

**T.C.
MANİSA CELAL BAYAR ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

**YÜKSEK LİSANS TEZİ
MATEMATİK ANABİLİM DALI
UYGULAMALI MATEMATİK BİLİM DALI**

**ROMATİZMAL HASTALIKLARA SAHİP BİREYLERE
UYGULANABİLECEK BİNARY LOJİSTİK REGRESYON
MODELİNİN GELİŞTİRİLMESİ**

Hussam ALSİTARİ

**Danışman
Prof. Dr. Ali KONURALP**



MANİSA-2023

HUSSAM
ALŞİTARİ

ROMATİZMAL HASTALIKLARA SAHİP BİREYLERE
UYGULANABİLECEK BİNARY LOJİSTİK REGRESYON
MODELİNİN GELİŞTİRİLMESİ

2023

TEZ ONAYI

Hussam ALSİTARİ tarafından hazırlanan "**Romatizmal Hastalıklara Sahip Bireylere Uygulanabilecek Binary Lojistik Regresyon Modelinin Geliştirilmesi**" adlı tez çalışması 23.06.2023 tarihinde aşağıdaki jüri üyeleri önünde Manisa Celal Bayar Üniversitesi Fen Bilimleri Enstitüsü **Matematik Anabilim Dalı**'nda **YÜKSEK LİSANS TEZİ** olarak savunulmuş ve **oybirliği** ile başarılı olarak kabul edilmiştir.

Danışman

Prof. Dr. Ali KONURALP
Manisa Celal Bayar Üniversitesi

Jüri Üyesi

Prof. Dr. Necati ÖZDEMİR
Balıkesir Üniversitesi

Jüri Üyesi

Doç. Dr. Gökşen BACAK TURAN
Manisa Celal Bayar Üniversitesi

TAAHHÜTNAME

Bu tezin Manisa Celal Bayar Üniversitesi Fen-Edebiyat Fakültesi Matematik Bölümü'nde, akademik ve etik kurallara uygun olarak yazıldığını ve kullanılan tüm literatür bilgilerinin referans gösterilerek tezde yer aldığını beyan ederim.

HUSSAM ALSİTARİ



İÇİNDEKİLER

	Sayfa
İÇİNDEKİLER	I
SİMGELER VE KISALTMALAR DİZİNİ.....	II
ŞEKİLLER DİZİNİ.....	III
TABLO DİZİNİ	IV
TEŞEKKÜR.....	V
ÖZET.....	VI
ABSTRACT.....	VII
1. GİRİŞ	1
2. REGRESYON.....	4
2.1. Lineer Regresyon	4
2.2. Lojistik Regresyon	5
2.2.1. Lineer ve Lojistik Regresyon Analizi Arasındaki Farklar.....	6
2.3. İkili (Binary) Lojistik Regresyon	7
2.4. Lojistik Regresyon Analizinde Parametre Tahmini	9
2.4.1. Maksimum Olasılık Tahmini.....	9
2.4.2. Oran ve Oran Oranı (OR) Kavramları	11
2.4.3. Parametrelerin Önem Testi	12
2.4.4. Değişken Seçim	13
2.5. Modelin Değerlendirilmesi	14
2.6. Çoklu Bağlantı Problemi	17
3. MATERYAL VE YÖNTEMLER.....	19
3.1. Veri Kısaltması.....	20
4. ARAŞTIRMA BULGULARI VE TARTIŞMA	22
4.1. Lojistik Regresyon Uygulaması	22
4.1.1. Bağımsız Değişken Ekleme Yolu	23
4.1.2. Model Conformity Test	24
4.2. Lojistik Regresyon Yorumlama	25
4.2.1. Çoklu Bağlantı İçin Test Etme	35
4.2.2. Lojistik Regresyon Modelinin Yorumlanması	37
5. SONUÇ VE ÖNERİLER	38
KAYNAKLAR	39
ÖZGEÇMİŞ	41

SİMGELER VE KISALTMALAR DİZİNİ

GLM General Linear Model (Genelleştirilmiş Doğrusal Model)

LR Logistic Regression (Lojistik Regresyon)

MLE Maximum Likelihood Estimation (Maksimum Olasılık Tahmini)

OLS Ordinary Least Square (En Küçük Kareler)

OR Odds Ratio (Olasılık Oranı)

SPSS Statistical Package for the Social Sciences (Sosyal Bilimler için İstatik Paketi)

-2LL -2log Probability

% Yüzde

W² Wald İstatistiği

X₁ Uzunluk (Hasta Uzunluğu)

X₂ Hastanın Cinsiyeti

X₃ Sigara İçme Durumu

X₄ Kan Grubu

X₅ Teşhis Tarihi

X₆ Ağırlık

X₇ Yaş

Y Romatizma Hastası

ŞEKİLLER DİZİNİ

	Sayfa
Şekil 2.1.a. Kesik doğrusal ilişki.....	6
Şekil 2.1.b. S-şekilli eğri.....	6
Şekil 3.1. Çalışma verisi (İdlib Merkez Hastanesi’nde eklem ağrısı ve şüpheli romatizmal hastalıklar)	19
Şekil 4.1. Lojistik regresyon uygulamasına giriş	21
Şekil 4.2. Lojistik regresyon ana iletişim kutusu	23
Şekil 4.3. Regresyon yöntemi	24
Şekil 4.4. Bazı testler almak için iletişim kutusu	24
Şekil 4.5. Gözlenen gruplar ve olasılıklar	34
Şekil 4.6. Çoklu bağlantı için test için iletişim kutusu	36
Şekil 4.7. İletişim kutusu	36

TABLO DİZİNİ

	Sayfa
Tablo 2.1. Lineer Regresyon ve Lojistik Regresyonda model uyumuna ilişkin temel kavramlar	15
Tablo 3.1. Hastanın durumu.....	20
Tablo 3.2. Uzunluk (kategorilere göre uzunluk).....	20
Tablo 3.3. Hastanın cinsiyeti	20
Tablo 3.4. Sigara içme durumu	20
Tablo 3.5. Kan grubu	21
Tablo 3.6. Kategorilere göre ağırlık	21
Tablo 3.7. Kategorilere göre yaş	21
Tablo 3.8. Kategorilere göre çömelme	21
Tablo 4.1. Bağımlı değişken kodlama.....	25
Tablo 4.2. Kategorik değişken kodlamaları	25
Tablo 4.3. Yineleme geçmişi	26
Tablo 4.4. Sınıflandırma tablosu.....	26
Tablo 4.5. Denklemdaki değişkenler	27
Tablo 4.6. Denklemden olmayan değişkenler.....	28
Tablo 4.7. Model katsayılarının Omnibus testleri.....	29
Tablo 4.8. Sınıflandırma tablosu.....	30
Tablo 4.9. Denklemdaki değişkenler	32
Tablo 4.10. Hosmer ve Lemeshow testi.....	34
Tablo 4.11. Coefficientsa	36

TEŞEKKÜR

Çalışmamın her aşamasında bana destek olan, bilgi ve tecrübesi ile lisansüstü öğrenim hayatımın tüm zorlu aşamalarında maddi manevi her yönden yardımcı olan, tecrübeleri ile beni aydınlatan ve desteğini hiç eksik etmeyen, kendisini tanımaktan büyük onur duyduğum danışman hocam Sayın Prof. Dr. Ali KONURALP'e, öğrenim hayatım boyunca beni maddi ve manevi olarak destekleyen ve ihtiyacım olan cesareti bana vererek hep yanımda olan eşime, anneme, babama ve kardeşlerime yürekten teşekkür ederim.

Hussam ALSİTARİ
Manisa, 2023



ÖZET

Yüksek Lisans Tezi

Romatizmal Hastalıklara Sahip Bireylere Uygulanabilecek Binary Lojistik Regresyon Modelinin Geliştirilmesi

Hussam ALSİTARİ

Manisa Celal Bayar Üniversitesi
Fen Bilimleri Enstitüsü
Matematik Anabilim Dalı
Uygulamalı Matematik Bilim Dalı

Danışman: Prof. Dr. Ali KONURALP

Bu çalışmada amacımız, yüksek sınıflandırma oranına sahip ve olası en az açıklayıcı değişkenleri kullanarak romatizmal hastalığa sahip bireyler için bir lojistik regresyon modeli oluşturmaktır. Böylece bu hastalıklarda uzmanlaşmış doktor, hastalığın olasılığını ve yüzdesini verecek bir lojistik regresyon modeli kullanarak romatizmal hastalığın varlığını en düşük maliyetle ve en düşük çabayla tahmin edebilecektir. Böylece doktor dikkatle incelenmesi gereken vakaları seçebilir ve gerekli laboratuvar testlerini yapabilir. Bu testlerin çok pahalı olması ve birçok ülkenin ekonomisini de etkileyecek şekilde önemli bir yere sahip olması dolayısıyla, bu modeli kullanmak hem süreci hızlandıracak hem de uygulanması gereken hastalara uygulanmasını sağlayarak ekonomik açıdan yarar sağlayacaktır. Ayrıca, bu tez çalışmasında lojistik modelin başarısız olacağı ve modelin nasıl çözüleceği ile ilgili çeşitli sorunları tartışacağız. Lojistik modeli tamamen yok eden iki ciddi soruna, tekillik sorununa (Singularity problem) ve çoklu doğrusal bağlantı sorununa (multicollinearity problem) değineceğiz. Böylece literatüre ve bu konularda çalışan bilim insanlarına önemli bir katkı sağlamış olacağız.

Anahtar Kelimeler: Lojistik Regresyon, Romatizmal Hastalık

2023, 41 sayfa

ABSTRACT

M.Sc. Thesis

Development of Binary Logistic Regression Model that can be Applied to Individuals with Rheumatic Diseases

Hussam ALSITARI

**Manisa Celal Bayar University
Faculty of Graduate School of Applied and Natural Sciences
Department of Mathematics**

Supervisor: Prof. Dr. Ali KONURALP

Our aim in this study is to create a logistic regression model for individuals with rheumatic disease using the least possible explanatory variables with a high classification rate. Thus, the physician specializing in the diseases will be able to predict the presence of rheumatic disease at the lowest cost and with the least effort, using a logistic regression model that will give the probability and percentage of the disease. Thus, the doctor can select the cases that need careful examination and perform the necessary laboratory tests. Since the tests are very expensive and have an important place in affecting the economy of many countries, using this model will both accelerate the process and provide economic benefits by ensuring that it is applied to the patients who need it. Also, in this thesis, we will discuss the various problems that the logistic model will fail and how to solve the model. We will address two serious problems that completely destroy the logistic model, the singularity problem and the multicollinearity problem. Thus, we will make an important contribution to the literature and scientists working on these issues.

Keywords: Logistic Regression, Rheumatic Disease

2023, 41 pages

1. GİRİŞ

Regresyon, bir veya daha fazla (bağımsız, belirleyici, ortak) değişkene bağlı bir değişkenin değerini tahmin etmeye çalışan istatistiksel bir süreçtir. Bir regresyonun sonucu genellikle bağımlı ve bağımsız değişken (değişkenler) arasındaki ilişkiyi özetleyen bir modeldir. Modele, ne kadar iyi modellendiğini açıklayan özet istatistikler eşlik eder [1].

Regresyon analizi, çıkarım ve tahmin için kullanılan, iki veya daha fazla değişken arasındaki temel ilişkiyi belirlemek için kullanılan istatistiksel bir tekniktir [2].

Regresyon analizi modelleri temel olarak lineer ve lineer olmayan regresyon modeli olmak üzere ikiye ayrılır. Lineer regresyon modelleri, bağımlı (cevap değişkeni) ile bağımsız (açıklayıcı) değişkenler arasında neden-sonuç ilişkisi kurarlar ve bu değişkenler de sürekli ve sürekli değildir. Doğadaki birçok sosyal fenomen sürekli ve niceliklidir; ayrık veya nitel olarak tanımlanırlar. Bir olay meydana geldiğinde ve kişi bir seçim yaptığında, kişi birini seçer ve diğerini seçmez. Örneğin; bir kişi çocuk sahibi olabilir, gezici olabilir, evlenebilir, boşanabilir, işe girebilir, işten istifa edebilir, bir adaya oy verebilir, suç işleyebilir veya bir topluma üye olabilir. Bu olayların her biri seçim içerir [3].

İki değerle (ikili, ikili) seçerek (evet/hayır veya 0/1) seçerek veya seçmeyerek bir olay seçerken, değişkenlerimizi aralıklı (sürekli) değişkenlerle ifade ederiz. Bu durumda Lineer Regresyon modellerini kullanmak yerine Lineer Olmayan Regresyon modellerinden Lojistik Regresyon analizini kullanmamız gerekir.

Lojistik Regresyon, bir veya daha fazla açıklayıcı değişken ile ikili veya orantılı yanıt değişkenleri arasındaki ilişkiyi araştıran istatistiksel bir tekniktir. Lojistik Regresyon modelinin amacı, bir veya daha fazla (ikili ve çoklu) kategori tahmin edicisi ve bir değişken kategorik listede olduğunda cevap (bağımlı) değişkeni anlamaktır [4,5].

Lojistik işlevi, 19. yüzyılda nüfus artışını belirleme çalışmaları sırasında ortaya çıkmıştır. Malthus, 1789'da insan nüfusunun geometrik bir süreçle artacağını ortaya çıkarmıştır. Alphonse Quetelet, gözlemlerinin ötesinde rastgele tahmin edilemeyen üstel bir büyüme olduğunu fark etmiştir. Quetelet'in öğrencisi Verhulst ise, bu konu üzerinde çalışarak 1838 ile 1847 yılları arasında üç makale yayınlamıştır. Bugün ilk Verhulst kavramı üstel olarak bilinmektedir. Lojistik işlevi, 1920'de Pearl ve Reed tarafından yapılan bir nüfus artışı çalışmasında yeniden keşfedilmiştir. Lojistik büyümesinin temel fikri basit ve etkili olup, aynı zamanda günlük hayatımızda nüfus artışı, teknoloji veya yeni ürün tanıtımlarının modellenmesinde de kullanılmaktadır. Probit modellerinin keşfi Gaddum (1993) ve Bliss(1934)'e dayanmaktadır.

Gaddum, biyolojik testlerin istatistiksel yorumu ve testlerin pratik yönleri hakkında kapsamlı ve güvenilir raporlar yazmıştır [6].

Bliss, ilk olarak sigmoid eğrisini (doz-yanıt) yanıt olarak pürüzsüz bir eğriye (Probit kavramı) dönüştürme fikrini geliştirmiştir. 1952'de Finney, Probit Analizi kitabını Bliss'in fikir ve verileriyle yazmıştır [7].

Fisher ve Yates, 1963 yılındaki çalışmalarında (İstatistiksel Tablolar, Biyolojik, Tarımsal ve Tıbbi Araştırma) iki terimli parametrelerin ikili verilere dönüştürülebileceğini öne sürmüşlerdir [8].

Berkson, 1944 yılında Logit konseptini geliştirerek, hesaplamaları daha basit olmasına rağmen, en büyük olasılığın Logit'te kullanılabileceğini savunmuştur [9]. Akabinde, Fisher'ındayardımla Probit, eğrilerde en büyük olasılığı gösteren çalışmalar yapmıştır. 1944 yılında Logit yerine lojistik kullanmayı önererek 1944 ile 1980 yılları arasında bir dizi makale yayınlamıştır [6].

Berkson'un Logit kavramı ilk olarak 1949 yılında Barnard tarafından log-olasılık kavramı olarak kullanılmıştır. Barnard ve arkadaşları 1962 yılındaki çalışmalarında ilk kez Fisher tarafından geliştirilen "olasılık çıkarımı" terimini kullanmışlardır [5].

Cox, 1958 yılında Lojistik Regresyonun geliştirilmesindeki kilit figürlerden biri haline gelmiştir ve 1970 yılında İkili Verilerin Analizi'ni yazmıştır. İlerleyen yıllarda Probit modeli pazarlama ve ekonomik çalışmalarında kullanılmaya başlanmıştır.

Farrell, 1954 yılında eski araba sahiplerinin gelirlerinin bir fonksiyonu olarak bir Probit modeli üzerinde çalışmıştır. 1960 ile 1970 arasında Logit, Probit kadar ün kazanmış ve yavaş yavaş, analitik özelliklerin biyolojik çalışmaların ötesinde çok daha büyük istatistiksel işlemlere izin verdiği kabul edilmiştir [6].

İki alternatif sonuç değişkeninin analizinde kullanılmak üzere birçok dağıtım işlevi önerilmiştir. Cox ve Snell, 1989 yılında bunlardan bazılarının bir lojistik dağıtım seçmenin iki ana nedeni olduğunu ileri sürmektedir; bunlardan ilki, Matematiksel terimlerle esnek ve kullanımı kolay işlev, ikincisi ise model parametrelerinin yorumlanma kolaylığıdır [10].

Genelleştirilmiş Lineer Model kavramı ilk olarak Nelder ve Wedderburn tarafından 1972'de geliştirilmiştir. 1974 yılında Nelder, yıllar içinde GenStat olarak adlandırılan ve GLIM yazılımı olarak adlandırılan, özel olarak tasarlanmış GLM yazılımını geliştiren ilk istatistiksel araçtır. Wedderburn, Quasi-Likelihood kavramını Genelleştirilmiş Lineer modelleme algoritmasının bir uzantısı olarak geliştirmiştir [5,11].

Lojistik Regresyon, ikili değişkenlerin hesaplanmasında yaygın olarak kullanılmaktadır. Olasılık oranlarını hesaplamak ve yorumlamak kolay olduğu için yazılım programlarında rahatlıkla kullanılabilir [12]. Öyle ki, bilimin her dalında (tıp, epidemiyoloji, psikoloji, ekonomi, işletme, pazarlama, sosyal araştırma vb.) yaygın kullanıma sahiptir. Günümüz piyasasında bulunan hemen hemen tüm paket programlarında lojistik regresyon hesaplamaları yapılabilir.

İkili değişken hesabına bağlı olan lojistik regresyon, yaygın olarak kullanılan bir yöntemdir. Olasılık oranındaki hesaplama ve yorumlamanın kolay olmasından dolayı popüler olup, yazılım programlarında rahatlıkla kullanılabilir [12].

2. REGRESYON

2.1. Lineer Regresyon

Regresyon analizi, aşağıdaki formdaki gibi bağımsız bir değişken ile bağımlı bir değişken arasındaki ilişkiyi gösterir. Bu denkleme (2.1) Basit Doğrusal Regresyon denir.

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon \quad (2.1)$$

Y: Bağımlı değişken,

X₁: Bağımsız değişken,

β_0 : Kesişme parametresi; Bağımsız değişken sıfır ise bağımlı değişkenin olacağı sabit değer,

β_1 : Eğim parametresi; bağımsız değişkendeki birim değişiklik karşılığında bağımlı değişkende meydana gelecek değişim miktarıdır. Eğim parametresi, bu ilişkinin yönünü ve büyüklüğünü belirleyen eğim parametresi, ε hata terimi olarak tanımlanır.

X₁ + X₂ + + X₃'ten fazla bağımsız değişken olması durumunda, çoklu regresyonun matematiksel modeli yazılabilir.

$$Y = \beta_0 + \sum_{i=1}^k \beta_i X_i + \varepsilon \quad (2.2)$$

Menard'a göre Lineer Regresyon analizinde aşağıdaki varsayımların yapılması gerekmektedir [13]:

1. Hata ($y_i - \hat{y}_i$), gerçek rastgele değişken ile tahmini değer arasındaki farka dayanır.

Bu değer pozitif, negatif veya sıfıra eşit olabilir.

2. Hataların beklenen değerleri $E(\varepsilon_i) = 0$ 'dır.

3. σ_ε^2 Hataların varyansı olarak tanımlanır. Bağımsız değişkenlere göre değişmeyecek olan hata teriminin sabit olması (Eş varyans) beklenmektedir.

$$Var(\varepsilon_i) = E[(\varepsilon_i - E(\varepsilon_i))^2] = \sigma_\varepsilon^2$$

4. Hataların dağılımı normaldir.

$$\varepsilon_i \sim N(0, \sigma_{\varepsilon_i}^2)$$

5. Otokorelasyonun yokluğu, yani ardışık hata terimleri birbirinden bağımsızdır.

$$E(\varepsilon_i, \varepsilon_j) = 0, \quad \forall i \neq j \text{ ve } i, j = 1, 2, \dots, n$$

6. Argüman ve hata terimleri arasında bir ilişki olmaması beklenir.

$$E(\varepsilon_i X_i) = 0$$

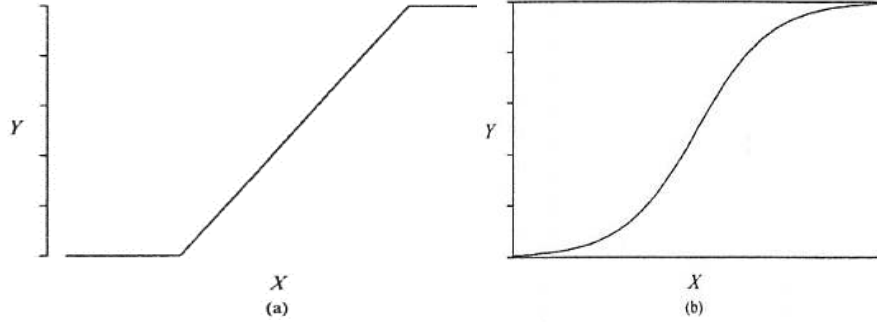
7. İkiden fazla bağımsız değişken arasındaki Lineer ilişkiye Çoklu Bağlantı Doğrusallık denir ve çoklu bağlantı olmaması arzu edilir.

Lineer Regresyonda $(\beta_0, \beta_1, \dots, \beta_k)$ En Küçük Kareler (OLS) yönteminde standartlaştırılmış regresyon katsayılarını tahmin etmek için en yaygın kullanılan yöntemdir.

2.2. Lojistik Regresyon

Lojistik Regresyon analizi, bağımsız değişkenler ile bağımlı değişken değeri arasındaki ilişkinin analizinden oluşan; bağımsız değişkenler ile bağımlı değer arasındaki ilişkiyi analiz etmek için kullanılır [4]. Lojistik Regresyon analizi, bağımlı değişkenler ile bağımsız değişkenler arasındaki neden-sonuç ilişkisinin incelenmesinde önemlidir [9]. Lojistik Regresyon analizi, istatistikte kullanılan diğer model teknikleriyle aynı amaca sahiptir ve en az bir değişken kullanarak en iyi uyumu sağlayan, kabul edilebilir bir model oluşturmayı hedefler.

Lojistik Regresyon analizin de incelenecek bağımlı değişkenler bazen iki seviyeden oluşur : pozitif-negatif, başarılı-başarısız, evet-hayır vb. Lojistik regresyon analizinde, bazı durumlarda test edilecek veriler, iki katlı "negatif pozitif", "başarılı başarısızlık" veya "evet veya hayır" gibi bağımlı değişkenler olabilir. İkili bağımlı değişkenin maksimum ve minimum değerler sırasıyla 0 ve 1'dir. Tanım olarak, olasılık 0'dan büyük ve 1'den küçüktür. Lineer Regresyondaki bağımsız değişkenlerin değeri, pozitif sonsuza veya negatif sonsuzluğa genişletilebilmektedir. Ancak Lojistik Regresyondaki eğimi ve gözlemlenen X değerine bağlı olarak, model 1 temelinde ve bağımlı değerler 0 altında tahmin edilebilmektedir. Bu değerlerin hiçbir anlamı yoktur ve tahmin yetersizdir [3].



Şekil 2.1.a. Kesik doğrusal ilişki

Şekil 2.1.b. S-şekilli eğri

Genel olarak incelenen olayın oluşumu 1 olarak ifade edilir; Doğrusal olmayan eğri, yavaş ve düzgün bir şekilde 0 ve 1'e yaklaşır. Bağımlı değişken iki sonuçlu olduğunda yanıt fonksiyonunun S şeklindeki eğrisi (Şekil 2.1.a ve Şekil 2.1.b).

Lojistik Regresyon genel olarak üç başlık altında toplanır:

1. İkili Lojistik Regresyon: Bağımlı değişken iki kategoride olduğunda gerçekleştirilir. Bağımlı değişken 0 ve 1 olarak kodlanmıştır (ölüm-hayat, başarı-başarısızlık, hiç-hayır, vb.).
2. Sıralı (Sıralı) Lojistik Regresyon: Bağımlı değişken sıralı olduğunda kullanılır. Bağımlı değişken birden fazla kategori içerir (örneğin; 0 = zayıf, 1 = kabul edilebilir, 2 = mükemmel vb.).
3. Çok Terimli Lojistik Regresyon: İki den fazla seviye içeren ve adlandırıldığı durumlarda kullanılan bağımlı değişken. Lojistik Regresyon analizini kodlayan sıra olarak belirli bir sırada veri düzenlemelerine gerek yoktur (evet-hayır-belki, evli-bekar-dul vb.).

2.2.1. Lineer ve Lojistik Regresyon Analizi Arasındaki Farklar

Lineer regresyon ve lojistik regresyon arasındaki benzerliklere rağmen, lineer regresyonu doğrudan sonuç değişkeninin kategorik olduğu bir duruma uygulayamayacağımızın iyi bir nedeni vardır. Bunun nedeni, lineer regresyon varsayımlarından birinin, değişkenler arasındaki ilişkinin lineer olmasıdır. Bu nedenle, lineer regresyonun geçerli bir model olması için, gözlemlenen verilerin lineer bir ilişki içermesi gerekir. Sonuç değişkeni kategorik olduğunda bu varsayım ihlal edilir. Bu problemi aşmanın bir yolu, logaritmik dönüşümü kullanarak veriyi

dönüştürmektir. Budönüşüm, lineer olmayan bir ilişkiyi lineer bir şekilde ifade etmenin bir yoludur. Aşağıda açıklanan (2.3) lojistik regresyon denklemi bu prensibe dayanmaktadır. Çoklu lineer regresyon denklemini logaritmik terimlerle (logit olarak adlandırılır) ifade eder ve böylece doğrusallık varsayımını ihlal etme sorununun üstesinden gelir [14].

$$P(Y) = \frac{1}{1+e^{-(b_0+b_1x)}} \text{ veya } P(Y) = \frac{1}{1+e^{-(b_0+b_1X_1+b_2X_2+\dots+b_nX_n)}} \quad (2.3)$$

(bir bağımsız değişken)

(çok bağımsız değişken)

Lojistik Regresyon analizinde bağımsız değişken(ler) dağılımı için bir ön koşul yoktur, Lineer Regresyon analizi ise bağımsız değişkenlerin çoklu normal dağılımlara sahip olmasını gerektirir [10]. Dahası, Wright'e göre, varyansların homojenliği, bağımlı değişkenin normal varyansı, hataların normal dağılımı varsayımı gerekli değildir [15].

2.3. İkili (Binary) Lojistik Regresyon

Bağımlı değişken, iki seçenekli (başarılı-başarısız, tam doğru, vb.) kategorik bir değişken ise, buna İkili Lojistik Regresyon denir. Bu tür değişkenler İkili veya Bernoulli varyantları olarak da adlandırılabilir.

İkili değişkenler pek çok alanda söz konusu bile olamaz. Örnekler arasında belirli bir hastalığın varlığı veya yokluğu, ameliyat sırasında bir hastanın ölümü veya hayatta kalması veya bir ürünün bir müşteri tarafından satın alınıp alınmadığı yer alır.

Modelimiz (2.1) göz önüne alındığında, lineer regresyon varsayımlarından türetilcek rastgele hata teriminin beklenen değeri 0'dır, burada,

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \varepsilon_i, \quad (i=1,2,\dots,n) \\ E(Y) &= E(\beta_0 + \beta_1 X_1) + E(\varepsilon_i) \\ Y &= \beta_0 + \beta_1 X_1 \end{aligned}$$

Burada, $P(Y)$ (2.3) bir Y olayının gerçekleşme olasılığını gösteriyorsa, yani $Y=1$ ise, $Y=0$, yani gerçekleşmeme olasılığı $1 - P(Y)$ olur. Buradan;

$$E(Y) = 1 \cdot P(Y) + 0 \cdot (1 - P(Y)) = P(Y) \quad (2.4)$$

elde edilir. Böylece $E(Y) = P(Y)$ olur,

$$P(Y) = b_0 + b_1 X_1 \quad (2.5)$$

elde edilir.

$X = x$ Y 'nin koşullu beklenen değeri

$$E(Y/x) = b_0 + b_1 x_1 \quad (2.6)$$

formunda yazılmıştır.

Modelde yer alan parametreler (β_0, β_1) ile lineerleştirebilirliği için koşullu beklenti değerinin dönüştürülmesi gerekmektedir.

Bir olayın olma olasılığı o olaya sahip olmama olasılığına bağlıdır $\frac{\pi(x)}{1-\pi(x)}$ denir.

Oranların bu dönüşümü Logit modeli ($g(x)$) olarak adlandırılır ve aşağıdaki şekilde temsil edilir.

$$g(x) = \text{logit} = \ln \left(\frac{P(x)}{1 - P(x)} \right)$$

$$g(x) = E(Y/x) = \frac{\exp(b_0 + b_1 x_1)}{1 + \exp(b_0 + b_1 x_1)} = \frac{e^{b_0 + b_1 x_1}}{1 + e^{b_0 + b_1 x_1}} \quad (2.7)$$

$$g(x) = [1 - \exp(b_0 + b_1 x_1)]^{-1}$$

Bunu bir matris gösterimi olarak gösterirsek,

$$\exp \left[\ln \left(\frac{P(x)}{1 - P(x)} \right) \right] = \exp(Xb)$$

$$\frac{P(x)}{1 - P(x)} = \exp(Xb)$$

$$\frac{1 - P(x)}{P(x)} = \frac{1}{\exp(Xb)}$$

$$\frac{1}{P(x)} - 1 = \exp(-Xb)$$

$$\frac{1}{P(x)} = 1 + \exp(Xb)$$

$$P(x) = \frac{1}{1 + \exp(Xb)}$$

Sonuç olacaktır [5].

$g(x)$, model parametreleriyle doğrusal ve sürekli, $-\infty$ ile $+\infty$ arasındaki değerleri alır. $P(x)$ arttıkça, $g(x)$ artar [10].

2.4. Lojistik Regresyon Analizinde Parametre Tahmini

Eliason ve Czepiel, ikili bağımlı değişken parametrelerini tahmin etmek için En Küçük Kareler regresyon analizini kullanmak için alternatif yaklaşımlardan biri olan En Küçük Olabilirlik Tahmini (Maksimum Olasılık Tahmini) kullanmayı önermektedir. En Küçük Kareler yönteminden farklı olarak, Maksimum Olasılık Tahmini, doğrusal ve doğrusal olmayan tüm modeller için kullanılabilen uygun bir yaklaşımdır [16,17].

2.4.1. Maksimum Olasılık Tahmini

En Küçük Kareler Yöntemi (EKK) ve Maksimum Olabilirlik (MO) yöntemi farklı yaklaşımlardır. Bağımlı değişken normal dağılım gösteriyorsa klasik analizde her iki yöntem de aynı sonucu verir. Ayrıca, ML tahmin (MLE) tüm Lineer ve Lineer olmayan yöntemler için uygun bir yaklaşımdır, ML yöntemi doğrusal modellerin yanı sıra Linear olmayan modellerin bileşenlerinin tahmin edilmesinde de uygulanabilir. Genel olarak MLE, Lojistik Regresyon analizinde yaygın olarak kullanılan bir tahmin yöntemidir. Çünkü Lojistik Regresyonda model lineer değildir. Ayrıca, MLE uygulandığında, bağımlı değişken sıralı, nominal veya aralıklı olmalıdır. ML eşitlik yanıt değişkeni (Y), olasılık dağılımından türetilmiştir. Her cevap, i popülasyonundaki Y varyantının binom sayısını temsil eder. Ortak olasılık yoğunluk fonksiyonu [17] başarı olasılığının $P_i^{y_i}$ olduğu bir binom dağılımında, başarısızlık $(1 - P_i)^{n_i - y_i}$ olasılık fonksiyonu Denklem (2.8)'de olduğu gibi yazılır.

$$f(y/b) = \prod_{i=1}^N \frac{n_i!}{y_i!(n_i - y_i)!} P_i^{y_i} (1 - P_i)^{n_i - y_i} \quad (2.8)$$

Binom dağılımına göre Logit modeli şu şekilde elde edilebilir:

$$\log\left(\frac{P_i}{1-P_i}\right) = \sum_{k=0}^k x_{ik} b_k \quad i = 1, 2, 3, \dots, N$$

$$\frac{P_i}{1-P_i} = e^{\sum_{k=0}^k x_{ik} b_k}$$

$$P_i = \left(\frac{e^{\sum_{k=0}^k x_{ik} b_k}}{1 + e^{\sum_{k=0}^k x_{ik} b_k}} \right) \quad (2.9)$$

Denklem (2.8) yerine Denklem (2.9) yazmak istersek, aşağıdaki gibi olabilirlik fonksiyonunu elde ederiz.

$$\prod_{i=1}^N (e^{\sum_{k=0}^k x_{ik} b_k})^{y_i} \left(1 - \frac{e^{\sum_{k=0}^k x_{ik} b_k}}{1 + e^{\sum_{k=0}^k x_{ik} b_k}}\right)^{n_i - y_i}$$

$$\prod_{i=1}^N (e^{y_i \sum_{k=0}^k x_{ik} b_k})^{y_i} (1 + e^{\sum_{k=0}^k x_{ik} b_k})^{n_i - y_i} \quad (2.10)$$

$$+ \sum_{k=0}^k x_{ik} b_k)^{n_i}$$

ML tahmin yöntemi, gözlemlenen verisetinin değerlerini gerçekleştirmeyi mümkün olan en yüksek hale getiren popülasyon parametrelerini bulmaya çalışır. Matematiksel denklemde (2.10) logaritmalar, log-olabilirlik fonksiyonu ile çalışmak daha kolay olacaktır.

$$L(b) = \sum_{i=1}^N y_i (\sum_{k=0}^k x_{ik} b_k) - n_i \log(1 + \sum_{k=0}^k x_{ik} b_k) \quad (2.11)$$

olur.

Denklem 2'deki olabilirlik fonksiyonunun birinci ve ikinci türevleri, $L(\beta)$ 'yi maksimize eden β değerleri bularak sifıra eşitlenir.

Birincitürev;

$$\frac{\partial L(b)}{\partial b_k} = \sum_{i=1}^N y_i x_{ik} - n_i \frac{1}{1 + e^{\sum_{k=0}^k x_{ik} b_k}} \frac{\partial}{\partial b_k} (1 + e^{\sum_{k=0}^k x_{ik} b_k})$$

$$\begin{aligned}
&= \sum_{i=1}^N y_i x_{ik} - n_i \frac{1}{1 + e^{\sum_{k=0}^K x_{ik} b_k}} e^{\sum_{k=0}^K x_{ik} b_k} \frac{\partial}{\partial b_k} (x_{ik} b_k) \\
&= \sum_{i=1}^N y_i x_{ik} - n_i \frac{1}{1 + e^{\sum_{k=0}^K x_{ik} b_k}} e^{\sum_{k=0}^K x_{ik} b_k} x_{ik} \\
&= \sum_{i=1}^N y_i x_{ik} - n_i P_i x_{ik}
\end{aligned}$$

İkinci türev aynı şekilde alınır şu şekilde elde edilecektir:

$$\begin{aligned}
\frac{\partial^2 L(b)}{\partial b_k \partial b_{k'}} &= \frac{\partial}{\partial b_{k'}} \sum_{i=1}^N y_i x_{ik} - n_i P_i x_{ik} \\
&= \frac{\partial}{\partial b_{k'}} \sum_{i=1}^N -n_i P_i x_{ik}
\end{aligned} \tag{17}$$

2.4.2. Oran ve Oran Oranı (OR) Kavramları

Pampl'e göre, olasılık kavramı LR parametrelerinin dönüşümün dengelir. Bu nedenle, bağımsız değişkenler, Oranlardan yüksek oranda bağımlı değişkenlerin log-olasılıklarından etkilenir. Oranlar üzerine argümanın bazı etki denklemleri gösterilmiştir [3].

$$\ln \left(\frac{P_i}{1-P_i} \right) = b_0 + b_1 X_1 \tag{2.12}$$

$$e^{\ln \left(\frac{P_i}{1-P_i} \right)} = e^{b_0 + b_1 X_1} \tag{2.13}$$

$$Odds = \left(\frac{P_i}{1-P_i} \right) = e^{b_0 + b_1 X_1}$$

$$Odds = e^{b_0} * e^{b_1 X_1} \tag{2.14} \quad [3]$$

Oranlar, oluşma olasılığı (P), Oran oranı (OR) iki oranın birbirine oranı olarak tanımlanabilir.

Newson'a göre, Odds oranı arşılaştırılabilir. Oran oranı ile Oran kavramı arasındaki temel fark, olasılık oranıdır, ancak Oran oranıdır [18].

Kleinbaum ve Klein'a göre, e^b eşittir. Odds oranı: $Y = 1$ olduğunda, tüm bağımsız değişkenler ve bu olasılık şu şekilde verilir [19]:

$$P_i = \frac{e^{b_0 + b_1 X_1}}{1 + e^{b_0 + b_1 X_1}} \quad (2.15)$$

($Y_i = 0$) X olasılığı:

$$1 - P_i = 1 - \frac{e^{b_0 + b_1 X_1}}{1 + e^{b_0 + b_1 X_1}} \quad (2.16)$$

$$1 - P_i = \frac{1}{1 + e^{b_0 + b_1 X_1}} \quad (2.17)$$

Oran değeri,

$$Odds = \left(\frac{P_i}{1 - P_i} \right) = \frac{\frac{e^{b_0 + b_1 X_1}}{1 + e^{b_0 + b_1 X_1}}}{\frac{1}{1 + e^{b_0 + b_1 X_1}}}$$

$$Odds = e^{b_0 + b_1 X_1} \quad (2.18)$$

2.4.3. Parametrelerin Önem Testi

Lojistik Regresyon analizinde, Omnibus testi ile modelin uygunluğunu test etmek için ve her bağımsız değişkenin anlamlı olup olmadığını görmek için Wald istatistiği kullanılır.

Cizek ve Fitzgerald'a göre, Wald istatistiği genellikle lojistik regresyon analizindeki tüm parametreler için istatistiksel anlamlılığını test etmek için ve aynı zamanda parametrelerin anlamlılığını test etmek için kurulan sıfır hipotezi için kullanılır [20].

$$W^2 = \left(\frac{\hat{b}}{S.E._b} \right)^2$$

Bağımsız değişken X için Lojistik Regresyon için lgilenilen parametreyi kabul edersek, $S.E._b$, b 'nin standart hata değeridir.

İlgili b parametresi ile ilgili hipotezler aşağıda gösterilmiştir.

$$H_0 : b = 0$$

$$H_1 : b \neq 0$$

Wald istatistiği ile belirlenen etkilerin anlamlı olup olmadığına göre parametreler modelden çıkarılır. Wald testi, numune hacmi büyükse anlamk azanır. Bu nedenle, bazı araştırmacıların Wald istatistiğinin kullanımıyla ilgili bazı uyarıları olduğu görülmektedir.

Menard, büyük katsayılarda standart hatanın artmasıyla Wald istatistiğinin değerinin azaldığını [21]; Agresti, Wald istatistiği yerine olasılık istatistiklerinin kullanılmasının küçük örneklem büyüklüklerinde daha uygun olduğunu belirtmiştir [9]. Ek olarak, Field, Wald istatistiği zaman zaman güvenilmez sonuçlar üretebildiğinden, olasılık oranı istatistiklerinin bu yöntemlerdeki en iyi varyans ölçüsü olduğunu vurgulamaktadır [14].

Katsayıların önemini ölçmek için kullanılan diğer bir test yöntemi Puan testidir. Halteman ve Tsiatistarafından geliştirilen Skore testi, $k-1$ serbestlik derecesi χ^2 dağılımını gösterir ve aşağıdaki denklemle hesaplanır:

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

2.4.4. Değişken Seçim

Lojistik Regresyon analizinde, tüm bağımsız değişkenlerin bağımlı değişkenlerini Doğrusal Regresyon analizinde olduğu gibi açıklamak mümkün değildir. Bu nedenle değişken tarama yöntemleri kullanılmaktadır. SPSS programında yer alan eliminasyon yöntemlerinde, Enter yöntemi tek seferde tüm değişkenleri içerir. İleri ve Geri (for ward and backward) geriye doğru adım adım (stepwise) seçim yöntemidir.

Lojistik Regresyon analizinde, tüm bağımsız değişkenlerin bağımlı değişkenlerini Lineer Regresyon analizine benzer şekilde açıklamak mümkün değildir. Lojistik Regresyon analizinde, tüm bağımsız değişkenlerin bağımlı değişkenlerini Doğrusal Regresyon analizine benzer şekilde açıklamak mümkün değildir. Bu nedenle değişken eleme yöntemleri kullanılmaktadır. Test yöntemlerinde yer alan SPSS, modele dahil edilen tüm değişken yöntemlerini aynı anda girecektir. İleri Adım, ileri adımdır ve Geri Adım, geriye doğru adım adım seçim yöntemidir. Bu adımlar, üç şekilde karar verilen değişken adına dahil edilmeyecektir. Olasılık Oranı istatistikleri - Koşul Endeksi ve üçüncü ve son kriter Wald istatistiğidir.

2.5. Modelin Değerlendirilmesi

Modelin genel uyumunun önemini test etmek için uyum iyiliği kullanılır. Lineer Regresyon modelleri için modelin uyumu için F testi ve R² kullanılmıştır. Lojistik Regresyonda seçilen parametreler için log-olabilirlik (LL) uyumu kullanılır [21].

Hipotezleri test etmek için kullanılan F testi aşağıdaki gibidir:

$$F = \left[\frac{\frac{SSR}{k}}{\frac{SSE}{(N-k-1)}} \right]$$

Hatanın Karesi Toplamı, $SSE = \sum (y_i - \hat{y}_i)^2$

Kare Regresyon Toplamı, $SSR = \sum (\hat{y}_i - \bar{y}_i)^2$

Karelerin toplamı $SST = SSR + SSE$ şeklinde hesaplanır. Aynı zamanda R^2 'nin değeri $R^2 = SSR/SST$ 'dir [21].

Lineer Regresyonlu Lojistik Regresyonda kullanılan uygunluk kriterleri Tablo 2.1'de verilmiştir. -2LL, Lojistik Regresyonda kullanılan bir model uygunluk kriteridir.

Tablo 2.1.Lineer Regresyon ve Lojistik Regresyonda model uyumuna ilişkin temel kavramlar [22]

Çoklu Regresyon	Lojistik Regresyon
Genel Kareler Toplamı	Başlangıç Modeli / Temel Model (Base model) için -2LL
Hata Kareler Toplamı	Amaçlanan Model (proposed model) için -2LL
Regresyon Kareler Toplamı	Başlangıç modeli - Amaçlanan model arasındaki -LL farkı
Model Uyumu için F Testi	-2LL farkı için χ^2 testi
Determinasyon Katsayısı (R^2)	Pseudo R^2 ölçümleri

-2LL veya -2log likelihood (-2log olasılık) bir model uyum indeksi'dir. Olasılık tahmincisinin iyi uyumu ne kadar iyi gösterdiğinin temel ölçüsü, çoklu regresyondaki karelerin toplamına benzer olan olasılık değeri'dir. Lojistik Regresyon, olasılık değerinin -2 log'sunu alarak yakalanan modelin kesişmesini ölçer. -2LL'nin alabileceği en küçük değer 0'dır ve tam eşleşme bunun tersidir. -2LL = 0 olduğunda, olasılık = 1. Sonuç olarak, -2LL değeri Çoklu Regresyonun R^2 değeri olarak düşünülebilir ve uyumdaki değişiklikler karşılaştırılarak yorumlanabilir. Olabilirlik değeri, bir modelden diğerine değişiklikler arasındaki farkın karşılaştırılmasıyla değerlendirilir[23]. Hosmer ve Lemeshow'a göre, Lojistik Regresyonda gözlenen değerleri bağımlı değişkenlerin beklenen değerleri ile karşılaştırmak için artıklar ve sapmalar kullanılır [10]. Bu, günlük olabilirlik işlevine dayanır.

D , sapma, -2LL, istatistik veya Sapma olarak adlandırılır;

$$D = -2 \ln (\text{olasılık veya takılan model}) \quad (2.20)$$

$$D = -2L(\sum_{i=1}^n [y_i \ln \left(\frac{p_i}{y_i}\right) + (1 - y_i) \ln \left(\frac{1-p_i}{1-y_i}\right)]) \quad (2.21)$$

Bir şekilde ifade edilen D istatistiği, karar verme sürecinde önemli bir role sahiptir. D istatistiği, Lineer Regresyondaki hata karelerinin toplamı ile aynı işleve sahiptir. Ön yargının minimum olması arzu edilir. Denklemden bağımsız bir model arasındaki değişim incelenir (2.21). Bu değişiklik G harfi ile gösterilir. G -istatistiği, 1 serbestlik dereceli χ^2 dağılımını gösterir.

$$G = D (\text{Değiştirilmiş model}) - D (\text{Yeni eklenen değişken modeli})$$

$$G = -2 \ln \left[\frac{\text{Possibility of model without variable}}{\text{Possibility of the variable model}} \right] \quad (2.22)$$

Hosmer-Lemeshow test istatistiği, uyumluluğu bir bütün olarak değerlendirir. Koordinasyonun uygunluğu konusunda karar verebilmek için gözlemlenen ve beklenen değerleri karşılaştırmak için geliştirilmiştir. T-2, ki-kare dağılımının serbestlik derecesini gösterir[10].

$$R_L^2 = \frac{-2LL(model)}{-2LL(orijinal)} \quad (2.23)$$

Cox ve Snell R^2 istatistikleri, çok değişkenli regresyon analizindeki R^2 istatistiğine benzer. Bağımlı değişkenlerle bağımsız değişkenler arasındaki ilişki hakkında bilgi verir.

$$R_{CS}^2 = 1 - e^{\left[-\frac{2}{n}(LL(yeni) - (LL(temel)))\right]} \quad (2.24)$$

Nagelkerke R^2 istatistiği, Cox ve Snell R^2 'nin 0-1 aralığında değerlere sahip olmasını sağlamak için geliştirilmiştir. Cox ve Snell, R^2 istatistik değerinin 1 - (sabit terimli olabilirlik modeli) $^{\frac{2}{N}}$ 'ya bölünmesiyle elde edilir. Elde edilen değer 0'a yakınsa o zaman model uygunluk eksikliğini gösterir.

McFadden R^2 istatistiği, tüm değişkenleri içeren modelin log olabilirliğinin sabit terimi modelin log-olabilirliğine oranıyla elde edilir.

$$R_N^2 = \frac{R_{CS}^2}{1 - e^{\left(\frac{2(LL(temel))}{n}\right)}} \quad (2.25)$$

2.6. Çoklu Bağlantı Problemi

Klasik Lineer Regresyon ve Lojistik Regresyon modellerinin varsayımlarından biri de bağımsız değişkenler arasında çoklu doğrusal bağlantı (multicollinearity) olmamasıdır. Çoklu lineer bağlantı problemi, çoklu regresyon analizinde, her bir gözlem için, bağımsız değişkenlerden biri veya birkaçı arasında, tam veya tama yakın lineer bir ilişki olması durumudur. Bu durum, lineer ve yansız tahminciler arasında en küçük varyansa sahip olan EKK tahmincisini ciddi derecede olumsuz etkilemektedir [14].

Çoklu lineer bağlantı kavramı, ilk kez 1934 yılında Ragnar Frisch, tarafından ortaya atılmıştır. Frisch analizlerinde, bağımlı ve bağımsız değişken ayırımına gitmeden tüm değişkenlerin hata içerdiği ve asıl problemin, gerçek değişkenler arasındaki farklı doğrusal ilişkileri tahmin etmek olduğunu ileri sürmüştür. Ekonometri literatüründe, çoklu doğrusal bağlantı probleminin kitleden kaynaklanan bir problem olmadığı, genelde örnekleme ile ilgili ve veri yetersizliğinden oluşan bir problem olduğu konusunda yaygın bir görüş vardır. Çoklu lineer bağlantı problemi bu özelliği ile diğer lineer regresyon varsayımlarından sapmalar sonucu ortaya çıkan problemlerden farklıdır.

Çoklu bağlantıya sahip olmanın bir yolu, tüm yordayıcı değişkenlerin bir korelasyon matrisini taramak ve herhangi bir korelasyonun çok yüksek düzeyde olup olmadığını görmektir (çok yüksek oranda, yukarıdaki korelasyonları kastediyorum .80 veya .90). Bu, iyi bir yöntemdir, ancak daha ince çoklu doğrusallık biçimlerini kaçırır. Neyse ki, SPSS, bir varyans enflasyon faktörü (VIF) olan çeşitli doğrusallık tanımları üretir. VIF, bir tahmincinin diğer tahmincilerle güçlü bir doğrusal ilişkiye sahip olup olmadığını gösterir. VIF'nin hangi değerinin endişeye neden olması gerektiğine dair kesin ve hızlı kurallar olmamasına rağmen, Myers, 1990 yılında 10 değerinin endişelenmek için iyi bir değer olduğunu öne sürmektedir. Dahası,

ortalama VIF 1'den büyükse, çoklu bağlantı, regresyon modelini saptırıyor olabilir [14]. VIF ile ilgili olan, tersi olan tolerans istatistiğidir ($1 / VIF$). Menard, 0.2'nin altındaki değerlerin kaygılanmaya değer olduğunu öne sürmesine rağmen, 0.1'in altındaki değerler ciddi sorunlara işaret etmektedir [21].

$$VIF = \frac{1}{1-R^2} ; \quad VIF \geq 0$$

Burada R^2 bağımsız değişken regresyon modelinin belirleme katsayısıdır [29].

- En büyük VIF 10'dan büyükse endişelenecek bir neden vardır [14].
- Ortalama VIF 1'den büyük ölçüde büyükse, bu durumda regresyon önyargılı olabilir [14].
- 0.1'in altındaki tolerans ciddi bir sorunu gösterir.
- 0.2'nin altındaki tolerans, olası bir sorunu gösterir [21,14].

Çoklu Bağlantı Problemi hakkında bir not:

Bu soruna yol açan tahmindeki önemsiz değişkenleri izole edebiliriz, ancak çalışma çok sayıda değişken içeriyorsa, 10 değişken olsun ve bu problem bu değişkenlerin dördü arasında ortaya çıktıysa, o zaman aşağıdaki düzeltici önlem yapılabilir:

Boweman ve O'connell , öngörücü değişkenlerden birini kaldırırken, eşit derecede önemli olan ve çoklu Bağlantı olmayan başka bir değişken koymanın yararlı olduğunu söylüyor. İki bilim adamı, bu sorunun hafifletilip azaltılamayacağını görmek için (mümkünse) daha fazla veri toplamayı öneriyor.

Başka bir tedavi önlemi: Bu problemde ortak olan bir grup değişken bulursak, bunun üzerinde bir faktör analizi çalışması yapmak ve hepsi için onun adına yeni bir değişken çıkarmak iyi bir fikirdir.

Tüm bu bilgiler ışığında bu çalışmadaki amacımız; Romatizmal hastalığın varlığını tahmin edebilen bir lojistik regresyon modelinin geliştirilmesidir.

3. MATERYAL VE YÖNTEMLER

Bu tez çalışmasında, İdlib Merkez Hastanesi'nde eklem ağrısı ve şüpheli romatizmal hastalıklar nedeniyle teşhis edilmiş 250 hastaya ait veriler kullanılarak lojistik regresyon modelinin çok çeşitli yorumlanmış değişkenlerle uygulanacağı lojistik bir model oluşturmaya ağırlık verilecektir Şekil (3.1) . Oluşturulan bu model çalışmada tanımlanan risk faktörlerine göre incelenecektir. (İdlib Sağlık Müdürlüğü Kayıt NF14*13)

Bu çalışmada bağımlı değişkeni romatizmal hastalık ($Y=RH$) olarak değerlendirip, bağımsız değişkenler üzerinde şu şekilde tanımlanan bir lojistik model oluşturacağız:

X_1 : Uzunluk (hasta uzunluğu)

X_2 : Hasta cinsiyeti

X_3 : Sigara içme durumu

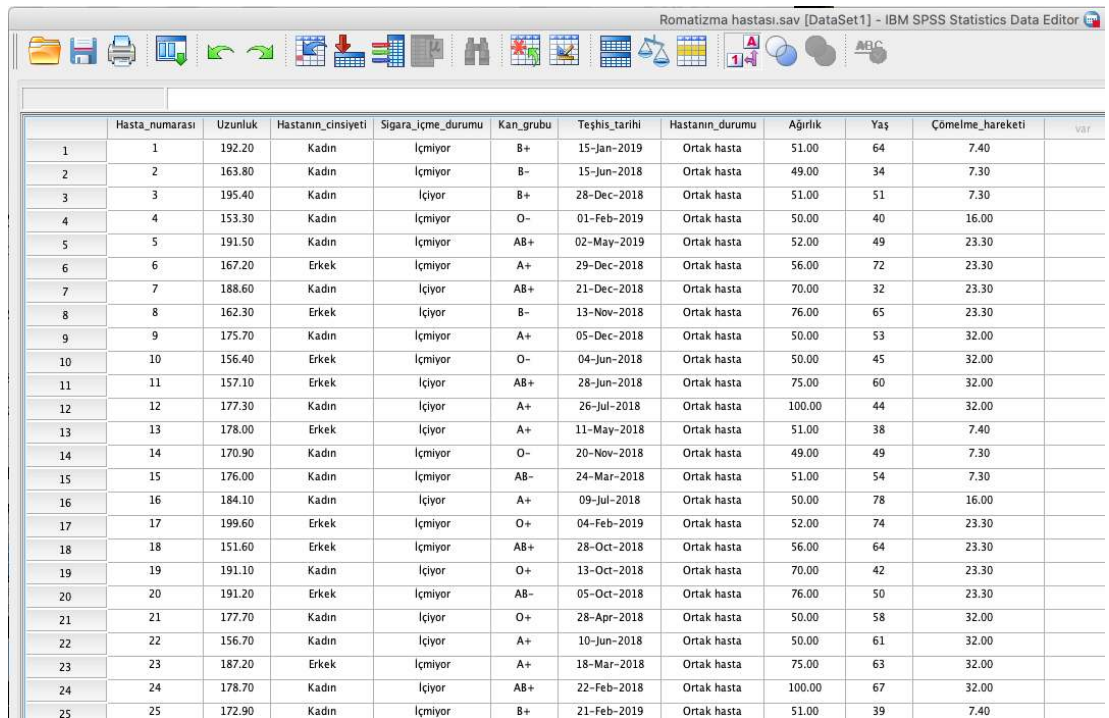
X_4 :Kan grubu

X_5 : Teşhis tarihi

X_6 : Ağırlık

X_7 : Yaş

X_8 : Çömelleme hareketi (çömelme ve tekrar yükselmek için gereken zaman)



	Hasta_numarası	Uzunluk	Hastanın_cinsiyeti	Sigara_içme_durumu	Kan_grubu	Teşhis_tarihi	Hastanın_durumu	Ağırlık	Yaş	Çömelleme_hareketi	var
1	1	192.20	Kadın	İçmiyor	B+	15-Jan-2019	Ortak hasta	51.00	64	7.40	
2	2	163.80	Kadın	İçmiyor	B-	15-Jun-2018	Ortak hasta	49.00	34	7.30	
3	3	195.40	Kadın	İçiyor	B+	28-Dec-2018	Ortak hasta	51.00	51	7.30	
4	4	153.30	Kadın	İçmiyor	O-	01-Feb-2019	Ortak hasta	50.00	40	16.00	
5	5	191.50	Kadın	İçmiyor	AB+	02-May-2019	Ortak hasta	52.00	49	23.30	
6	6	167.20	Erkek	İçmiyor	A+	29-Dec-2018	Ortak hasta	56.00	72	23.30	
7	7	188.60	Kadın	İçiyor	AB+	21-Dec-2018	Ortak hasta	70.00	32	23.30	
8	8	162.30	Erkek	İçiyor	B-	13-Nov-2018	Ortak hasta	76.00	65	23.30	
9	9	175.70	Kadın	İçmiyor	A+	05-Dec-2018	Ortak hasta	50.00	53	32.00	
10	10	156.40	Erkek	İçmiyor	O-	04-Jun-2018	Ortak hasta	50.00	45	32.00	
11	11	157.10	Erkek	İçiyor	AB+	28-Jun-2018	Ortak hasta	75.00	60	32.00	
12	12	177.30	Kadın	İçiyor	A+	26-Jul-2018	Ortak hasta	100.00	44	32.00	
13	13	178.00	Erkek	İçiyor	A+	11-May-2018	Ortak hasta	51.00	38	7.40	
14	14	170.90	Kadın	İçmiyor	O-	20-Nov-2018	Ortak hasta	49.00	49	7.30	
15	15	176.00	Kadın	İçmiyor	AB-	24-Mar-2018	Ortak hasta	51.00	54	7.30	
16	16	184.10	Kadın	İçiyor	A+	09-Jul-2018	Ortak hasta	50.00	78	16.00	
17	17	199.60	Erkek	İçmiyor	O+	04-Feb-2019	Ortak hasta	52.00	74	23.30	
18	18	151.60	Erkek	İçmiyor	AB+	28-Oct-2018	Ortak hasta	56.00	64	23.30	
19	19	191.10	Kadın	İçiyor	O+	13-Oct-2018	Ortak hasta	70.00	42	23.30	
20	20	191.20	Erkek	İçmiyor	AB-	05-Oct-2018	Ortak hasta	76.00	50	23.30	
21	21	177.70	Kadın	İçiyor	O+	28-Apr-2018	Ortak hasta	50.00	58	32.00	
22	22	156.70	Kadın	İçiyor	A+	10-Jun-2018	Ortak hasta	50.00	61	32.00	
23	23	187.20	Erkek	İçmiyor	A+	18-Mar-2018	Ortak hasta	75.00	63	32.00	
24	24	178.70	Kadın	İçiyor	AB+	22-Feb-2018	Ortak hasta	100.00	67	32.00	
25	25	172.90	Kadın	İçmiyor	B+	21-Feb-2019	Ortak hasta	51.00	39	7.40	

Şekil 3.1. Çalışma verisi (İdlib Merkez Hastanesi’nde eklem ağrısı ve şüpheli romatizmal hastalıklar)

3.1. Veri Kısaltması

Tablo 3.1. Hastanın durumu

		Frequency	Percent	ValidPercent	CumulativePercent
Valid	Ortak hasta	106	42.4	42.4	42.4
	Romatizma		57.6	57.6	100.0
	hastası	144			
	Total	250	100.0	100.0	

Tablo 3.2.Uzunluk (kategorilere göre uzunluk)

		Frequency	Percent	ValidPercent	CumulativePercent
Valid	150-160	57	22.8	22.8	22.8
	160.01-1	54	21.6	21.6	44.4
	170.01-1	47	18.8	18.8	63.2
	180.01-1	49	19.6	19.6	82.8
	190.01-2	43	17.2	17.2	100.0
	Total	250	100.0	100.0	

Tablo 3.3. Hastanın cinsiyeti

		Frequency	Percent	ValidPercent	CumulativePercent
Valid	Erkek	92	36.8	36.8	36.8
	Kadın	158	63.2	63.2	100.0
	Total	250	100.0	100.0	

Tablo 3.4.Sigara içme durumu

		Frequency	Percent	ValidPercent	CumulativePercent
Valid	İçmiyor	119	47.6	47.6	47.6

İçiyor	131	52.4	52.4	100.0
Total	250	100.0	100.0	

Tablo 3.5.Kan grubu

	Frequency	Percent	ValidPercent	CumulativePercent
Valid O-	32	12.8	12.8	12.8
O+	33	13.2	13.2	26.0
A-	34	13.6	13.6	39.6
A+	32	12.8	12.8	52.4
B-	27	10.8	10.8	63.2
B+	20	8.0	8.0	71.2
AB-	24	9.6	9.6	80.8
AB+	48	19.2	19.2	100.0
Total	250	100.0	100.0	

Tablo 3.6.Kategorilere göre ağırlık

	Frequency	Percent	ValidPercent	CumulativePercent
Valid 45-55	78	31.2	31.2	31.2
56-65	17	6.8	6.8	38.0
66-75	50	20.0	20.0	58.0
76-85	24	9.6	9.6	67.6
86-95	8	3.2	3.2	70.8
96-105	73	29.2	29.2	100.0
Total	250	100.0	100.0	

Tablo 3.7.Kategorilere göre yaş

	Frequency	Percent	ValidPercent	CumulativePercent
Valid 30-45	36	14.4	14.4	14.4
46-60	71	28.4	28.4	42.8
61-75	82	32.8	32.8	75.6
76-90	55	22.0	22.0	97.6

91-105	6	2.4	2.4	100.0
Total	250	100.0	100.0	

Tablo 3.8.Kategorilere göre çömelme

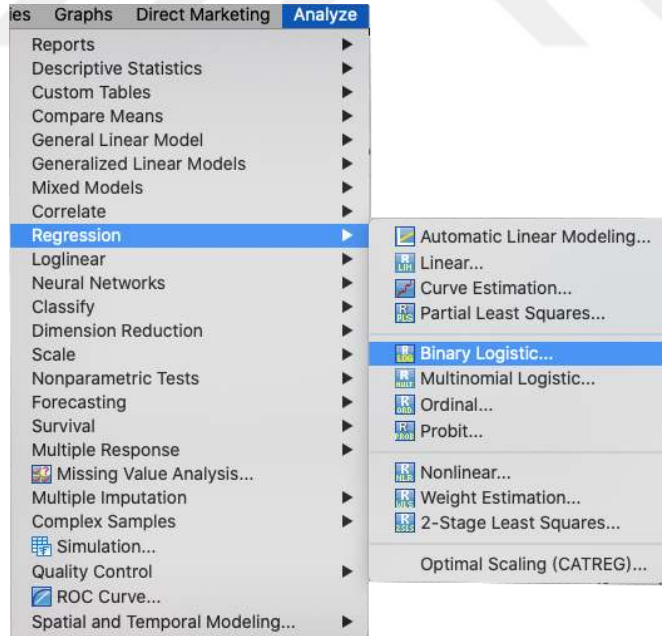
	Frequency	Percent	ValidPercent	CumulativePercent
Valid 0-10	96	38.4	38.4	38.4
10.1-20	42	16.8	16.8	55.2
20.01-30	50	20.0	20.0	75.2
30.01-40	62	24.8	24.8	100.0
Total	250	100.0	100.0	

4. ARAŞTIRMA BULGULARI VE TARTIŞMA

4.1. Lojistik Regresyon Uygulaması

Veriler SPSS editörüne girildikten sonra aşağıdaki gibi Analyze menüsüne girilir.

Analyze → Regression → Binary logistic

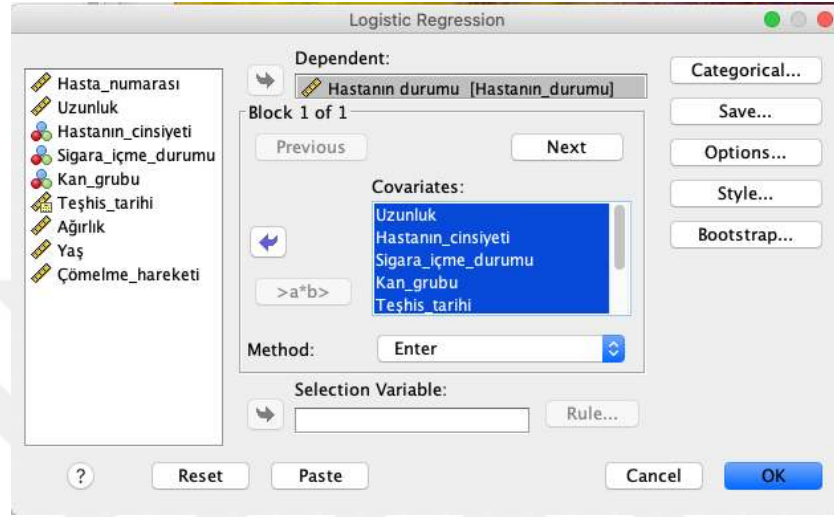


Şekil 4.1.Lojistik regresyon uygulamasına giriş

Bağımlı bir değişkeni (veya sonuç değişkenini) yerleştirmek için bir alan vardır. Bu çalışmada, Sonuç değişkeni (bağımlı değişken) bir kişinin hasta olup olmadığına bakılmaksızın Hastanın durumu üzerine tıklayıp Bağımlı kutusuna

sürükleyebilir veya üzerine tıklayabiliriz. Eş değişkenleri (tahmin değişkenleri) belirtmek için bir kutu da vardır. Ardından bağımlı değişkenleri Covariates alanına aktaracağız (Şekil 4.2).

Not: Teşhis tarihi ve hasta numarası değişkeni, çalışmada önemsiz olmaları nedeniyle Covariates'ye aktarılmayacaktır.



Şekil 4.2. Lojistik regresyon ana iletişim kutusu

4.1.1. Bağımsız Değişken Eklemenin Birkaç Yolu:

Zorunlu Giriş Yöntemi: Değişkenlerin modele zorunlu girişi.

Piramit Yöntemi: Bu şekilde, değişkenleri önceki çalışmalara dayanarak modele giriyoruz.

Adım Adım Yöntem: Programın otomatik olarak gerçekleştirdiği testlere dayalı olarak değişkenleri aşamalı olarak veya geriye doğru tanıttığı adım adım yöntem.

Forward LR yöntemini seçiyoruz. Romatizma hastası değişkeninin modelini belirlemek için İleri Adım Yöntemi (The Forward Stepwise LR) yöntemi seçilmiş ve modelin anlamlılığını test eden Omnibus testleri Tablo 4.8'de görülmektedir. Model Ki-Kare istatistiği genel olarak lojistik regresyon modelini test eder. Sabit terim dışındaki tüm Logit katsayılarının sıfıra eşit olup olmadığını test eder. Blok Ki-Kare istatistiği, model olarak bir blok değişkeni dahil ederek model Ki-Kare istatistiğindeki değişikliği gösterir.

$H_0 = \beta_1 = \beta_2 = \dots = \beta_k = 0$ Modelindeki tüm değişkenlerin katsayıları sıfıra eşittir.

$H_1 = \beta_1 \neq \beta_2 \neq \dots \neq \beta_k \neq 0$ Modelindeki tüm değişkenlerin katsayıları sıfırdan farklıdır.

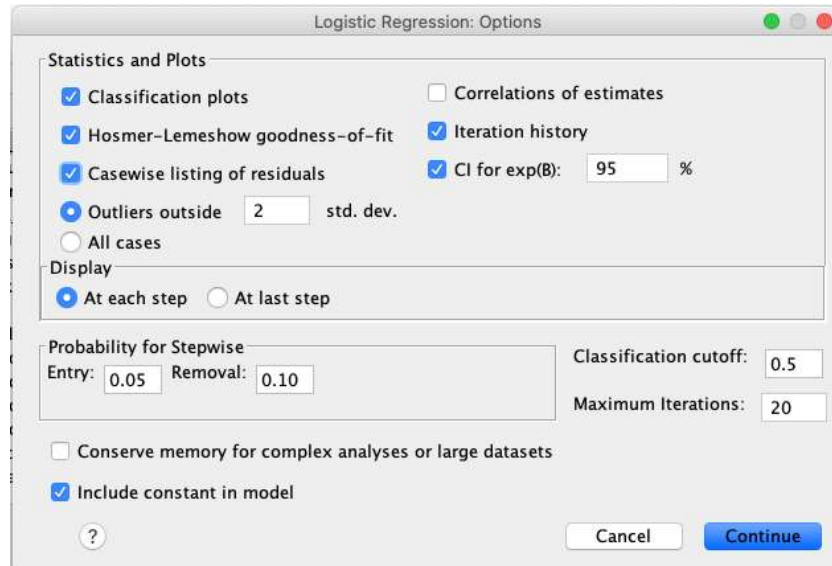
Daha sonra Tablo 6.8’de açıklayacağız.



Şekil 4.3. Regresyon yöntemi

4.1.2. Model Uygunluk Testi :

Model uygunluğu için kullanılacak testlerden biri Hosmer-Lemeshow testidir (Şekil 4.3) ve genel olarak lojistik regresyon modeli test edilir. Option listesinden seçilir.



Şekil 4.4. Bazı testler almak için iletişim kutusu

Çalışmanın sonuçlarını görmek için OK’e tıklanır.

4.2. Lojistik Regresyon Yorumlama

SPSS Tablo 4.1 bize iki şey anlatır. Önce bize sonuç değişkenimizi nasıl kodladığımızı anlatıyor ve bu tablo bize sadece 0 = Ortak hasta ve 1 = Romatizma hastası olduğunu hatırlatıyor. İkincisi, SPSS'nin kategorik öngörücüleri nasıl kodladığını bize anlatmasıdır. Parametre kodlamaları, kategorik yordayıcı değişkenler (Hastanın cinsiyeti, Sigara içme durumu, Kan grubu) için de verilmiştir (Tablo 4.2).

Tablo4.1.Bağımlı değişken kodlama

Original Value	Internal Value
Ortak hasta	0
Romatizma hastası	1

Tablo 4.2. Kategorik değişken kodlamaları

		Frequency	Parameter coding						
			(1)	(2)	(3)	(4)	(5)	(6)	(7)
Kan_grubu	O-	32	1.000	.000	.000	.000	.000	.000	.000
	O+	33	.000	1.000	.000	.000	.000	.000	.000
	A-	34	.000	.000	1.000	.000	.000	.000	.000
	A+	32	.000	.000	.000	1.000	.000	.000	.000
	B-	27	.000	.000	.000	.000	1.000	.000	.000
	B+	20	.000	.000	.000	.000	.000	1.000	.000
	AB-	24	.000	.000	.000	.000	.000	.000	1.000
	AB+	48	.000	.000	.000	.000	.000	.000	.000
Sigara içme durumu	İçmiyor	131	1.000						
	İçiyor	119	.000						
Hastanın cinsiyeti	Erkek	92	1.000						
	Kadın	158	.000						

Bu ilk analiz için ileri aşamalı bir yöntem (Forward LR) Şekil 4.2 talep ettik ve bu nedenle ilk model sadece regresyon denklemindeki sabit kullanılarak türetilti. SPSS Tablo 4.3 bize sadece sabit dahil edildiğinde model hakkında bilgi verir (yani tüm yordayıcı değişkenler çıkarılır).

Yineleme Geçmişi etiketli tablo bize bu temel modelin (log-olabilirlik) değerinin (bkz. Bölüm 2.5 Uyum iyiliği) 340.775 olduğunu söyler. Bu, en temel modelin veriye uyumunu temsil eder.

Yalnızca sabiti dahil ederken, bilgisayar modeli her katılımcıyı sonuç değişkeninin tek bir kategorisine atamaya dayandırır. Bu çalışmada SPSS, hastanın Romatizma hastası olduğunu veya her hastanın Ortak hasta olduğunu tahmin etmeye karar verebilir. Bu kararı keyfi olarak verebilir, ancak modelin gözlemlenen verileri ne kadar iyi tahmin ettiğini en üst düzeye çıkarmaya çalışmak çok önemli olduğundan, SPSS her hastanın gözlemlenen vakaların çoğunun düştüğü kategoriye ait olduğunu tahmin edecektir.

Tablo 4.3.Yineleme geçmişi

Iteration		-2 Log likelihood	Coefficients
			Constant
Step 0	1	340.775	.304
	2	340.775	.306
	3	340.775	.306

Tablo 4.4'teki çalışmada 144 romatizma hastası, 106'sı ortak hasta olan hasta vardı. Bu nedenle, SPSS her hastanın tedavi edildiğini öngörürse, bu tahmin 250'de 144 kez doğru olacaktır (yani yaklaşık %58). Ancak SPSS her hastanın ortak hasta olduğunu tahmin etseydi, bu tahmin 250 kişide sadece 106 kez doğru olurdu (yaklaşık %42). Bu nedenle, mevcut iki seçenektan tüm hastaların romatizma hastası olduğunu tahmin etmek daha iyidir. Çünkü bu daha fazla sayıda doğru tahminle sonuçlanır. Çıktı, bu temel durumdaki model için bir olasılık tablosunu gösterir.

Tablo 4.4.Sınıflandırma tablosu

		Predicted		
		Hastanın durumu		PercentageCorrect
		Ortak hasta	Romatizma hastası	
Step 0	Ortak hasta	0	106	.0
	Romatizma hastası	0	144	100.0
	OverallPercentage			57.6

Gördüğünüz gibi SPSS, tüm hastaların Romatizma hastası olduğunu tahmin etmiş, bu da Ortak hasta olan hastalarda%0, Romatizma hastası olduğu görülenlerde %100 doğrulukla sonuçlanmaktadır. Genel olarak, model hastaların %57,5'ini doğru bir şekilde sınıflandırır.

SPSS Tablo 4.5 modeli (Eşitlikteki Değişkenler) özetlemektedir ve bu aşamada bu 0,306 sabitinin (b_0) değerinin alıntılmasını gerektirir. Denklemden olmayan Değişkenler etiketli tablo bize artık ki-kare istatistiğinin 129.228 olduğunu söyler ve bu $p<.05$ 'te anlamlıdır (bu istatistiği Genel İstatistikler olarak etiketler (Overall Statistics)). Bu istatistik bize modelde yer almayan değişkenlerin katsayılarının sıfırdan önemli ölçüde farklı olduğunu, başka bir deyişle, bu değişkenlerden bir veya daha fazlasının modele eklenmesinin modelin tahmin gücünü önemli ölçüde etkileyeceğini söyler. Kalan ki-kare olasılığı .05'ten büyük olsaydı, modelden hariç tutulan tüm değişkenlerin modele zorlanmasının modelin tahmin gücüne önemli bir katkı yapmayacağı anlamına gelirdi.

Tablo 4.5.Denklemdaki değişkenler

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.306	.128	5.731	1	.017	1.358

Tablo 4.6.Denkleimde olmayan deęişkenler

		Score	df	Sig.
Step 0	Variables			
	Uzunluk	.005	1	.944
	Hastanincinsiyeti(1)	.630	1	.427
	Sigaraicmedurumu(1)	33.373	1	.000
	Kangrubu	9.235	7	.236
	Kangrubu(1)	.361	1	.548
	Kangrubu(2)	.141	1	.708
	Kangrubu(3)	.350	1	.554
	Kangrubu(4)	.047	1	.828
	Kangrubu(5)	1.107	1	.293
	Kangrubu(6)	.060	1	.806
	Kangrubu(7)	4.392	1	.036
	Aęırlık	73.377	1	.000
	Yaş	13.683	1	.000
	Çömelmehareketi	28.834	1	.000
	OverallStatistics	129.228	13	.000

SPSS Tablo 4.7; yeni modelle ilgili özet istatistikleri gösterir (daha önce gördüğümüz The rest of the variables deęişkenlerin geri kalanı içerir). Yeni modelin genel uyumu, mantıksal olabilirlik istatistięi kullanılarak deęerlendirilir (bkz. Bölüm 2.5 Uyum iyilięi). SPSS'de, log-olabilirliğini bildirmek yerine, deęer -2 ile çarpılır (ve bazen $-2LL$ olarak anılır): bu çarpma yapılır çünkü $-2LL$ yaklaşık olarak ki-kare daęılımına sahiptir ve bu nedenle deęerleri tek başına şans eseri elde etmeyi bekleyebileceğimiz deęerlerle karşılaştırmayı mümkün kılar. Log-olabilirlik istatistięinin büyük deęerlerinin, uygun olmayan istatistiksel modelleri gösterdiğini unutmayın. Analizin bu aşamasında, $-2LL$ 'nin deęeri, modele yalnızca sabitin dahil edildięi deęerden daha düşük olmalıdır (çünkü $-2LL$ 'nin düşük deęerleri, modelin sonuç deęişkenini daha doęru tahmin ettięini gösterir). Sadece sabit dahil edildiğinde, $-2LL = 340.775$ (bkz. Tablo 4.3) ama şimdi (sigara içme durumu, aęırlık, yaş, çömelme hareketi) dahil edildiğinde bu deęer 170.513'e düşürülmüştür.

Bu azalma, bize modelin tahmin etmede daha iyi olduğunu söylüyor. Modelin sonuç değişkenini ne kadar daha iyi tahmin ettiği sorusu, şu anda olduğu gibi model ile sadece sabitin dahil edildiği model arasındaki farkı ölçen model ki-kare istatistiği (the model chi-square statistic) kullanılarak değerlendirilebilir. Bölüm 2.5 Uyum iyiliği 'de yeni modelin log-olasılığını alarak ve ondan temel modelin log-olabilirliğini çıkararak bir modeldeki değişikliğin önemini değerlendirebileceğimizi gördük. Ki-kare istatistiği modelinin değeri bu prensibe göre çalışır ve şu şekildedir: Yeni model için -2LL eksi orijinal model için -2LL ($340.775 - 170.513 = 170.262$). Bu değer ki-kare dağılımına sahiptir ve bu nedenle istatistiksel önemi kolayca hesaplanabilir. Bu çalışmada değer 0,05 düzeyinde anlamlıdır ve bu nedenle modelin genel olarak bir hastanın Romatizma hastası olup olmadığını tahmin ettiğini söyleyebiliriz, yalnızca sabitin dahil edildiğinden önemli ölçüde daha iyidir.

Modelin tahmin gücündeki son aşamada bu yana gelişmeyi gösteren step statistic adı verilen ikinci bir istatistik var. üç veya dört aşamanın olduğu karmaşık modellerde bu istatistik, son adımdan bu yana modelin tahmin gücündeki iyileşmenin bir ölçüsünü verir. İyileştirme istatistiği önemliyse, modelin artık sonucu son adımda olduğundan önemli ölçüde daha iyi tahmin ettiğini gösterir ve ileriye doğru bir regresyonda bu, bir tahmincinin tahmin gücüne katkısının bir göstergesi olarak alınabilir. Model benzer şekilde, block istatistiği, son blocktan bu yana olan değişikliği sağlar.

Tablo 4.7. Model katsayılarının Omnibus testleri

		Chi-square	df	Sig.
Step 1	Step	81.529	1	.000
	Block	81.529	1	.000
	Model	81.529	1	.000
Step 2	Step	56.068	1	.000
	Block	137.598	2	.000
	Model	137.598	2	.000
Step 3	Step	21.151	1	.000
	Block	158.748	3	.000
	Model	158.748	3	.000
Step 4	Step	11.514	1	.001
	Block	170.262	4	.000
	Model	170.262	4	.000

SPSS Tablo 4.8; Mevcut model, ortak hasta olan 84 hastayı doğru şekilde sınıflandırmakta, ancak 22 hastayı yanlış sınıflandırmaktadır (vakaların%79,2'sini doğru şekilde sınıflandırmaktadır). Model ayrıca romatizma hastası olan 125 hastayı doğru bir şekilde sınıflandırır, ancak diğer 19 hastayı yanlış sınıflandırır (vakaların%86,6'sını doğru bir şekilde sınıflandırır). Sınıflandırmanın genel doğruluğu, bu nedenle, bu iki değerin ağırlıklı ortalamasıdır (%83,6). Dolayısıyla, yalnızca sabit dahil edildiğinde, model hastaların%57,6'sını doğru bir şekilde sınıflandırdı, ancak şimdi, değişkenlerin geri kalanı bir deyskanlar olarak dahil edildiğinde, bu%83,6'ya yükseldi.

Tablo 4.8. Sınıflandırma tablosu

Observed			Predicted		
			Hastanın durumu		Percentage Correct
			Ortak hasta	Romatizma hastası	
Step 1	Hastanın durumu	Ortak hasta	70	36	66.0
		Romatizma hastası	25	119	82.6
	OverallPercentage				75.6
Step 2	Hastanın durumu	Ortak hasta	88	18	83.0
		Romatizma hastası	25	119	82.6
	OverallPercentage				82.8
Step 3	Hastanın durumu	Ortak hasta	83	23	78.3
		Romatizma hastası	23	121	84.0
	OverallPercentage				81.6
Step 4	Hastanın durumu	Ortak hasta	84	22	79.2
		Romatizma hastası	19	125	86.8
	OverallPercentage				83.6

Çıktının bir sonraki bölümü (SPSS Tablo 4.9) çok önemlidir.Çünkü bize modele dahil edilen yordayıcılar için katsayıların tahminlerini söyler. Çıktının bu bölümü bize bu noktada modele dahil edilen değişkenlerin katsayılarını ve istatistiklerini verir (yani sigara içme durumu, ağırlık, yaş, çömelme hareketi ve sabiti).

B-değeri, lineer regresyondaki b-değeri ile aynıdır: bunlar, bir durumun belirli bir kategoriye girme olasılığını belirlemek için denklem (2.3) 'de değiştirmemiz gereken değerlerdir.

Lineer regresyonda, b'nin değerinin, tahmin değişkenindeki bir birim değişiklikten kaynaklanan sonuçtaki değişikliği temsil ettiğini gördük. Bu katsayının lojistik regresyondaki yorumu, yordayıcı değişkendeki bir birimlik bir değişiklik ile ilişkili sonuç değişkeninin Doğal logaritma(Logit) değişimini temsil etmesi açısından çok benzerdir. Sonucun Logit'i, Y'nin meydana gelme olasılığının doğal logaritmasıdır. Önemli istatistik, ki kare dağılımına sahip olan ve bize bu öngörünün b katsayısının sıfırdan önemli ölçüde farklı olup olmadığını söyleyen Waldistatistiğidir.Katsayı sıfırdan önemli ölçüde farklıysa, tahmin edicinin sonucun tahminine (Y) önemli bir katkı yaptığını varsayabiliriz. Bölüm 2.4.3'te Wald istatistiğine rastladık ve dikkatli kullanılması gerektiğini gördük.Çünkü regresyon katsayısı (b) büyük olduğunda, standart hata şişirme eğilimindedir ve Wald istatistiğinin hafife alınmasına neden olur (bkz.Menard,1995).Bununla birlikte, bu veriler için, müdahalenin yapılmasının (veya yapılmamasının) hastanın Romatizma hastası olup olmadığının önemli bir yordayıcısı olduğuna işaret ediyor gibi görünmektedir (Wald istatistiğinin öneminin (Sig) .05'ten az olduğuna dikkat edin). Bölüm (2.5) 'de Hosmer ve Lemeshow'un ölçüsünün (R_L^2), modelin ki-kare'yi orijinal $-2LL$ 'ye bölerek hesaplandığını gördük.

$$R_L^2 = \frac{-2LL(model)}{-2LL(orijinal)} = \frac{170.262}{340.775} \approx 0.5$$

Yani modeldeki değişkenler (sigara içme durumu, yaş, ağırlık, çömelme hareketi) romatizmal hastalık insidansının %50'sini açıklamaktadır.

Tablo 4.9. Denklemdaki değişkenler

ccc		B	S.E.	Wald	df	Sig.	Exp (B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1a	Ağırlık	.067	.009	58.462	1	.000	1.070	1.051	1.088
	Constant	-4.524	.628	51.832	1	.000	.011		
Step 2b	Ağırlık	.098	.013	57.274	1	.000	1.103	1.075	1.131
	Çömelmehareketi	-.143	.023	37.696	1	.000	.867	.828	.907
	Constant	-3.956	.709	31.169	1	.000	.019		
Step 3c Sigara içme durumu(1)	Ağırlık	.099	.014	50.457	1	.000	1.104	1.074	1.135
	Çömelmehareketi	-.152	.025	35.882	1	.000	.859	.818	.903
	Constant	-4.702	.801	34.491	1	.000	.009		
	Sigara içme durumu(1)	1.672	.381	19.206	1	.000	5.321	2.519	11.236
Step 4d Sigara içme durumu(1)	Ağırlık	.104	.015	48.911	1	.000	1.110	1.078	1.143
	Yaş	.044	.014	10.416	1	.001	1.045	1.017	1.073
	Çömelmehareketi	-.155	.027	33.784	1	.000	.856	.813	.902
	Constant	-7.757	1.350	33.041	1	.000	.000		
	Sigara içme durumu(1)	1.634	.395	17.073	1	.000	5.124	2.360	11.122
	Yaş	.044	.014	10.416	1	.001	1.045	1.017	1.073


 ≤ 0.05

Bu nedenle, SPSS tarafından seçilen dört değişkeni kabul edeceğiz ve bunları modele aşağıdaki gibi yerleştireceğiz:

$$P(\text{Romatizma hastası}) = \frac{1}{1 + e^{-(-7.757 + 1.634X_3 + 0.104X_6 + 0.044X_7 - 0.155X_8)}}$$

X₃ Sigara içme durumu

X₆ Ağırlık

X₇ Yaş

X₈ Çökelme hareketi

Bakmamız gereken son şey, bölüm 2.4.2'de açıklanan olasılık oranıdır (SPSS çıktısındaki Exp (B)). Bu çalışmanın yordayıcıdaki birim değişikliğinden kaynaklanan Odds'deki değişikliği hesaplamak için, öncelikle modelde bağımsız değişkenlere sahip olmadıkları için romatizma hastası olan bir hastanın Odds'ni hesaplamalıyız. Daha sonra romatizma hastasının olasılıklarını bağımsız değişkenlerle hesaplıyoruz. Son olarak, bu iki olasılıktaki orantılı değişikliği hesaplıyoruz.

İlk olasılık kümesini hesaplamak için, bağımsız değişkenler olmadığından hastanın romatizma hastası olma olasılığını hesaplamak için Denklem (2.3)'ü kullanmamız gerekir.

$$P(\text{Romatizma hastası}) = \frac{1}{1 + e^{7.76}} \approx 0.0004$$

$$P(\text{Romatizma hastası değil}) \approx 1 - 0.0004 = 0.9996$$

$$Odds = \frac{0.0004}{0.9996} \approx 0.0004$$

Şimdi, tahmin değişkeni her bağımsız değişken için bir birim değiştikten sonra aynı şeyi hesaplıyoruz. Sigara içmeye başlasın

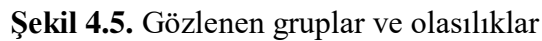
$$P(\text{Romatizma hastası}) = \frac{1}{1 + e^{7.76 - (1.634)(1)}} \approx 0.002$$

$$P(\text{Romatizma hastası değil}) \approx 1 - 0.002 = 0.998$$

$$Odds = \frac{0.002}{0.998} \approx 0.002$$

$$\Delta Odds = \frac{0.002}{0.0004} \approx 5$$

Hastanın sigara içtiğini bilerek romatizma hastalığı olasılığının bu değişken olmadan romatizma olma olasılığının 5 kat daha az olduğunu söyleyebiliriz. Kırmızı çerçeveye denk gelen Tablo 4.9'daki olasılık oranına ($\Delta Odds$)dikkat edin. Seçeneklerde, olasılık oranı için bir güven aralığı talep ettik ve bu aynı zamanda çıktı Tablo 4.9'da bulunabilir.



Not: Romatizma hastaları R sembolü ile ve romatizma olmayan O sembolü ile sembolleştirilir.

Tablo 4.10.HosmerveLemeshowtesti

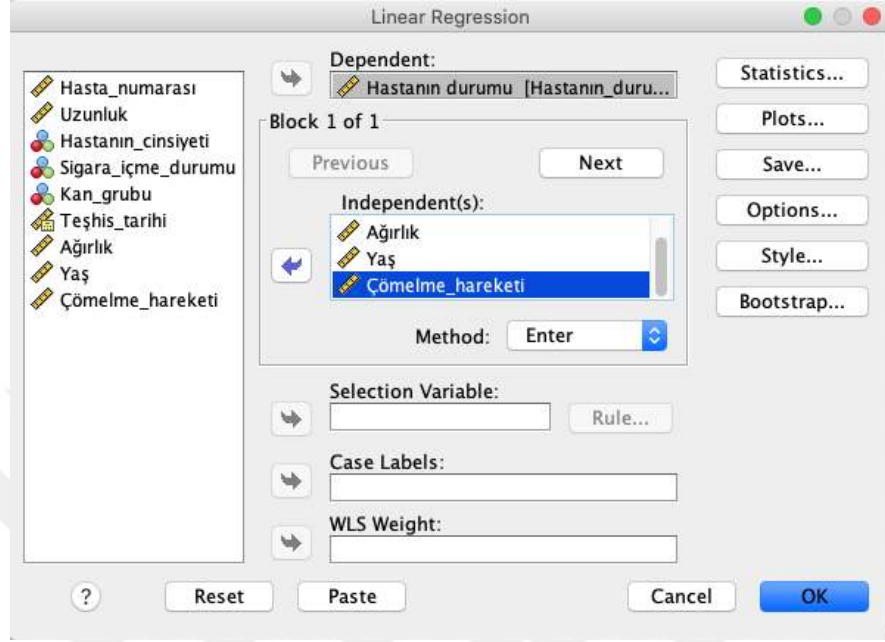
Step	Chi-square	df	Sig.
1	71.433	7	.000
2	42.613	8	.000
3	22.244	8	.004
4	5.204	8	.736

Model uygunluğu için kullanılacak testlerden biri Hosmer-Lemeshow testidir (Tablo 4.10) ve lojistik regresyon modeli genel olarak test edilir. Hosmer ve Lemeshowchi kare uyum iyilik testi, geleneksel ki kare yöntemi ile hesaplanan Omnibus testine daha güçlü bir alternatiftir. Hosmer ve Lemeshow ki kare uyum iyiliği (Tablo 4.10) Model Ki-Kare = 5.204 ve Önem Seviyesi(sig) eşittir 0.736. Veri hizalaması için Lojistik Regresyon modelinin yeterli olduğuna karar verildi (sig<0.05).

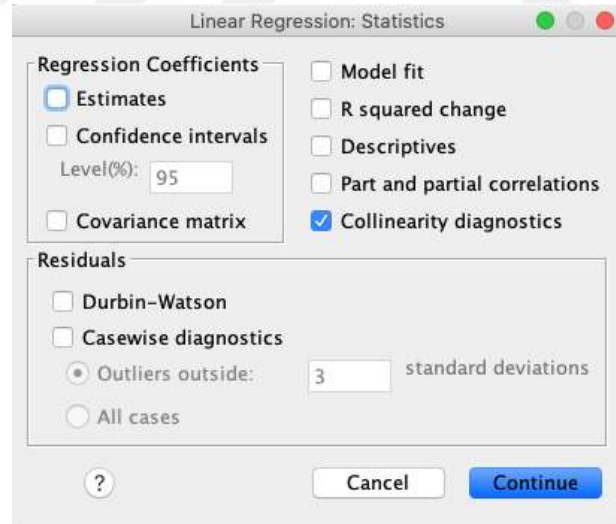
4.2.1. Çoklu Bağlantı İçin Test Etme:

Bölüm 2.6'da çoklu doğrusallığın bir regresyon modelinin parametrelerini nasıl etkileyebileceğini gördük. Lojistik regresyon, eşdoğrusallığın önyargı etkisine aynı derecede eğilimlidir ve bir lojistik regresyon analizini takiben eşdoğrusallığı test etmek önemlidir. Ne yazık ki, SPSS'nin lojistik regresyonda doğrusallık teşhisi üretme seçeneği yoktur (bu, test etmenin gereksiz olduğu yanılsamasını yaratabilir!). Bununla birlikte, tolerans ve VIF gibi istatistikleri, aynı sonucu ve öngörücüleri kullanarak yalnızca doğrusal bir regresyon analizi çalıştırarak elde edebilirsiniz. **Analyze -Regression- Linear** seçerek Doğrusal Regresyon iletişim kutusuna erişin.Tamamlanan iletişim kutusu Şekil 4.6.'de gösterilmektedir. Nihai modelde seçilen değişkenleri tanımlıyor ve onları Independent bölümüne taşıyoruz. Çalışma departmana aktarılır Dependent. Şekil 4.7'de çok sayıda seçenek belirtmek gereksizdir (bu tekniği yalnızca doğrusallık testlerini elde etmek için kullanıyoruz), ancak Statistic seçeneğine tıklayıp ardından iletişim kutusunda (Collinearitydiagnostics) ögesini seçmeniz önemlidir. Tüm varsayılan seçenekleri

kapatın, Doğrusal Regresyon iletişim kutusuna dönmek için tıklayın ve ardından analizi çalıştırmak için Ok üzerine tıklayın.



Şekil 4.6.Çoklu bağlantı için test için iletişim kutusu



Şekil 4.7.İletişim kutusu

Lineer regresyon analizinin sonuçları SPSS Tablo 4.11'de gösterilmektedir.

Tablo 4.11.Coefficientsa

Model		Collinearity Statistics	
		Tolerance	VIF
1	Sigara içme durumu	.938	1.066
	Ağırlık	.948	1.055
	Yaş	.982	1.019
	Çökme hareketi	.987	1.013

Bölüm 2.6'da doğrusallığı değerlendirmek için çeşitli kriterler gördük. Özetlemek gerekirse, Menard (1995) 0.1'den küçük bir tolerance değerinin neredeyse kesinlikle ciddi bir doğrusallık problemine işaret ettiğini öne sürmektedir. Myers (1990) ayrıca 10'dan büyük bir VIF değerinin endişe kaynağı olduğunu ve bu verilerde değerlerin hem endişeli hem de önceki için 70'in üzerinde olduğunu öne sürmektedir. VIF değeri tüm değişkenler için 10'dan küçük olduğundan çoklu doğrusallık problemine düşme korkusu yoktur.

4.2.2. Lojistik Regresyon Modelinin Yorumlanması

Modelden memnun kaldığımızda, tahmin için kullanılabilir. Bu model için tahmin şöyle olacaktır:

$$P(Y) = \frac{1}{1 + e^{-(7.757 - 1.634X_3 + 0.104X_6 + 0.044X_7 - 0.155X_8)}}$$

Bu tahmin denklemini kullanarak, bir hasta aşağıdakilere sahip olduğu için romatizma ile tahmin edilebilir:

Sigara durumu: sigara içiyor yani $X_3=1$

Ağırlık : 85 kg yani $X_6 = 85$

Yaş :70 sene yani $X_7 = 70$

Çökme hareketi: 30 saniye yani $X_8 = 30$

Bu değerleri denkleme yerleştirmek aşağıdakileri verir:

$$P(\text{Romatizma hastası}) = \frac{1}{1 + e^{-(7.757 + 1.634*1 + 0.104*85 + 0.044*70 - 0.155*30)}} = 0.651$$

Böylelikle bu bilgilerin varlığında hastanın romatizma geçirme şansı%65,1'dir. Örneğin hasta sigara içmiyorsa, yüzde aşağıdaki gibi olur:

$$P(\text{Romatizma hastası}) = \frac{1}{1 + e^{-(-7.757+1.634*0+0.104*85+0.044*70-0.155*30)}} = 0.267$$

Romatizma hastası olma riskinin düşük olması, sigara içmenin enfeksiyon olasılığı üzerinde önemli bir etkisinin olduğunu gösterir.

5. SONUÇ VE ÖNERİLER

Çalışmaya, kategorik bir sonucumuz olduğunda neden lineer regresyon kullanamadığımızı, bunun yerine ikili lojistik regresyon (iki sonuç kategorisi) kullanmamız gerektiğini inceleyerek başladık. Daha sonra, regresyon denklemine ve ne anlama geldiğine bakarak bazı lojistik regresyon teorisine baktık. Daha sonra modeli değerlendirmeye geçtik ve log-olabilirlik istatistiği ve ilgili ki-kare testi hakkında konuştuk. Regresyonda R^2 'nin eşdeğerlerini elde etmenin farklı yöntemlerinden bahsedildi (Hosmer&Lemeshow, Cox&Snell ve Nagelkerke). Wald istatistiğini ve Exp(B)'yi de keşfettik. Çalışmanın geri kalanı, lojistik regresyonları gerçekleştirmek için SPSS kullanan örneğe baktık.

KAYNAKLAR

- [1] Guido, J. J., Winters, P. C., Rains, B. A., 2014. LogisticRegression Basics. University of Rochester Medical Center, Rochester, NY. <http://www.lexjansen.com/nesug/nesug06/an/da26.pdf>.
- [2] Campbell, D ., Campbell, S., Introduction to regression and data analysis.In: Statlab Workshop Series. 2008, 1-14.
- [3] Pampel, F.C., Logistic regression a primer. Sage University Paper series on Quantitative Applications in the Social Sciences. 2000, 96.
- [4] Walker, J., Methodology application: logistic regression using the codes data. Developed for department of transportation, National Highway Traffic Safety Administration (NHTS). 1996, <http://www.nrd.nhtsa.dot.gov/Pubs/96843.pdf> . Access time : 12.02.2016.
- [5] Hilbe, J.M., Logistic regression models. Chapman & Hall/CRC. 2009, 656.
- [6] Cramer, J.S. The origins of logistic regression. Tinbergen Institute Discussion Paper. 2002, 119, 4.
- [7] Vincent, K., Probit analysis San Francisco State University, San Francisco: 2008, <http://userwww.sfsu.edu/efc/classes/biol7110/probit/ProbitAnalysis.pdf>. Access date : 02.06.2016.
- [8] Fisher, R.A., Yates, F. Statistical tables for biological, Agricultural and Medical Research. 1963, 155.
- [9] Agresti, A. An introduction to categorical data analysis. Wiley, Florida, United States, 1996, 744.
- [10] Hosmer, D.W., Lemeshow, S., Application of logistic regression. Johnson Wiley & Sons, New York, 2000, 528.
- [11] Nelder, J.A., Wedderburn, R.W.M., Generalized linear models. Journal of the Royal Statistical Society. Series A (General). 1972, 135(3), 370-384.
- [12] Hosmer, D.W., Hosmer, T., Lemeshow, S., Lemeshow, S., A comparison of goodness-of-fit tests for the logistic regression model. Statistics in Medicine. 1997, 16, 965-980.
- [13] Menard, S., Applied logistic regression analysis (quantitative applications in the social sciences). Sage Publications, 2002, 120.
- [14] Field, A., Discovering Statistics using SPSS. SAGE Publication. 2005, 779.

- [15]Wright, R.E., LogisticRegression. American Psychological Association, Washington, 1995, 217-244.
- [16]Eliaison, S.R.,Maximum likelihoo destimation logicandpractice. Sage University Paper Series on Quantitative Applications in the SocialSciencies. 1993, 132.
- [17] Czepiel, S.A. Maximum likelihoo destimation of logistic regression models the oryan dimplementation.2002, 83, <http://czep.net/stat/mlelr.pdf>. Access date: 02.02.2016.
- [18]Newsom, J., Data analysis II: logistic regression. 2003, Fall2015. http://web.pdx.edu/~g3jn/da2/ho_logistic.pdf. Access date : 02.02.2016
- [19]Kleinbaum, D.G., Klein, M., Logistic regression: a self-learningtext.USA: Springer. 2002, 513.
- [20]Cizek, G.J., Fitzgerald, S.M. An introduction to logistic regression. Measurementand Evaluation in Counselingand Development. 1999, 31(4), 223-245.
- [21] Menard, S., Applied logistic regression analysis on quantitative applications in the socialsciences. Sage Publications. 1995, 106.
- [22] Çokluk, Ö., Lojistik regresyon analizi: kavram ve uygulama. Eğitim Bilimleri: Teorik ve Uygulama. 2010, 10, 3, 1357-1407.
- [23]Hair, J.F., Black, W.C., Babin, B., Anderson, R.E., Tatman, R.L. Multivariate Data Analysis. UpperSaddleRiver. 2006, 497.
- [29]Abdul Rahman İsmail,King FahdNational Library, Kingdom of SaudiArabia2001
- [30] Hussam Alsitari , Abdulrahman Alaita, Creating a Binary Logistic Regression Model for Predicting Outcomes in Individuals with Rheumatic Diseases, 2023