim noktaları arasına düşen denekleri içerir [28]. $0 \leq P(x_i) \leq 1$ olduğundan t=10 olarak alınırsa her bir risk grubunun genişliği 1/t=1/10=0,1 olarak alınır ve 10 tane grup oluşturulur. Ancak burada grup sayısı denek sayısına göre değişkenlik gösterebilir.

Çizelge 3.2. Olasılıkların persentillerine göre hesaplanan onlu risk grupları için ve beklenen frekanslar

Risk grupları	Kesim noktası	Durum=1		Durum=0		Toplam
		Gözlemlenen değer	Beklenen değer	Gözlemlenen değer	Beklenen değer	
1	0-0,1	g_{11}	b_{11}	g_{01}	b_{01}	n_1
2	0,11-0,2	g_{12}	b_{12}	g_{02}	b_{02}	n_2
3	0,21-0,3	g_{13}	b_{13}	g_{03}	b_{03}	n_3
4	0,31-0,4	g_{14}	b_{14}	g_{04}	b_{04}	n_4
5	0,41-0,5	g_{15}	b_{15}	g_{05}	b_{05}	n_5
6	0,51-0,6	g_{16}	b_{16}	g_{06}	b_{06}	n_6
7	0,61-0,7	g_{17}	b_{17}	g_{07}	b_{07}	n_7
8	0,71-0,8	g_{18}	b_{18}	g_{08}	b_{08}	n_8
9	0,81-0,9	g_{19}	b_{19}	g_{09}	b_{09}	n_9
10	0,91-1,0	g_{110}	b_{110}	g_{01}	b_{010}	n_{10}

Bu $\hat{H}_g$ istatistik değerine karşılık gelen p değeri hesaplanır. Sıfır hipotezi istatistik model uyumunun iyi olduğu şeklinde kurulmuştur. Bu sebeple p değerinin 0,05'ten büyük olması istenir. Bu örnekte p>0,05 olması durumunda modelin veri setine uyduğu söylenebilir.

Eğer bir risk grubunda gözlemlenen ile beklenen değer arasındaki farkın büyük olduğu gözlemlenmişse öncelikle model uyumunda istenilen sonucu elde etmemiz için bu durumun standartlaştırılması gerekmektedir.

Tahmin ettiğimiz beklenen frekansın büyük olması durumunda standart normal dağılım uygulanmaktadır. Bu durumu anlamak içinde 1,96 sayısı referans olarak alınır. Gözlemlenen ve beklenen değer arasındaki farkın 1,96 dan büyük olması önemli bir fark olarak ele alınmaktadır.

Göze sayısı 5'ten az olduğu taktirde var olan uyum iyiliği testlerinde hatalar gözlemlenmektedir ve bu durum iyi sonuç elde edilememesine yol açar. Bu nedenle verilerin normalleştirilmesi gerekmektedir.

Bu aşamada pearson ki kare ile düzeltme yapılır. Z değerine göre de modelin veriyi açıklayıp açıklamadığı test edilir. Hosmer Lmeshow başka bir yaklaşım önermektedir. Hosmer normalde veri setini 10'a bölerek yaparken bu sefer de 7-8 e bölünmüş ve kıyaslanma bu şekilde yapılmıştır.

J bağımsız değişken olarak aldığımızda bu yaklaşımda adımlar şu şekilde [2];

1. Modelden olasılık değerleri alınıyor. $\hat{\pi}_j$, j = 1, 2, 3, …, J

2. u_j değişkeni oluşturuluyor. $u_j = m_j\hat{\pi}_j\ (1 - \hat{\pi}_j\ )$, j = 1, 2, 3, …, J

3. c_j değişkeni oluşturuluyor. $c_j = \frac{(1-2\hat{\pi}_j\ )}{u_j}$, j=1, 2, 3,…, J

4. Pearson ki kare istatistiği değeri hesaplanır;

$$X^2 = \sum_{j=1}^{j} \frac{(y_j - m_j\hat{\pi}_j)^2}{u_j} \qquad (3.15)$$

5. bağımsız değişken x üzerinde c ağırlığı kullanılacaksa adım 3ten, u ağırlığı kullanılacaksa da adım 2 den yararlanılır. Bu regresyonda bağımsız değişken sayısı olarak aldığımız J'nin örneklem büyüklüğü not edilir. Regresyonun hata terimler toplamını RSS ile gösterelim. Bazı paket programlarında örneğin STATA gibi, özet ölçümler ağırlıklandırılır. Hata kareler toplamı, düzeltilmiş RSS den elde edilen ağırlıklandırmalar ile çarpılmalıdır.

6. Uygunluğu sağlamak amacıyla varyans için düzeltme faktörü hesaplanır;

$$A = 2(J - \sum_{j=1}^{j} \frac{1}{m_j}) \tag{3.16}$$

7. Standartlaştıracak olursak;

$$Z_{x^2} = \frac{[x^2-(J-p-1)]}{\sqrt{A+RSS}} \tag{3.17}$$

8. Standart normal dağılım kullanarak p değeri hesaplanır.

Yukarıda bahsettiğimiz analizi yürütmek için toplu veri kümesi oluşturmak gerekmektedir. Bu durumu bazı yazılım programları ile yapmak kolaydır ancak programlar harici yapmak imkansız görünmektedir.

Son durumda; bir veri setimizi oluşturulurken diğer bir yandan da bu yeni verilerle lojistik regresyona dönüşüm yapılır.

Herhangi paket programda adımlar şu şekildedir;

i. Modelde değişkenlerin temel etkilerini tanımlayın. Bağımsız değişkenler tanımlanır.
ii. Bağımlı değişkenin toplam değeri hesaplanır. Her bağımsız değişken için y_j ve m_j tanımlanır.
iii. Bağımsız değişkenler, y_j ve m_j yeni bir veri seti halinde sunulur. Standardize edilmiş veri setine normal dağılım uygulanır.

$$z_s = \frac{(S-\sum_{j=1}^{j} m_{j\hat{\pi}_j}(1-\hat{\pi}_j))}{\sqrt{A+RSS^*}} \tag{3.18}$$

A düzeltme faktörü adım 6 da tanımlandı. RSS*, $d_j = (1 - 2\hat{\pi}_j)$ lineer regresyon modelinde hata kareler toplamıdır ve ağırlık $u_j = m_j\, \hat{\pi}_j(1- \hat{\pi}_j)$ olarak belirlenmiştir [2].

Hosmer (1997) normalize edilmiş pearson ki kare de, normalize edilmiş Z_s ve Z_{x^2} karşılaştırılmış ve bunun sonucunda iki istatistik değerinin çoğu durumda benzer durum sergilediği gözlemlenmiştir [2].

Ancak çok küçük ve çok büyük olasılık durumları istisna olarak belirtilmiştir.(Örnek olarak: $\hat{\pi}_j$, nin büyük olma durumunda y=0 ya da $\hat{\pi}$ nin küçük olma durumunda y=1) Bu gibi durumlarda, pearson sonucunda oldukça büyük X^2 değeri elde edilir. Aslında tek bir pearson değeri bile uygun durumu reddetmekte yeterli olabilir. Bir diğer yan etkisi de bu durumun tam tersidir. Yani oldukça büyük X^2 değerine varyans şişirme uygulayarak bu durumundan kurtulabiliriz. Bu ayarlamalarda da S yi kullanmak tercih edilir. Bir çok paket programında hesaplanan Z_S ve Z_{x^2} değerlerinde çok az bir fark bulunmuştur.

Weesie (1998) standart normal dağılım kullanılarak Pearson ki-kare istatistiğinin önemini hesaplamak için Windmeijer (1990) tarafından önerilen bir yöntem uygulayan bir STATA programı yazmıştır [2].

Yukarıda tarif edildiği gibi, (1988) Stukel genelleştirilmiş bir lojistik modeli parametreleri sıfıra eşit olup olmadığını test etmek için serbestlik derecesi 1 ya da 2 olan istatistik önerilmiştir [2].

Bu temel bir lojistik regresyon modeli varsayımını test etmek için bu anlamda uyum iyiliği testleri Hosmer-Lemeshow ve Osius-Rojek yardımcı olacaktır [2].

Bu test için herhangi bir paket program kullanılmamıştır ancak aşağıdaki adımlarla kolayca test edilebilir [2]:

1. Modelden olasılık değerleri alınır. $\hat{\pi}_j$, j = 1, 2, 3, …, J
2. Tahmini logaritma hesaplanır.

$$\hat{g}_j = \text{In}\left(\frac{\hat{\pi}_j}{1-\hat{\pi}_j}\right) = x_j\,\hat{\beta},\ j = 1, 2, 3,\ldots,J \qquad (3.19)$$

3. İki yeni ortak değişken hesaplanır.

$$z_{1j} = 0.5 * \hat{g}_j{}^2 * \text{I}\,(\hat{\pi}_j \geq 0.5) \qquad (3.20)$$

ve

$$z_{2j} = -0.5 * \hat{g}_j{}^2 * \text{I}\,(\hat{\pi}_j < 0.5) \text{ dir.} \qquad (3.21)$$

Tüm bu değerlerle sadece bir değişken oluşturulur 0.5'den daha az ya da daha fazla olduğu durumda not edilir [2].

4. Modele z1 ve z2 değişkeni eklenmesi için Score testi yapılır. Score testi iyi bir performans sağlamadığı taktirde kısmi olabilirlik oran testi kullanılabilir[2].

Sonuç olarak standartlaştırılmış veri, model uyumu analizini destekleyici bir yol izlemektedir. Model uyumunun yeterliliği hakkında karar veremediğimiz takdir de sınıflandırma tablolarına bakılır. Ayrıca $\hat{C}$ ve $\hat{H}$ istatistikleri Hosmer ve lemeshow (1982) tarafından yapılan bir simülasyon çalışmasında $\hat{H}$ istatistiğinin $\hat{C}$ istatistiğine göre daha etkili olduğu görülmüştür [27].

3.3. Sınıflandırma Tabloları

Bu tablo; sonuç değişkeni y ile $\hat{y}$ değerleri tahmini lojistik olasılıklardan elde edilen dikotom değişkeninin çapraz-sınıflaması ile elde edilir.

Dikotom değişken (Bir kategorik değişkenin sadece iki özelliği olması(ölü/canlı vb)) elde etmek için kesme noktası belirlenir «c» ve her bir tahmini olasılık (estimated probability) c ile karşılaştırılır [2].

1. Tahmini olasılık değeri c'den büyükse 1; küçükse 0 değerini alır.
2. c için kullanılan en sık değer 0,5'tir.
3. En son olarak beklenen değerler ile gözlenen gerçek değerlerin çapraz tablosu oluşturulur [2].

Tahmini olasılık değeri c'den büyükse 1; küçükse 0 değerini alacağını belirtmiştik. Bu yöntemle tahmini olasılık değerleri grup üyelerini belirlemek için kullanılır.

Eğer model grup üyelerini tam olarak belirlerse, bu modelin uygun olduğu söylenebilir.

Diğer bir deyişle; sınıflandırma tabloları, uyum iyiliği için kullanılan bir başka ölçüttür. Bu tablo, bağımlı değişken ile bağımsız değişkenin çapraz sınıflandırılması ile elde edilir. Sınıflandırma tablosunda, bağımlı değişkeninin gözlenen ve kestirilen lojistik

olasılıklarından türetilen (0 veya 1) değerleri yer almaktadır. Türetilen bağımlı değişken değerlerinin elde edilmesinde bir kesim değerinin (c) tanımlanması zorunludur. Kestirilen olasılık değeri c' yi aştığında türetilen bağımlı değişken 1, aksi durumda 0 değerini alacaktır.

Kesim değeri için en çok kullanılan değer 0,5' dir. Ele alınan kestirim değeri 0,5 değerini geçtiğinde 1, aksi durumda 0 grubuna atama yapılmaktadır. Bu şekilde yapılan gruplamanın dezavantajları olması kaçınılmazdır. Örneğin, kestirim değeri olarak 0,52 ile 0,48 arasında neredeyse bir fark olmadığı halde, 0,5 kesim değerine göre atama yapıldığında, neredeyse birbirine eşit olan bu iki değer farklı gruplara atanmaktadır. Diskiriminant analizinde olduğu gibi amaç sınıflandırma olduğunda sınıflandırma tablolarının kullanımı daha anlamlı olacaktır. Ancak uyum iyiliğini değerlendirme konusunda bize yeterli bilgi vermemektedir. Bu yöntemle; duyarlılık, seçicilik ve doğruluk hesaplanır [2].

Seçicilik ve duyarlılık lojistik regresyon modelinin ayırıcılık gücü üzerine ölçütlerdir.

Çizelge 3.3. Sınıflandırma tablosu [2]

		Beklenen Durum		
Gözlenen Durum		Düzey=1	Düzey=0	Toplam
	Düzey=1	22	19	41
	Düzey=0	103	356	459
	Toplam	125	375	500

Tabloya göre duyarlılık; 22/125 = 17,6% Tabloya göre seçicilik; 356/375=94,93%

Doğru sınıflama yüzdesi; 100* ((22+356)/ 500)=75,6

Pozitif Kestirim Değeri : 22/41 = 0,54

Negatif Kestirim Değeri : 356/459= 0,77

Bir çok lojistik regresyon modelinde görülen tipik bir sınıflandırma tablosu örneğimizde doğru sınıflandırma yüzdesi %75,6 olarak bulunmuştur. Tabloya göre düzey = 1 olduğunda

(bu durum duyarlılık olarak nitelenmekte), % 17,6 doğruluk payının olduğu gözlenmekte iken, gelişme durumumuz 0 olduğunda (bu durum da seçicilik olarak ifade edilmekte), % 94,93 doğruluk payı içermektedir.

Sınıflandırma 2 bileşenli gruplandırmada, grupların büyüklüğüne duyarlıdır ve genelde daha büyük olan grup lehinedir. Bu durum model uyumundan bağımsızdır.

2x2 sınıflandırma tablosundan elde edilen ölçümler (duyarlılık ve seçicilik gibi) model performansını değerlendirmede kullanılmamalıdır çünkü ağırlıklı olarak örnekteki olasılık dağılımına bağlıdır.

Bu nedenle eğer 2 model duyarlılık ve seçicilik açısından karşılaştırılacaksa bu modelin birinin diğerine üstünlüğünden çok örneklemin dağılımı ile ilgilidir.

Ayrıca kesim noktaları ve olasılık değerleri(π) ile sınıflandırma tablo yüzdelerini, gruba dahil edilme aralığını değiştirmek mümkündür. Bu durumda duyarlılık ve seçicilik değerleri de değişmektedir.

Aşağıda benzer 2 örnekte π değerlerinin farklı alınması ile birlikte değerlerin etkisi gözlemlenmiştir.

Çizelge 3.4. Sınıflandırma tablosu [2]

		Beklenen Durum		
Gözlenen Durum		Düzey=1	Düzey=0	Toplam
	Düzey=1	39	2	41
	Düzey=0	23	436	459
	Toplam	62	438	500

Bu tabloya Lojistik regresyon modeline dayalı çalışma düzeylerinde sınıflandırma modelindegözlenen değerler yukarıdaki gibi olduğunda (62, 438) kesim noktası 0,5 iken π<0,50 olanların $\hat{\pi}$ =0.05'i ve $\hat{\pi} \geq$0,5 olanların olanların $\hat{\pi}$ =0,95 ise'i düzey=1 olarak tahmin edilmiştir.

Beklenen düzey=0 olanların yani kesim noktası 0,5 iken $\hat{\pi} < 0.50$ olanların %5'i (459 x 0,05=23) yanlış tahmin edilmiştir.

Beklenen düzey=1 olanların yani kesim noktası 0,5 iken $\hat{\pi} \geq 0.5$ olanların %95'i (41 x 0.95=39) doğru tahmin edilmiştir.

Yani gözlenen değerlere baktığımızda gerçekte fracture=1 olan 62 kişinin 39'u doğru tahmin edilirken 23'ü yanlış tahmin edilmiştir.

Tabloya göre duyarlılık: 39/62=62,9% Tabloya göre seçicilik:436/438=99,5%

Pozitif Kestirim Değeri : 39/41 = 0,95

Negatif Kestirim Değeri : 436/459= 0,95

Çizelge 3.5. Sınıflandırma tablosu [2]

		Beklenen Durum		
Gözlenen Durum		Düzey=1	Düzey=0	Toplam
	Düzey=1	23	18	41
	Düzey=0	207	252	459
	Toplam	230	270	500

Bu tabloya Lojistik regresyon modeline dayalı çalışma düzeylerinde sınıflandırma modelinde gözlenen değerler yukarıdaki gibi olduğunda (230, 270) kesim noktası 0.5 iken $\hat{\pi}$ <0,50 olanların $\hat{\pi}$ =0,45'i ve $\hat{\pi}$≥0,5 olanların ise $\hat{\pi}$ =0,55'i düzey=1 olarak tahmin edilmiştir.

Beklenen düzey=0 olanların yani kesim noktası 0,5 iken $\hat{\pi} < 0.50$ olanların %45'i (459 x 0,45=207) yanlış tahmin edilmiştir.

Beklenen düzey=1 olanların yani kesim noktası 0,5 iken $\hat{\pi} \geq 0.5$ olanların %95'i (41 x 0,55=23) doğru tahmin edilmiştir.

Yani gözlenen değerlere baktığımızda gerçekte fracture=1 olan 230 kişinin 23'ü doğru tahmin edilirken 207'si yanlış tahmin edilmiştir.

Yani duyarlılık ve seçicilik gözlenen değerlerin oranına bağlı olarak değişmektedir.

Tabloya göre duyarlılık:23/230=10,0% Tabloya göre seçicilik:252/270=93,3%

Pozitif Kestirim Değeri : 23/41 = 0,56
Negatif Kestirim Değeri : 252/459= 0,55

Sonuç olarak sınıflandırma tabloları, sınıflandırma eğer analizin bir hedefi ise uygundur. Bunun dışındaki durumlarda kullanıldığında modelin uyumunu değerlendirmek sınıflandırma tablolarına ek analizlerin yapılması ihtiyacını doğuracaktır.

Sınıflandırma tablosunda kestirilen olasılıklar grup üyeliğini tahmin etmek için kullanılır;

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

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

$B_0 = \text{In}\,[Q_1/(1-Q_1)] - \mu^2/2$ ile verilir. (3.22)

Bu varsayımlar altında lojistik regresyonun hatalı sınıflandırma olasılığı (PMC),

$$\text{PMC} = Q_1\,\Phi\left\{\frac{1}{\beta_1}\,In\,\left[\frac{(1-Q_1)}{Q_1}\right] - \frac{\beta_1}{2}\right\}$$

$$+(1 - Q_1)\,\Phi\left\{\frac{1}{\beta_1}\,In\,\left[\frac{Q_1}{(1-Q_1)}\right] - \frac{\beta_1}{2}\right\} \quad (3.23)$$

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

Grup büyüklüğü arttıkça doğru sınıflandırma olasılığının güvenirliği artar. Bu sebeple sınıflandırma tablosu diğer yöntemlere yardımcı olarak kullanılabilir [11, 14]

3.4. Pseduo R^2 İstatistikleri (McFadden R^2, Cox ve Snell R^2, Nagelkerke)

Lojistik regresyon analizinde R²'nin, doğrusal regresyondaki olduğu gibi en küçük kareler yöntemi ile hesaplanması olanaksızdır. Bu nedenle yalancı (pseudo) R² istatistikleri geliştirilmiştir.

Model uyumunun değerlendirilmesinde rutin kullanımı önerilmemektedir. (Doğrusal regresyondan farklı olarak genellikle küçük değerler çıktığından değerlendirme yapılırken dikkatli olunmalıdır.) Modelin açıklayıcılığını belirtmek amacıyla kullanımı önerilmemektedir. Modeller arasında seçim yapma sürecinde karşılaştırmada kullanılmalıdırlar.

Hatta bu nedenle bazı yazarlar sonuçlar sunulurken R² değerlerinin verilmemesini önermektedir [2].

Bu istatistiklerden bazıları mcfadden R^2, cox ve snell R^2 ve nagelkarke R^2 istatistikleridir.

Düzeltilmiş McFadden R^2 istatistiğinde sadece sabit terimli modelin log-olabilirliği kareler toplamı olarak, tüm değişkenleri içeren modelin log-olabilirliği ise hata kareler toplamı olarak değerlendirilir. Burada $\hat{L}$ kestirilen olabilirliği ifade etmektedir. Olabilirlik oranı tüm değişkenleri içeren modelin, sadece sabit terimi içeren modele göre açıklayıcılık düzeyini göstermektedir [21].

Cox Ve Snell R^2, Olabilirlik esasına göre çoklu R^2 ölçüsüne benzemektedir. İstatistiğin maksimum değerinin 1'den küçük olması bu istatistiğin yorumunu güçleştirmektedir [23]

Nagelkarke R^2, Cox ve Snell R^2 istatistiğinin 0 – 1 aralığında değerler almasını sağlamak maksadıyla geliştirilmiştir. Bunu sağlamak için Cox ve Snell R2 istatistiği maksimum değeri olan 1 – $L(Ms)^2/N$'e bölünmelidir. Eğer tüm model sonucu iyi bir

şekilde tahmin eder ve olabilirliği 1'e eşit olursa, Nagelkarke R^2'de 1'e eşit olacaktır [21, 27]

3.5. Roc Eğrisi Altında Kalan Alan

ROC (Receiver Operating Characteristics Curves), ROC eğrisi olarak tanımlanmaktadır. ROC doğru pozitiflerin, yanlış pozitiflere olan bölümü olarak da ifade edilmektedir. Tıp'ta kullanım alanının yaygın olması tanı testlerinin performanslarının değerlendirilmesi ve karşılaştırılmasından kaynaklanmaktadır. Klinikte ölçülen değişken sürekli ise, kişileri hasta ve sağlıklı olarak ayrılma işlemi zorlaşır. Klinik çalışmaların koşullarına bağlı olarak incelenen tanı testinin optimum etki noktası değişmektedir [2].

Seçilen farklı eşik değerleri için bulunan farklı duyarlılık-belirleyicilik özelliklerine bağlı olarak ara seçenekler belirlenerek, ROC eğrileri (Receiver operating characteristic curves) oluşturulmuştur. ROC eğrisi yöntemi; testin ayırt etme gücünün belirlenmesinde, testlerin etkinliklerinin karşılaştırılmasına, uygun pozitiflik eşiğinin belirlenmesine, laboratuar sonuçlarının kalitesinin izlenmesinde kullanılabilmektedir [2].

Tıbbi araştırmalarda, ROC eğrileri, tanısal testlerin doğruluğunu analiz etmeyi sağlar. Ayrıca tanısal testlerde pozitif ve negatif test sonuçları arasında ayrım için en iyi eşik veya "cut-off" değerini belirlemek için kullanılmaktadır. Tanısal test hemen hemen her zaman duyarlılık (sensitivity) ve belirleyicilik (specificity) arasında önemli bir denge yöntemidir. ROC eğrileri bu dengeyi açıklayan bir grafik gösterimi sağlar.

ROC eğrisi en yüksek doğruluk veren kesim (cut-off) noktasını belirler. Eğri ile duyarlılık ve belirleyicilik arasında optimal bir ilişki ile cut off değerinin saptanmasına olanak sağlar. Bir kesim değeri çok düşük alındığında, çok yüksek bir duyarlılık sağlar. Kesim değeri çok yüksek alınırsa da çok yüksek bir seçicilik sağlamaktadır.

Tanı testinin hiçbir veriyi atlamamaktadır. Duyarlılık yüksek alındığında pozitif sonuçlar yakalanırken belirleyicilik ihmal edilmiş olur. Bunun sonucunda ise çok fazla yanlış pozitif sonuçların elde edilir. Çok yüksek kesim değerinin alınması ise; duyarlılığın ihmal edildiği, yüksek belirleyiciliğin sağlaması durumu olarak belirtilebilir.

Sınıflandırma tablosunda duyarlılık ve özgüllük sadece ek bilgi verir, model uyumuna karar vermede doğrudan etkili değildir

Çizelge 3.6. Kesim noktasi 0.60 alindiğinda siniflandirma tablosu [2]

		Beklenen Durum		
Gözlenen Durum		Düzey=1	Düzey=0	Toplam
	Düzey=1	4	7	41
	Düzey=0	121	368	459
	Toplam	125	375	500

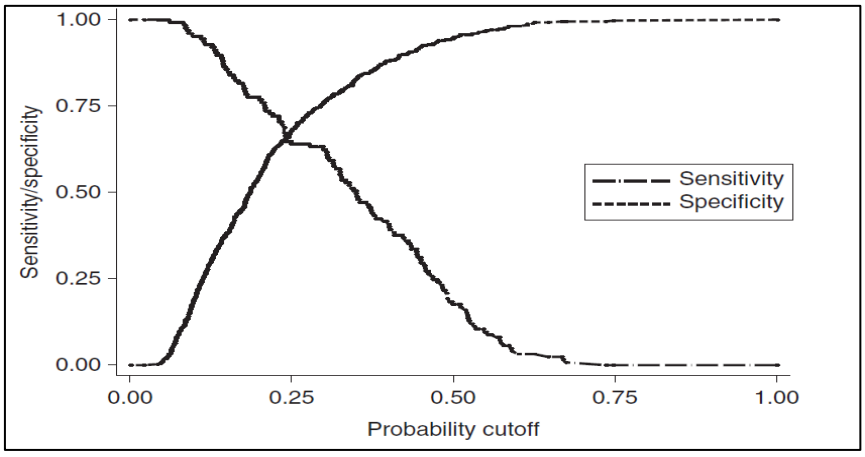
Sınıflandırma tablosunda kesme değerini 0,5 almıştık; bu değeri 0,6 alırsak;

Tabloya göre duyarlılık:4/125=3,2% Tabloya göre seçicilik: 368/375:=98,13%

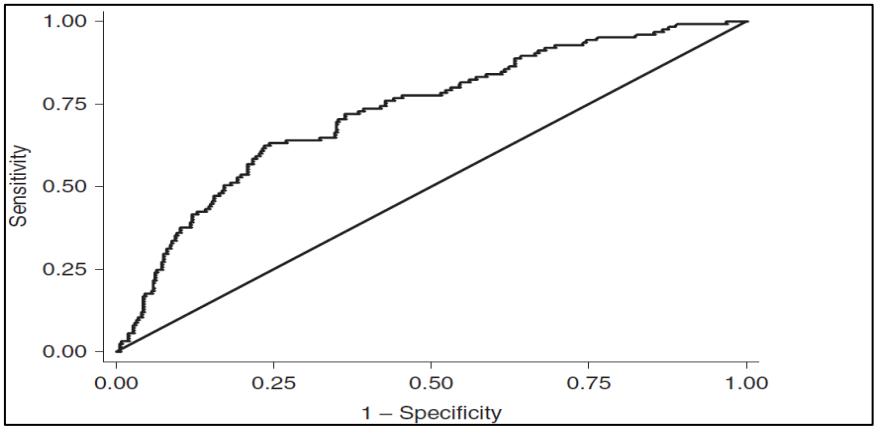
Çizelge 3.7. Kesme değerini 0,05 ile 0,75 arasındaki tüm olası değerleri için duyarlılık, seçicilik ve 1-seçicilik özet tablosu [14]

Kesim Noktası	Duyarlılık	Seçicilik	1-Seçicilik
0,05	100	1,6	98,4
0,10	95,2	19,7	88,3
0,15	84,8	38,4	61,6
0,20	76,8	55,2	44,8
0,25	64,0	68,5	31,5
0,30	62,4	76,5	23,5
0,35	48,8	82,9	17,1
0,40	39,2	88,0	12,0
0,45	29,6	92,3	7,7
0,50	17,6	94,9	5,1
0,55	8,8	96,8	3,2
0,60	3,2	98,1	2,9
0,65	2,4	99,5	0,5
0,70	0,0	99,5	0,5
0,75	0,0	100,0	0,0

Amacımız en iyi kesme noktasını seçmek olduğunda duyarlılık ve seçicilik için her ikisini de birlikte en yüksek olduğu noktayı seçmeyi tercih ederiz.



Şekil 3.1. Mümkün olan tüm kesim noktalarına karşı duyarlılık ve seçicilik [2]



Şekil 3.2. Mümkün olan tüm kesim noktalarına karşı seçicilik ve 1-seçicilik [2]

ROC eğrileri ayırt ediciliği göstermekle beraber, farklı testlerin performans açısından karşılaştırılmasında, eğri altında kalan alana (AUC) gereksinim olur. AUC değeri bir tanı testinin doğruluğunu gösteren bir ölçümdür. O halde tanı testine ilişkin performans değerlendirmeleri için eğri altındaki alanın belirlenmesi gerekir.

AUC değerleri için genelleme yapacak olursak;
AUC = 0,5 alana sahipse ayrım yapılamaz, yorumlanamaz.
$0.7 \leq AUC < 0,8$ ise kabul edilebilir ayrım olduğu düşünülür.
$0.8 \leq AUC < 0,9$ ise mükemmel ayrımdır.
$AUC \geq 0,9$ ise çok iyi ayrımdır [16].

4. BULGULAR

Koroner arter hastalıklar dünya genelinde, mortalite ve morbiditenin majör nedeni olma yolunda gittikçe artan bir rol üstlenmektedir[28]. Kalp damarlarının daralması veya tıkanması olarak tanımlanmaktadır. Bu hastalığın ilerlemesi ileri evrede kalp krizi ve ölüm ile sonlanabilmektedir [29].

Koroner arter hastalıkları çeşitli sebeplerden dolayı damarlarınızda ateroskleroz denilen plakların oluşmasıyla meydana gelir. Damarlarda oluşan ateroskleroz kanınızın akışını engellemektedir [30].

Bu çalışmada, koroner arter hastalığını etkileyen risk faktörlerini belirlemek amacıyla çok değişkenli lojistik regresyon analizi yapılmıştır. Çalışmaya 499 koroner arter hastası, 404 kontrol grubu (sağlıklı grup) olmak üzere toplam 903 hasta dahil edilmiştir [27]. Tüm olguların risk faktörü olduğu araştırılan yaş, cinsiyet, sigara kullanımı, diyabet varlığı, hipertansiyon varlığı ve kolesterol (HDL ve LDL), değerleri çalışmaya dahil edilmiştir.

Lojistik regresyon yöntemlerinden ikili lojistik regresyon yöntemi kullanılmıştır. Bağımlı değişken olarak koroner arter hastalığı olan ve olmayan grup alınmış olup, koroner arter hastalar 1 ile kontrol grubu ise 0 ile kodlanmıştır. Tüm olgular için koroner arter hastalığını diğer bağımsız değişkenlerin etkileyip etkilemediği araştırılarak verinin modele uyumunu değerlendirmek amacıyla pearson ki kare istatistiği, sapma istatistiği, pseudo değerleri (cox ve snell, nagelkerke, mc fadden), hosmer lemeshow testleri, sınıflandırma tabloları ve roc eğrisi altında kalan alan hesaplanmıştır.

Verilen analizinde SPSS programı kullanılmıştır. İstatistiksel anlamlılık düzeyi %5 olarak alınmıştır.

Öncelikle aşağıdaki tabloda bulunan aday değişkenlerin hangilerinin modelin oluşturulması açısından önemli olduğuna karar verilecektir. Kategorik değişkenlerde “0” olarak kodlanan gruplar referans kategori olarak alınmıştır.

Çizelge 4.1. Lojistik regresyon analizine dahil edilecek değişkenler

Bağımlı Değişken		
Y	Koroner Arter Hastalığı	1: Koroner Arter Hastalığı Olanlar 0: Kontrol Grubu
Bağımsız Değişkenler		
X_1	Yaş	
X_2	Cinsiyet	1: Erkek 0: Kadın
X_3	Sigara Kullanımı	1: Sigara Kullananlar 0: Sigara Kullanmayanlar
X_4	Diyabet Varlığı	1: Diyabeti Olanlar 0: Diyabeti Olmayanlar
X_5	Hipertansiyon Varlığı	1: Hipertansiyonu Olanlar 0: Hipertansiyonu Olmayanlar
X_6	LDL	
X_7	HDL	

4.1. Araştırmada Kullanılan Bağımsız Değişkenlere Ait Tanımsal İstatistikler

Çalışmaya dahil edilen hastalara ilişkin değişkenlere ait veriler tabloda gösterilmektedir.

Çizelge 4.2. Gruplar arası bağımsız değişkenlerin karşılaştırması

		Sayı	AO	SS	Med.	Min.	Maks.	P değeri
Yaş	Hasta	499	63,20	10,32	63,00	31,00	87,00	<0,001
	Kontrol	404	56,52	11,81	56,00	24,00	86,00	
LDL	Hasta	499	120,92	38,98	117,00	24,51	255,00	0,004
	Kontrol	404	128,30	36,30	128,24	36,00	287,00	
HDL	Hasta	499	40,89	11,22	40,00	10,00	104,00	<0,001
	Kontrol	404	46,24	13,12	44,73	20,00	119,00	
Student t testi kullanılmıştır.								

Tablo incelendiğinde, koroner arter hastalığı olanlarda yaş ortalaması, student t testi sonuçlarına göre kontrol grubuna göre %1 anlamlılık düzeyinde daha yüksektir. Diğer bir deyişle koroner arter hastalığı olanların yaş ortalaması kontrol grubunun yaş ortalamasına göre daha yüksek olup bu durum yaşın p koroner arter hastalığı oluşma riskini açıklamada etkili olabileceğini göstermektedir.

Gruplar arasında LDL değerleri değerlendirecek olursa, student t testi uygulanmış olup, koroner arter hastalığı gelişen gurubun LDL ortalaması kontrol grubuna göre daha düşük olup bu farklılık %5 anlamlılık düzeyinde istatistiksel olarak anlamlı bulunmuştur.

Gruplar arasında HDL değerlendirildiğinde, student t testi uygulanmış olup uygulanmış olup, koroner arter hastalığı olan grubun HDL ortalaması kontrol grubuna göre daha düşük olup bu farklılık %5 anlamlılık düzeyinde istatistiksel olarak anlamlı bulunmuştur.

Çizelge 4.3. Gruplar arası bağımsız değişkenlerin karşılaştırması

		Kontrol Grubu		Hasta Grubu		x^2	P değeri
		n	*%*	*n*	*%*		
Cinsiyet	Kadın	220	54,5%	134	26,9%	71,362	<0,001
	Erkek	184	45,5%	365	73,1%		
Sigara Kullanımı	Yok	240	59,4%	252	50,5%	7,139	0,008
	Var	164	40,6%	247	49,5%		
Diyabet Varlığı	Yok	313	77,5%	284	56,9%	42,129	<0,001
	Var	91	22,5%	215	43,1%		
Hipertansiyon Varlığı	Yok	176	43,6%	133	26,7%	28,364	<0,001
	Var	228	56,4%	366	73,3%		

Tablo incelendiğinde, koroner arter hastalığı olanlarda kontrol grubuna göre erkeklerin oranı daha yüksektir ve %5 anlamlılık düzeyinde gruplar arasında istatistiksel olarak anlamlı farklılık vardır (p:<0,001).

Tablo incelendiğinde, koroner arter hastalığı olanlarda kontrol grubuna göre sigara kullananların oranı daha yüksektir ve %5 anlamlılık düzeyinde gruplar arasında istatistiksel olarak anlamlı farklılık vardır (p:0,008).

Tablo incelendiğinde, koroner arter hastalığı olanlarda kontrol grubuna göre diyabet hastası olanların oranı daha yüksektir ve %5 anlamlılık düzeyinde gruplar arasında istatistiksel olarak anlamlı farklılık vardır (p:<0,001).

Tablo incelendiğinde, koroner arter hastalığı olanlarda kontrol grubuna göre erkeklerin oranı daha yüksektir ve %5 anlamlılık düzeyinde gruplar arasında istatistiksel olarak anlamlı farklılık vardır (p:<0,001).

Çizelge 4.4. Bağımsız değişkenler arasındaki ilişki

		Yaş	LDL	HDL
Yaş	r	1	-0,151	0,093
	p		<0,001	0,005
	n			
LDL	r		1	0,066
	p			0,047
	n			
HDL	r			1
	p			
	n			

Bağımsız değişkenler birbiriyle ilişki olduğu durumda çoklu bağlantı sorunu oluşmaktadır. Bu sebeple öncelikle sürekli değişkenler arasındaki korelasyon değerlerine bakılmış olup pearson korelasyon değerlerine göre sürekli olan değişkenler arasındaki korelasyonlar anlamlı bulunmuştur. Ancak anlamlılık düzeyleri çok yüksek değildir be sebeple tüm değişkenler lojistik regresyona dahil edilmiştir.

4.2. Modelin Anlamlılığı ve Uyum İyiliği Sonuçları

Analize dahil edilmesi gereken, bağımlı değişkeni en iyi açıklayan bağımsız değişkenlerle model kurulduktan sonra modelin veriye uyumunu değerlendirmek amacıyla pearson ki kare ve sapma istatistiği, pseudo değerleri (cox ve snell, nagelkerke, mc fadden), hosmer lemeshow testleri, sınıflandırma tabloları ve roc eğrisi altında kalan alan kullanılmıştır. Bu çalışmada lojistik regresyon analizine yaş, cinsiyet, sigara kullanımı, diyabet varlığı, hipertansiyon varlığı, HDL (iyi kolesterol) ve LDL (kötü kolesterl) risk faktörü olarak modele dahil edilmiştir. Çoklu lojistik regresyon analizinde backward LR (geriye doğru elemeli) yöntemi kullanılmıştır. Yöntem sonucu 2 adım oluşturulmuş olup 2. Adım sonuçları verilmiştir.

Pearson ki kare istatistiği;

H_0 = Model verileri itibariyle uygundur.
H_1 = Model verileri itibariyle uygun değildir.

$$r\,(y_j, \hat{\pi}_j) = \frac{y_j - m_j\hat{\pi}_j}{\sqrt{m_j\hat{\pi}_j\,(1-\hat{\pi}_j)}} \quad (4.1)$$

$$x^2 = \sum_{j=1}^{j}\left[r\,(y_j,\ \hat{\pi}_j)\right]^2 \quad (4.2)$$

Pearson artığı istatistiği

$$\sum_{j=1}^{918}\left[\frac{(0-(-0{,}015))}{0{,}015\,(1)}\right]^2 + \cdots + \left[\frac{0-(-0{,}978)}{(0{,}978)(1)}\right]^2 = 11{,}30$$

Bu artıklara dayalı j-p-1 serbestlik dereceli ki-kare istatistiği;

$x^2{}_h$ = 11,30

$x^2{}_{22,0.05}$ =33,92

$x^2{}_h < x^2{}_{8,0.05}$ olduğundan H_0 kabul edilir.

veriler model ile uyumludur.

Sapma istatistiği;

H_0 = Model verileri itibariyle uygundur.
H_1 = Model verileri itibariyle uygun değildir.

$$d\,(y_j, \hat{\pi}_j) = \pm 2\left(\left[y_j In\left(\frac{y_j}{m_j\hat{\pi}_j}\right) + (m_j - y_j) In\left(\frac{m_j - y_j}{m_j(1-\hat{\pi}_j)}\right)\right]\right)^{1/2} \quad (4{,}3)$$

$$D = \sum_{j=1}^{j} d\,(y_j, \hat{\pi}_j)^2 \quad (4.4)$$

D = 25,20

$x^2{}_h$ = 25,20

$x^2{}_{22,0.05}$ =33,92

$x^2{}_h < x^2{}_{8,0.05}$ olduğundan H_0 kabul edilir.

veriler model ile uyumludur.

Hosmer Lemeshow Testi;

Çizelge 4.5. Koroner arter hastalığını etkileyen faktörlerin belirlenmesinde sabit denek sayılı onlu risk grupları için gözlenen ve beklenen frekanslar

Risk grupları	Durum=0		Durum=1		Toplam
	Gözlemlenen değer	Beklenen değer	Gözlemlenen değer	Beklenen değer	
1	81	79,710	9	10,290	90
2	74	67,128	16	22,872	90
3	60	56,889	30	33,111	90
4	39	47,777	51	42,223	90
5	35	41,563	55	48,437	90
6	33	35,585	57	54,415	90
7	29	28,469	61	61,531	90
8	23	22,129	67	67,871	90
9	18	15,740	72	74,260	90
10	12	9,010	81	83,990	93

Uyum iyiliğine karar vermek için, yukarıdaki denklemde verilen formül yardımıyla çizelge 4.5'deki veriler kullanılarak, Hosmer-Lemeshow «Ĉ» istatistiği hesaplanır. C istatistiği 8 (10-2) serbestlik dereceli ki-kare dağılımı sahiptir.

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_{1k} - \hat{e}_{1k})^2}{\hat{e}_{1k}} - \frac{(o_{0k} - \hat{e}_{0k})^2}{\hat{e}_{0k}} \qquad (4.5)$$

$$\hat{C}_g = \sum_{k=0}^{1} \sum_{l=1}^{10} \frac{(o\,kl - eki)}{ekl} = \frac{(81-79{,}71)^2}{79{,}71} + \frac{(9-10{,}29)^2}{10{,}29} + \ldots + \frac{(12-9{,}010)^2}{9{,}10} + \frac{(81-83{,}99)^2}{83{,}99} =$$

10,638

$x^2{}_h = \hat{C}_g = 10{,}638$

$x^2{}_{8,0.05} = 15{,}507$

$x^2{}_h < x^2{}_{8,0.05}$ olduğundan H_0 kabul edilir.

Serbestlik derecesi 8 olan modelin p değeri de p=0,223 bulunmuştur. P değeri >0,05 H_0 kabul edilir. Modelin veriye uyumunun iyi olduğu anlaşılmaktadır.

Çizelge 4.6. Koroner arter hastalığını etkileyen faktörlerin belirlenmesinde tahmini olasılıkların persentillerine göre onlu risk grupları için gözlenen ve beklenen frekanslar

Risk grupları		Durum=0		Durum=1		Toplam
		Gözlemlenen değer	Beklenen değer	Gözlemlenen değer	Beklenen değer	
1	0-0,1	34	33,67	2	2,33	36
2	0,11-0,2	58	54,92	7	10,08	65
3	0,21-0,3	51	47,18	12	15,82	63
4	0,31-0,4	58	55,66	27	29,34	85
5	0,41-0,5	50	56,1	53	46,90	103
6	0,51-0,6	53	61,71	84	75,29	137
7	0,61-0,7	39	40,04	75	73,96	114
8	0,71-0,8	33	32,02	94	94,98	127
9	0,81-0,9	24	19,48	104	108,52	128
10	0,91-1,0	4	3,22	41	41,78	45

Uyum iyiliğine karar vermek için, yukarıdaki denklemde verilen formül yardımıyla çizelge 4.6'daki veriler kullanılarak, Hosmer-Lemeshow «H» istatistiği hesaplanır. H istatistiği 8 (10-2) serbestlik dereceli ki-kare dağılımı sahiptir.

$$\widehat{H}_{\mathrm{g}} = \sum_{k=0}^{1}\sum_{l=1}^{6}\frac{(o\,kl-eki)}{ekl} = \frac{(34-33{,}67)^2}{33{,}67} + \frac{(2-0-2{,}33)^2}{0-2{,}33} + \ldots + \frac{(4-3{,}22)^2}{3{,}22} + \frac{(41-41{,}78)^2}{41{,}78} \qquad (4.6)$$

$= 7{,}89$

$x^2{}_h = \widehat{H}_{\mathrm{g}} = 7{,}89$

$x^2{}_{8,0.05} = 15{,}507$

$x^2{}_h < x^2{}_{8,0.05}$ olduğundan H_0 kabul edilir.

Serbestlik derecesi 8 olan modelin P değeri >0,05 H_0 kabul edilir. Modelin veriye uyumunun iyi olduğu anlaşılmaktadır.

Sınıflandırma Tablosu;

Çizelge 4.7. Sınıflandırma tablosu sonuçları

		Beklenen Durum		
		Düzey=0	Düzey=1	Toplam
Gözlenen Durum	Düzey=0	251	153	404
	Düzey=1	101	398	499
	Toplam	352	551	903

Tabloya göre duyarlılık; 398/551= 72,2%

Tabloya göre seçicilik; 251/352 = 71,3%

Doğru sınıflama yüzdesi; 100 * (251 + 398) / 903 = 71,87

Bir çok lojistik regresyon modelinde görülen tipik bir sınıflandırma tablosu örneğimizde doğru sınıflandırma yüzdesi %71,87 olarak bulunmuştur. Tabloya göre gelişme durumumuz 1 olduğunda (bu durum duyarlılık olarak nitelenmekte) ,% 72,2 doğruluk payının olduğu gözlenmekte iken, gelişme durumumuz 0 olduğunda (bu durum da seçicilik olarak ifade edilmekte), % 71,3 doğruluk payı içermektedir. Sınıflandırma tablosundan elde edilen duyarlılık, seçicilik ve doğru sınıflama yüzdesine göre modelin veriye uyumunun orta düzeyde olduğu görülmüştür.

Pseduo R^2 istatistikleri (Cox ve Snell R^2, Nagelkerke R^2);

Çizelge 4.8. Pseudo değerleri istatistikleri sonuçları

Adım	-2 Log likelihood	Cox & Snell R^2	Nagelkerke R^2
1	1003,884	0,232	0,310

Modelin açıklayıcılığını belirtmek amacıyla kullanımı önerilmemektedir. Modeller arasında seçim yapma sürecinde karşılaştırmada kullanılmalıdırlar. Tüm model sonucu iyi bir şekilde tahmin eder ve olabilirliği 1'e eşit olursa, Nagelkarke R^2'de 1'e eşit olacaktır. 0,20 ile 0,40 arasındaki R^2 değerleri için modelin uyumu yeterlidir denebilir.

Ancak bizim analiz sonuçlarımız için bakılacak olursa cox & snell R^2 ve Nagelkerke R^2 değeri 0,3'ün üstünde bulunmuştur. Sonuç olarak verinin modele uyumunu yeterli olduğu anlaşılmıştır.

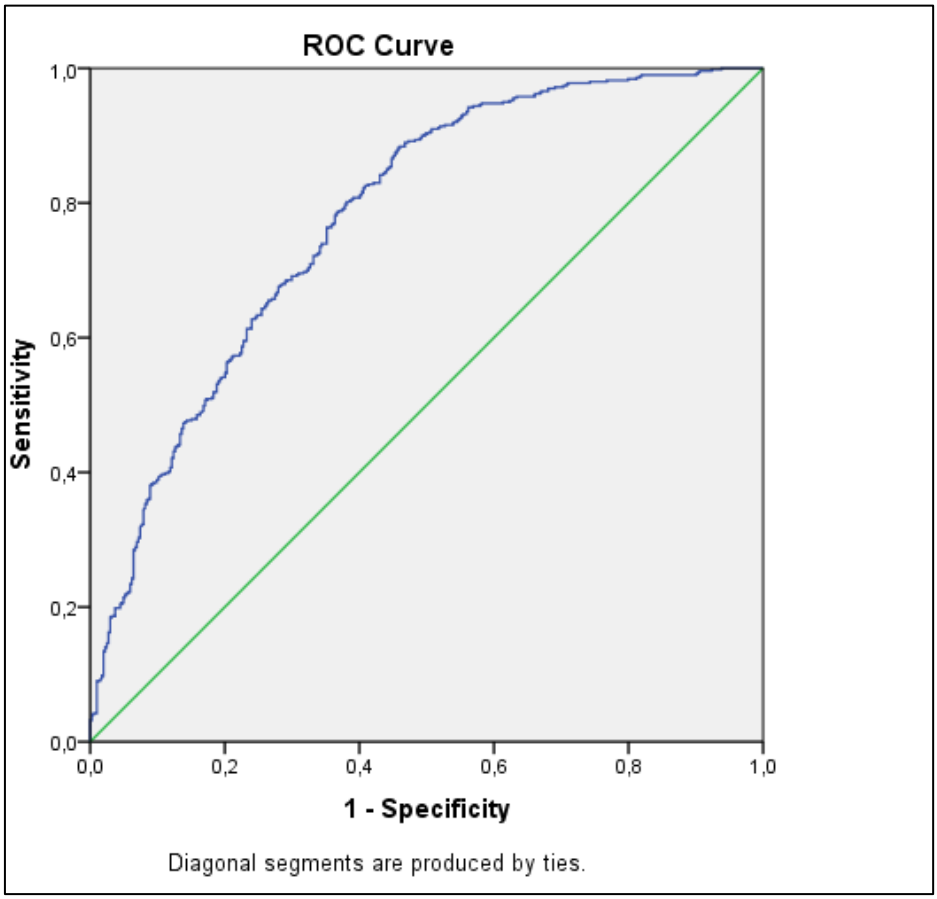
Roc Eğrisi Altında Kalan Alan;

Seçilen farklı eşik değerleri için bulunan farklı duyarlılık-belirleyicilik özelliklerine bağlı olarak, ROC eğrisi (Receiver operating characteristic curves) oluşturulmuştur.

Tahmin edilen olasılık değerlerine ($\hat{\pi}$) göre uygulanan roc eğrisi sonuçlarına göre eğri altında kan alan %78 bulunmuştur. Bu sonuç orta derecede iyidir. Tahmin edilen olasılık değerlerinin %78 düzeyinde ayırt edici olduğunu göstermektedir.

Çizelge 4.9. ROC eğrisi sonuçları

AUC	SH	P değeri	Güven Aralığı	
0,778	0,015	<0,001	0,748	0,808



Şekil 4.1. Roc eğrisi

Roc analizi sonucunda aşağıda kesme değerlerine denk gelen duyarlılık ve özgüllük değerleri yer almaktadır. Bu sonuçlar neticesinde tahmin edilen olasılık değerlerine ($\hat{\pi}$)

göre en iyi kesme noktası %69,2 duyarlılıkta ve %66,2 özgüllükte 0,501 olasılık değerinin, %64,6 duyarlılıkta ve %70,8 özgüllükte 0,522 olasılık değerinin ve %63,1 duyarlılıkta ve %72,3 özgüllükte 0,531 olasılık değerinin en iyi ayrım noktaları olduğu görülmektedir.

Çizelge 4.10. ROC eğrisi sonuçlarına göre hesaplanan kesme değerler

Cut Off	Duyarlılık	Özgüllük
0,4336	0,882	0,542

Çizelge 4.11'de tahmin edilen olasılık değerlerine ($\hat{\pi}$) göre bazı duyarlılık ve özgüllük değerleri için hesaplanan kesme değerler verilmiştir.

Çizelge 4.11. Bazı kesme değerlerine göre duyarlılık ve özgüllük değerleri özet tablosu

Kesme Değeri	Duyarlılık	1-Özgüllük	Özgüllük
0,0000000	1,000	1,000	0,000
0,0192145	1,000	0,998	0,002
0,0244721	1,000	0,995	0,005
0,0261802	1,000	0,993	0,007
0,0284053	1,000	0,990	0,010
0,0326432	1,000	0,988	0,012
0,0404693	1,000	0,985	0,015
0,0467559	1,000	0,983	0,017
0,0489670	1,000	0,980	0,020
0,0520933	1,000	0,978	0,022
0,0549242	1,000	0,975	0,025
0,0561440	1,000	0,973	0,027
0,0565322	1,000	0,970	0,030
0,0573272	1,000	0,968	0,032
0,0585768	1,000	0,965	0,035
0,1303489	0,990	0,886	0,114
0,1322412	0,990	0,884	0,116
0,1332897	0,990	0,881	0,119
0,1334179	0,990	0,879	0,121
0,1335952	0,990	0,876	0,124
0,1341829	0,990	0,874	0,126
0,1356305	0,990	0,871	0,129

Çizelge 4.11. (devam) Kesme değerlerine göre duyarlılık ve özgüllük değerleri özet tablosu

Kesme Değeri	Duyarlılık	1-Özgüllük	Özgüllük
0,3746768	0,914	0,527	0,473
0,4510483	0,858	0,448	0,552
0,4527056	0,856	0,448	0,552
0,4548641	0,854	0,448	0,552
0,4560389	0,854	0,446	0,554
0,4566418	0,852	0,446	0,554
0,4572699	0,852	0,443	0,557
0,4575068	0,850	0,443	0,557
0,4583638	0,850	0,441	0,559
0,4596110	0,848	0,441	0,559
0,4601185	0,846	0,441	0,559
0,4605550	0,846	0,438	0,562
0,4611342	0,844	0,438	0,562
0,4615249	0,844	0,436	0,564
0,4617530	0,842	0,436	0,564
0,4619233	0,842	0,433	0,567
0,4628007	0,842	0,431	0,569
0,4637740	0,840	0,431	0,569
0,4643679	0,838	0,431	0,569
0,4648351	0,836	0,431	0,569
0,4653331	0,834	0,431	0,569
0,4664292	0,832	0,431	0,569
0,4673635	0,830	0,431	0,569
0,4676541	0,830	0,428	0,572
0,4683199	0,830	,426	0,574
0,4690927	0,830	0,423	0,577
0,4702997	0,830	0,421	0,579
0,5488586	0,699	0,324	0,676
0,5504282	0,699	0,322	0,678
0,5516675	0,697	0,322	0,678
0,5526893	0,697	0,319	0,681
0,5546976	0,697	0,317	0,683
0,5562631	0,695	0,317	0,683
0,5568494	0,695	0,314	0,686
0,5580034	0,695	0,312	0,688
0,5593296	0,695	0,309	0,691
0,5596373	0,693	0,309	0,691
0,5605490	0,693	0,307	0,693
0,5617239	0,691	0,307	0,693
0,5623670	0,691	0,304	0,696
0,5634701	0,691	0,302	0,698

4.3. Çoklu Lojistik Regresyon Analiz Sonuçları

Çizelge 4.12. Koroner arter hastalığını etkileyen faktörler içi lojistik regresyon analizi sonuçları

	β	Sh.	Wald	p	Exp(β)	95% Güven Aralığı	
Sabit	-3,624	0,557	42,323	<0,001	0,027		
Yaş	0,066	0,008	68,539	<0,001	1,068	1,051	1,085
Cinsiyet	1,260	0,171	53,948	<0,001	3,524	2,518	4,932
Sigara Kullanımı	0,368	0,165	4,989	0,026	1,445	1,046	1,996
Diyabet Varlığı	0,654	0,168	15,058	<0,001	1,923	1,382	2,675
Hipertansiyon	0,488	0,169	8,352	0,004	1,629	1,170	2,268
HDL	-0,036	0,007	28,207	<0,001	0,965	0,952	0,978
LDL							

Cinsiyet: Referans Kategori kadın. β ; beta, Sh. Standart hata, Sd.; Serbestlik derecesi, Exp(B); odds oranı

Çoklu lojistik regresyon denklemi;

$$\hat{y}_i = -3{,}624 + 0{,}066x_1 + 1{,}260x_2 + 0{,}368x_3 + 0{,}654x_4 + 0{,}488x_5 - 0{,}036x_7$$

Backward LR metodu kullanılarak uygulanan çoklu lojistik regresyon analizinde yaş cinsiyet, sigara kullanımı, diyabet varlığı, hipertansiyon varlığı, HDL (iyi kolesterol) koroner arter hastalığını öngörmede yeterli bulunmuştur ($p<0{,}05$). Analiz sonucunda LDL (kötü kolesterol) koroner arter hastalığını öngörmede yeterli bunmamış olup 2. Adımda analizden çıkarılmıştır.

Çoklu lojistik regresyon sonuçlarına göre yaşın bir birim artması koroner arter hastalığının görülme riskini 1,068 kat arttırmaktadır.

Erkeklerde kadınlara oranla korener arter hastalığı görülme riski 3,524 kat fazlayken, sigara kullanımı da hastalığın görülme riskini 1,445 kat arttırmaktadır.

Diyabet hastalarında hastalığın görülme riski 1,923 kat artmakta iken hipertansiyonu olanlarda hastalığın görülme riski 1,629 kat artmaktadır.

HDL (iyi kolesterol)'nin bir birim artması (1-0,965=0,035) koroner arter hastalığının görülme riskini %3,5 azaltmaktadır.

5. SONUÇLAR VE ÖNERİLER

İstatistiğin temel amaçlarından biri olan tahminde bulunmayı sağlayan, günümüzde tıp, biyoloji, ekonomi vb. alanlarda kullanılabilen lojistik regresyon modelleri, hem bağımlı değişken ile bağımsız değişkenler (ler) arasındaki ilişkiyi tanımlayan en uygun modeli belirlemeyi sağlayan hem de bilinen bağımsız değişken (ler) yardımıyla bilinmeyen bağımlı değişken hakkında kestirim yapmamıza olanak sağlayan son derece önemli modellerdendir. Diğer regresyon analizlerine oranla varsayımlarının azlığı ve yorumlamasının kolaylığı gün geçtikçe çalışmalarda kullanımını arttırmaktadır.

Lojistik regresyon analizinde verinin modele uyumunu değerlendiren çeşitli testler kullanılmaktadır. Bu testlerin benzer sonuçlar verip vermediği tezin amacını oluşturmaktadır.

Tezin uygulama bölümünde bağımlı değişken iki kategoriden oluştuğu için binary lojistik regresyon analizi kullanılmıştır.

Yapılan analiz sonuçlarına bakılacak olursa Çoklu lojistik regresyon sonuçlarına göre yaşın bir birim artması koroner arter hastalığının görülme riskini 1,068 kat arttırmaktadır. Erkeklerde kadınlara oranla korener arter hastalığı görülme riski 3,524 kat fazlayken, sigara kullanımı da hastalığın görülme riskini 1,445 kat arttırmaktadır. Diyabet hastalarında hastalığın görülme riski 1,923 kat artmakta iken hipertansiyonu olanlarda hastalığın görülme riski 1,629 kat artmaktadır. HDL (iyi kolesterol)'nin bir birim artması (1-0,965=0,035) koroner arter hastalığının görülme riskini %3,5 azaltmaktadır.

Araştırmaya göre koroner arter hastalığı erkeklerde kadınlara göre daha yüksek oranda görüldüğü ve bir risk faktörü olduğu anlaşılmıştır. Araştırma sonucunda koroner arter hastalığı için risk faktörü olabilecek değişkenler belirlenmiştir.

Koroner arter hastalığı için risk faktörleri araştırılmak amacıyla uygulanan lojistik regresyon analizinde modelin uyum iyiliğinin incelenmesi konusunda, pearson ki kare ve sapma istatistikleri sonuçlarına göre modelin veriye uyumlu olduğu anlaşılmıştır. Hosmer lemeshow testi sonucuna göre sabit denek sayılı risk gruplarına hesaplanan istatistik değeri $\hat{C}_g$ = 10,638 ve olasılıkların persentillerine göre hesaplanan istatistik

değeri $\widehat{H}_g$ = 7,89 bulunmuştur. Olasılıkların persentillerine göre hesaplanan istatistik değeri daha düşük bulunmuştur. Hosmer ve lemeshow (1982) tarafından yapılan bir simülasyon çalışmasında $\widehat{H}$ istatistiğinin $\hat{C}$ istatistiğine göre daha etkili olduğu tespit edilmiş olup [14] çalışmamızda benzer sonuçlar verdiği değerlendirilmiştir.

Uygulama da model uyumunu değerlendirmede doğru sınıflama oranı %71,87 hesaplanmıştır. Yani modelin tahmininde her 100 gözlemden 72'si doğru tahmin edilmiştir. Bu oran yüksek olmasa da yeterlidir.

Pseudo değerlerinin diğer metodlara göre iyi bir ölçüt olmadığı bilinmekle birlikte bu çalışmada pseudo istatistiklerine göre verinin modele uyumlu olduğu değerlendirilmiştir.

Verinin modele uyumunu değerlendirme açısından bize pek bilgi vermediği fakat sonuçları destekleyici olduğu ayrıca roc analizi sonucunda eğri altında kalan alan %77,8 olarak tespit edilmiş ve orta düzeyde olduğu gözlemlenmiştir.

Çalışma genel olarak değerlendirildiğinde, uyum iyiliği testlerinden kikare istatistiği, sapma istatistiği ve hosmer lemeshow testlerinin birbirine paralel sonuçlar verdiği sınıflandırma tablosu ve roc eğrisi analizlerinin sonuçları, kikare istatistiği, sapma istatistiği ve hosmer lemeshow testlerinin sonuçlarını destekler nitelikte olduğu ancak pseudo değerlerinin model uyumunu değerlendirme de yapılan literatür araştırmalarında iyi bir test olmadığı belirtilse de çalışmamızda diğer testlerle benzer sonuçlar verdiği anlaşılmıştır.

KAYNAKLAR

1. Kleinbaum, D. G. and Hedeker, D. (1996). Logistic regression: a self-learning text. *Statistical Methods in Medical Research*, *5*(1), 103.

2. Hosmer, W. D. and Lemeshow, S. (2013). *Applied logistic regression*, John Wiley & Sons, Canada, 398.

3. Arabacı, Ö. (2002). *Lojistik regresyon analizi ve bir uygulama denemesi*, yayınlanmamış Yüksek Lisans Tezi, Uludağ Üniversitesi Sosyal Bilimler Enstitüsü, Bursa.

4. Johnson, D. E. (1998). *Applied multivariate methods for data analysts*. Pacific Grove, California: Duxbury Press, 48.

5. Kaşko, Y. (2007). *Çoklu bağlantı durumunda ikili (binary) lojistik regresyon modelinde gerçekleşen 1.tip hata ve testin gücü*, Yüksek Lisans Tezi, Ankara Üniversitesi Fen Bilimleri Enstitüsü, Ankara

6. Alpar, R. (2013). *Uygulamalı çok değişkenli istatistiksel yöntemler*, Ankara: Detay Yayıncılık, 637-712.

7. Akçay, A. (2009). *Broiler coccidiosis'inde risk faktörlerinin lojistik regresyon analizi ile belirlenmesi*, Yayınlanmamış Doktora, Tezi Ankara Üniversitesi Sağlık Bilimleri Enstitüsü, Ankara

8. Elhan, A. H. (1997). *Lojistik regresyon analizinin incelenmesi ve tıpta bir uygulaması,* Yayınlanmamış Yüksek Lisans Tezi, Ankara Üniversitesi Sağlık Bilimleri Enstitüsü, Ankara

9. Tatlıdil, H. (1996). *Uygulamalı çok değişkenli istatistiksel analiz*, Ankara: Akademi Matbaa, 100-150.

10. Agresti, A. (2002). *Categorical data analysis,* John Willey & Sons, 2. Baskı, New York, Inc., Publication, 1.

11. Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., Tatham, R. L. (2006). Multivariate data analysis 6th Edition. *Pearson Prentice Hall.* New Jersey. humans: Critique and reformulation. *Journal of Abnormal Psychology*, *87*, 49-74.

12. Boucher, K. M., Slattery, M. L., Berry, T. D., Quesenberry, C. and Anderson, K. (1998). Statistical methods in epidemiology: a comparison of statistical methods to analyze dose–response and trend analysis in epidemiologic studies. *Journal of clinical epidemiology*, *51*(12), 1223-1233.

13. Güç, K. (2015). *Türkiye'de resmi kurumlara duyulan güvenin kategorik regresyon ve lojistik regresyon analizi ile incelenmesi*. Yüksek Lisans Tezi, Gazi Üniversitesi Fen Bilimleri Enstitüsü, Ankara.

14. Başarır, G. (1990). *Çok değişkenli verilerde ayrımsama sorunu ve lojistik regresyon analizi,* Yayınlanmamış Doktora Tezi, Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, Ankara

15. Çokluk, Ö., Şekercioğlu, G. ve Büyüköztürk, Ş. (2014). *Sosyal bilimler için çok değişkenli istatistik: SPSS ve LISREL uygulamaları*. (3.Baskı). Ankara: Pegem Akademi Yayınları, 80-100.

16. Şensoy, Z. E. (2009). *Nonlineer lojistik regresyon ve uygulaması*, Yüksek Lisans Tezi Marmara Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul.

17. Yazar, G. N. (2018). *Aile içi kadına yönelik şiddeti etkileyen faktörlerin lojistik regresyon ile analizi*. Yüksek Lisans Tezi, Uludağ Üniversitesi Sosyal Bilimler Enstitüsü, Bursa

18. İnternet: Rush, S. (2001). Logistic regression: the standard method of analysis in medical research. URL: https://pdfs.semanticscholar.org/7fa6/1d9639ff581fe6bfccdb96c59 6a07b0fb57.pdf,s.4, Son Erişim Tarihi: 17.05.2016.

19. Allison, P. D. (2000). *Logistic regression using the sas system:Theory and application*, SAA Institute Inc.USA

20. Çolak, C. (2001). *Lojistik regresyon analizi ve sağlık bilimlerinde bir uygulama,* Yüksek Lisans Tezi, İnönü Üniversitesi Sağlık Bilimleri Enstitüsü, Malatya.

21. Kara, Ö. S. (2015). *Lojistik regresyon analizi ve kadın işgücü üzerine bir uygulama*. Yüksek Lisans Tezi, Uludağ Üniversitesi Sosyal Bilimler Enstitüsü, Bursa.

22. Oğuzlar, A. (2005). Lojistik regresyon analizi yardımıyla suçlu profilinin belirlenmesi. *Atatürk Üniversitesi İktisadi ve İdari Bilimler Dergisi*, *19*(1), 21-35.

23. Kalaycı, Ş. (2010). *SPSS uygulamalı çok değişkenli istatistik teknikleri*. Ankara: Asil Yayın Dağıtım, 5.

24. İyit, N. (2003). *Lineer Olmayan Lojistik Regresyon Analizinde Model Kurma Stratejileri ve Bir Uygulaması*. Yüksek Lisans Tezi, Selçuk Üniversitesi Fen Bilimleri Enstitüsü, Konya.

25. Pregibon, D. (1981). Logistic regression diagnostics. *The Annals of Statistics*, *9*(4), 705-724.

26. Şansh, Ş. and Ulutagay, G. (2006). *Logistic regression analysis to determine the factors that affect "Green Card* - Usage For Health Services, JFS.

27. Ekici, B, Tanınd,ı A, ve Sayın, I. (2016). *The effects of glomerular filtration rate on the severity of coronary heart disease*. Turk Kardiyoloji Derneği Araştırmaları, 44, 123-129.

28. Charles H, Hennekens, MD,DrPH. Increasing burden of cardiovascular disease. Current knowledge and future firections for research on risc factors. Circulation. 1998; 97: 1095-1102.

29. Yalçın, H. (2018). *Koroner arter hastalığı tanı ve yönetiminde nükleer kardiyoloji*. Türkiye Nükleer Tıp Derneği.

30. Kasapoğlu, E. S., ve Enç, N. (2017). Koroner Arter Hastaları İçin Bir Rehber. *Journal of Cardiovascular Nursing*, *8*(15), 1-7.

ÖZGEÇMİŞ

Kişisel Bilgiler

Soyadı, adı : ARSLANOĞLU, Esra
Uyruğu : T.C.
Doğum tarihi ve yeri : 07.05.1990, Ankara
Medeni hali : Bekar
Telefon : 0 (534) 926 51 40
Faks : 0 (312) 202 37 10
e-mail : esra-arslanoglu@hotmail.com.tr

Eğitim

Derece	Eğitim Birimi	Mezuniyet Tarihi
Yüksek Lisans	Gazi Üniversitesi / İstatistik	Devam Ediyor
Lisans	Gazi Üniversitesi / İstatistik	2013
Lise	Keçiören İncirli Lisesi	2007

İş Deneyimi

Yıl	Yer	Görev
2016-Halen	Sosyal Güvenlik Kurumu	Denetmen Yardımcısı

Yabancı Dil

İngilizce

Yayınlar

1. Arslanoğlu, E. ve Çelik, B. (2019). *Lojistik regresyon analizinde model uyumunun değerlendirilmesinde kullanılan yöntemlerin karılaştırılması*. 6. Multidisipliner Çalışmalar Kongresi, Gaziantep.

Hobiler

Yüzme, gitar, dans

GAZİ GELECEKTİR..