**T.C.**
**YILDIZ TEKNİK ÜNİVERSİTESİ**
**FEN BİLİMLERİ ENSTİTÜSÜ**

# LOJİSTİK REGRESYON VE BANKACILIK VERİLERİ ÜZERİNE BİR UYGULAMA

**ELMİRA KOCABAŞ**

**YÜKSEK LİSANS TEZİ**
**İSTATİSTİK ANABİLİM DALI**
**İSTATİSTİK PROGRAMI**

**DANIŞMAN**
**DOÇ. DR. FİLİZ KARAMAN**

**İSTANBUL, 2014**

**T.C.**
**YILDIZ TEKNİK ÜNİVERSİTESİ**
**FEN BİLİMLERİ ENSTİTÜSÜ**

# LOJİSTİK REGRESYON

Elmira KOCABAŞ tarafından hazırlanan tez çalışması 03.03.2014 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

**Tez Danışmanı**

Doç. Dr. Filiz Karaman
Yıldız Teknik Üniversitesi

**Jüri Üyeleri**

Doç. Dr. Filiz KARAMAN
Yıldız Teknik Üniversitesi _____

Yrd. Doç. Dr. Serpil KILIÇ
Yıldız Teknik Üniversitesi _____

Yrd. Doç. Dr. Y. Barış ALTAYLIGİL
İstanbul Üniversitesi _____

# ÖNSÖZ

Çalışmamın her aşamasında karşılaştığım zorlukların giderilmesinde anlayış ve desteğini esirgemeyen değerli hocam Doç. Dr. Filiz Karaman'a ve tüm bölüm hocalarıma en içten duygularımla sonsuz teşekkürlerimi sunarım.

Bugüne kadarki başarılarımın asıl mimarları olan sevgili eşime, aileme ve fikirleriyle her zaman katkı sunan arkadaşlarıma koşulsuz bir biçimde verdikleri destek, sevgi ve saygılarından ötürü teşekkürü bir borç bilirim.

Mart, 2014

Elmira KOCABAŞ

# İÇİNDEKİLER

Sayfa

# ŞEKİL LİSTESİ

# ÇİZELGE LİSTESİ

# ÖZET

## LOJİSTİK REGRESYON VE BANKACILIK VERİLERİ ÜZERİNE BİR UYGULAMA

Elmira KOCABAŞ

İstatistik Anabilim Dalı

Yüksek Lisans Tezi

Tez Danışmanı: Doç. Dr. Filiz KARAMAN

Bankaların faaliyetlerinin devamlılığı üstlendikleri kredi riskinin etkin şekilde yönetimine dayanır.

Bu çalışma ile kredi başvurusunda bulunan müşterilerin gelecekte iyi ya da kötü olma olasılıklarının tahmin edilmesi amaçlanmış ve etkili değişkenler belirlenmiştir.

İlk olarak, çok değişkenli istatistiksel analiz yöntemleri ve bunlardan biri olan lojistik regresyon yöntemi ile istatistiksel modelleme konusuna yer verilmiştir.

Sonrasında ise bankacılık verileri ile bir uygulama yapılarak yorumlanmıştır.

Lojistik regresyon modeli SPSS 15.0 kullanılarak geliştirilmiştir.

**Anahtar Kelimeler:** Lojistik Regresyon, skorkart, modelleme, ikili lojistik regresyon

## ABSTRACT

# LOGISTIC REGRESSION AND AN APPLICATION OF BANKING DATA

Elmira KOCABAŞ

Department of Statistics

MSc. Thesis

Adviser: Doç. Dr. Filiz KARAMAN

Banking business permanency depends on effective risk management.

With this study it is intented to predict the probability of an account to be good or bad in the future and determine the significant variables.

Firstly, multivariate statistical analysis methods and logistic regression, which is one of them and the statistical modelling technique, are considered. After that an application of banking data is interpreted.

Logistic regression model is developed by using SPSS 15.0.



**Keywords:** Logistic regression, scorecard, modelling, binary logistic regression

# BÖLÜM 1

# GİRİŞ

## 1.1 Literatür Özeti

Lojistik regresyon analizi son dönemlerde literatür taramasından da anlaşılabileceği gibi kullanımı hızla yaygınlaşan bir yöntemdir. Bağımlı değişkenin iki kategorili olduğu durumlarda bağımsız değişkenlerle arasında ilişki kurulabilmesi ve neden sonuç ilişkilerinin incelenmesinde önemli bir rol oynamaktadır [1].

Amaçlarından birisi sınıflandırma, diğeri ise bağımlı ve bağımsız değişkenler arasındaki ilişkileri araştırmak olan lojistik regresyon analizinde, bağımlı değişkeni kategorik veri oluşturmakta ve kesikli değerler almaktadır. Bağımsız değişkenlerin ise hepsinin veya bazılarının sürekli ya da kategorik değişkenler olmasına ilişkin bir zorunluluk bulunmamaktadır [2].

Lojistik regresyon analizi, regresyon analizinin normallik, ortak kovaryansa sahip olma gibi bir kısım varsayımlarının sağlanamaması durumunda, diskriminant analizi ve çapraz tablolara alternatif bir yöntemdir. Bağımlı değişkenin 0 ve 1 gibi iki düzey ya da ikiden fazla düzey içeren kesikli bir değişken olması durumunda da uygulanabilir olmasının yanında, matematiksel olarak esnekliği ve kolay yorumlanabilirliği, bu yönteme olan ilgiyi arttırmaktadır [3], [4].

Lojistik regresyon analizi, sınıflama ve atama işlemi yapmaya yardımcı olan bir regresyon yöntemidir. Normal dağılım varsayımı, süreklilik varsayımı önkoşulu yoktur. Bağımlı değişken üzerinde açıklayıcı değişkenlerin etkileri olasılık olarak elde edilerek, risk faktörlerinin olasılık olarak belirlenmesi sağlanır [1], [4].

Lojistik regresyon denkleminde p incelenen olayın gözlenme olasılığını göstermektedir. İncelenen bir olayın olasılığının kendi dışında kalan diğer olayların olasılığına oranına odds değeri denir [5]. İncelenen iki farklı olayın odds değerlerinin birbirine oranına ise

odds oranı denir. Lojistik regresyon denkleminde odds oranı, Exp (β) olarak ifade edilir. Olasılık oranı (Odds), bir olayın meydana gelme olasılığının meydana gelmeme olasılığına oranı olduğuna göre; exp ($\beta_p$) Y değişkeninin $X_p$ etkisi ile kaç kat daha fazla ya da yüzde kaç oranında fazla gözlenme olasılığına sahip olduğunu belirtir.

Bağımlı değişkenin dikotomik olduğunda odds oranı 0,5' ten büyük ise gözlem 1 olarak nitelendirilen gruba, odds oranı 0,5'ten küçükse gözlem 0 olarak nitelendirilen gruba atanmaktadır [6].

Kredi skorlama modelleri, borçlunun gözlemlenebilen özelliklerinden hareketle temerrüt etme ihtimalini hesaplamak ve borçluları farklı risk sınıflarına ayırmak için kullanılır. Kredi skorlama modelleri aracılığıyla üretilen skorlar, kredi fiyatlarının müşteriler bazında farklılaştırılmasına ve iyi kötü müşterilerin ayrıştırılabilmesine imkan sağlamaktadır [7].

## 1.2 Tezin Amacı

Teknolojik gelişmelere paralel olarak insanoğlunun çevresel olayları algılama, yorumlama ve analiz yeteneği devamlı gelişmektedir. Günümüzde verilerin incelenerek anlamlı hale dönüştürülmesi ve yorumlanmasına bir çok iş alanında ihtiyaç duyulmaktadır. Bu alanlardan biri olan bankalarda sorunlu kredilerin önceden tahmini karlılık ve kaynak verimliliği açılarından büyük önem arz etmektedir. Bankacılıkta riskin doğru yönetilmesi ve kaynakların verimli kullanılması potansiyel risklerin önceden tahmin edilerek gerekli aksiyonların bugünden alınmasına bağlıdır. Bu çalışmanın amacı, lojistik regreyon analizi yöntemini kullanarak kredi başvurularının değerlendirilmesinde en yüksek açıklama yüzdesine sahip değişkenleri belirlemek ve müşterinin gelecekteki ödeme performansının iyi mi kötü mü olacağını tahminlemektir.

Bankalar işlemleri nedeniyle karşılaştıkları risklerin izlenmesi ve kontrolü sağlamak amacıyla faaliyetlerini kapsamı ve yapısıyla uyumlu , esas ve usulleri Bankacılık Düzenleme ve Denetleme Kurumu tarafından çıkarılacak yönetmelikle belirlenecek etkin bir iç denetim sistemi ile risk kontrol ve yönetim sistemi kurmakla yükümlüdürler. Bu kapsamda müşteriler, risklerine göre sınıflandırılmalı, sıralanmalı ve gelecekteki temerrüte düşme olasılık tahminleri yapmalıdır.

## 1.3 Hipotez

Kredi başvurusunda bulunan müşterilerin başvurdukları anda riskli olup olmadıklarını tahmin eden modeller kurulmasında genellikle lojistik regresyon kullanılır. Varsayımlarının az olması, sonuçlarının kolay yorumlanabilmesi gibi nedenlerle lojistik regresyon oldukça tercih edilen bir analiz tekniğidir.

Bu tezde çok değişkenli istatistiksel analiz yöntemlerinden biri olan lojistik regresyon hakkında bilgi verildikten sonra lojistik regresyon ile bir model uygulaması yapılmıştır. Model, banka kredisine başvuran müşterilerin geçmişteki ödeme performanslarını baz alarak iyi ya da kötü olma olasılığına göre onay ya da red kararı verecektir.

# BÖLÜM 2

## ÇOK DEĞİŞKENLİ İSTATİSTİKSEL ANALİZ YÖNTEMLERİ

Çok değişkenli istatistik, iki veya daha fazla değişkeni aynı anda analiz etmeye yarayan istatistiksel yöntemleri tanımlamaktadır. Çok değişkenli istatistiksel analizde sistem içerisinde birbiriyle ilişki halinde çok sayıda değişken yer almaktadır. Ayrıca, gerçek hayatta ve modern bilimsel çalışmalarda temel alınan birim ve değişken sayısı birden fazla olup, bu birim ve değişkenlerin de karşılıklı etkileşimleri söz konusudur. Bu nedenle çok değişkenli istatistiksel analiz tekniklerine ihtiyaç duyulur ve bu yaklaşımı tek değişkenli istatistiksel analizlerden üstün kılan temel özellik; tek değişkenli istatistiksel analizlerde veri olarak kabul edilen bir çok faktörün çok değişkenli analizlerde birer değişken olarak sisteme dahil edilebilmesidir [3], [8].

Veriler arasındaki ilişkilerin saptanabilmesi açısından çok değişkenli verilerin gruplandırılması konusunda bir çok teknik geliştirilmiştir. Gözlemleri verilerin yapısında bulunan olası gruplara atamak için birkaç yöntem vardır. Bu yöntemler:

1. Diskriminant analizi
2. Kümeleme Analizi
3. Faktör Analizi
4. Lojistik Regresyon Analizi
5. Çok Boyutlu Ölçekleme Analizidir.

Tüm bu analizlerle gözlemleri gruplara ayırmak amaçlanmaktadır. Kümeleme analizi istatistiksel anlamda birbirinden farklılık gösteren gruplar yaratır. Kümeleme analizi gerçekleştirilirken verilerin grup sayısı bilinmemektedir. Gruplara daha sonradan

katılacakların hangi kıstaslara göre sınıflandırılacağını belirlemez. Lojistik regresyon analizi ve diskriminant analizlerinde ise grup sayısı bilinmektedir. Lojistik regresyon analizi ve diskriminant analizi ile yeni gözlemlerin hangi gruplara atanacağı belirlenmektedir [9], [11].

Diskriminant analizi, gözlemlerin hangi gruplara atanacağını belirlemede kullanılan ve çeşitli varsayımlara ( normal dağılım, ortak kovaryansa sahip olmama gibi) dayanan kullanışlı bir yöntemdir. Kümeleme analizinin bireyleri nasıl kümelediğini öğrenerek her bir grup için bir formül çıkarır. Bu fonksiyonlar aracılığıyla gruplar arası ayırıma en fazla etki eden ayırıcı değişkenleri belirlemektedir. Gruplara atanacak bireyler, bu formüller aracılığıyla sınıflandırılabilmektedir [9] , [11].

Lojistik regresyon analizi ise çeşitli varsayımların sağlanamadığı durumlarda diskriminant analizine alternatif olurken, bağımlı değişkenin 0 ve 1 gibi ikili ya da daha çok düzey içeren kesikli değişken olması durumunda normallik varsayımının bozulması nedeniyle doğrusal regresyon analizine alternatif bir yöntem olarak kullanılmaktadır. Bu analizin temel amacı bağımsız değişkenler ile bağımlı değişkenler arasındaki ilişkiyi ortaya koymaktır [9] , [19].

# BÖLÜM 3

## LOJİSTİK REGRESYON ANALİZİ

İstatistiksel uygulamalarda, bağımsız değişken ile hedef değişkeni arasında bir ilişki kurmak amacıyla bir çok regresyon yöntemi geliştirilmiştir. Doğrusal regresyon analizinde bağımlı değişken nicel değişkenler ile ifade edilmektedir. Ancak bağımlı değişken iki veya ikiden fazla kategorik değer aldığında süreksiz olduğu için normallik varsayımı korunamamaktadır [9]. Bağımsız değişken kategorik olduğunda istenilen sonuçları elde edebilmek için alternatif olarak lojistik dağılımlar kullanılır. Lojistik regresyon, bağımsız değişkenlerin bağımlı değişkenler üzerindeki etkilerini olasılık olarak hesaplar, olasılık kurallarına uygun sınıflama ve atama yapma imkanı verir. Lojistik dağılımları kullanmamızın üç temel sebebi vardır:

1. Basit ve çoklu doğrusal regresyon yöntemlerinin uygulanabileceği veri setlerinde;
    - Bağımlı değişkenin normal dağılım göstermesi,
    - Bağımsız değişkenlerin normal dağılım gösteren populasyonlardan hatasız ölçümler olarak belirlenmesi,
    - Bağımsız değişkenler arasında çoklu bağlantı (multicollinearity) olmaması,

Hata teriminin, tüm bağımsız değişkenler için sıfır ortalamalı ve aynı sabit varyanslı normal dağılım göstermesi $\varepsilon \cong N(0, \sigma^2)$,

- Hata terimleri arasında otokorelasyon olmaması,
- Hata terimleri ile bağımsız değişkenler arasında bir korelasyon olmaması varsayımları gerekmektedir.

Lojistik regresyonda bu ön koşullar aranmaz.

2. Matematiksel olarak incelediğimiz zaman lojistik dağılımlara ait fonksiyonlar çok kolay ve esnek fonksiyonlardır.
3. Lojistik dağılımlar ile elde edilen fonksiyonlar, anlamlı biçimde yorumlanmaya elverişlidir.

Lojistik regresyon analizi, son yıllarda tıp, biyoloji ve ekonomi alanlarında yaygın olarak kullanılmaktadır. Özellikle bankacılık alanında kredi skorlamada çok sıkça kullanılan bir tekniktir. Tıp alanında ise tedavi gören hastaların hastalıklarına göre sınıflandırılmasında elverişli bir teknik olarak karşımıza çıkmaktadır. Son yıllarda tıp alanındaki birçok bilimsel yayında lojistik regresyon analizi kullanılmıştır.

Lojistik regresyon analizi, özellikle bağımlı değişkenin kategorik olduğu durumlarda kullanılmaktadır. Lojistik regresyon analizi bağımlı değişkenin şekline göre 3 grupta incelenmektedir [19]:

1. **İkili (Binary) Lojistik Regresyon Analizi:** Bağımlı değişkenin iki kategorili olduğu durumlarda gerçekleştirilir. Bağımlı değişken 0 ve 1 şeklinde kodlanır ve bu sayılara birer özellik atanır. Örneğin bir ülkede yaşayan vatandaşların genç ve yaşlı şeklinde iki gruba ayrılması gibi durumlarda bağımlı değişken 0 ve 1 şeklinde kodlanmaktadır.

2. **Sıralı (Ordinal) Lojistik Regresyon Analizi:** Bağımlı değişkenin sıralı olduğu durumlarda kullanılır. Bağımlı değişken ikiden fazla kategori içermektedir. Bağımlı değişken sıralayıcı ölçekte kodlandığı zaman kendi sırasına uygun şekilde düzenlenmelidir. Örneğin beyin tümörleri Dünya Sağlık Örgütü (DSÖ) kriterlerine dayanarak en iyiden en kötüye 4 aşamalı olarak derecelendirme sisteminde sınıflandırılır; Evre I, Evre II, Evre III, Evre IV.

3. **Çok Terimli (Multinominal) Lojistik Regresyon Analizi:** Bağımlı değişkenin ikiden fazla düzey içerdiği ve isimsel olarak nitelendirildiği durumlarda kullanılmaktadır. Verilerin kodlanmasında sıralı lojistik regresyon analizinde olduğu gibi belirli bir sıra ile düzenleme zorunluluğu bulunmamaktadır.

## 3.1 Lojistik Regresyon Analizinin Kullanım Alanları ve Tercih Sebepleri

Doğru bir yöntemle iyi bir regresyon modeli kurmak hayati önem taşımaktadır. Çünkü elde edilen modele göre geleceğe yönelik kararlar alınacaktır. Lojistik regresyon analizi, diğer analizlere göre ağır varsayımlar içermediği için sıkça kullanılmaktadır. Özellikle sosyal bilimler, biyoloji, tıp, ekonomi, tarım ve veterinerlik sahalarında sıkça karşılaşmaktayız.

Lojistik regresyon analizi 20. yüzyılın ortalarından itibaren kullanılmaya başlanmıştır. İlk olarak 1944 yılında biyolojik araştırmalarda kullanılabileceği konusunda Berkson tarafından ortaya atılmıştır. İlerleyen yıllarda Cox (1970) lojistik model üzerine bir çok çalışma yapmıştır. Pregibon(1981) ikili lojistik modellerin kullanımı konusunda, Lesaffre ve Albert (1989) ise bağımsız değişken ikiden fazla değer aldığı durumlarda lojistik modele ilişkin etkin ve aykırı gözlemlerle belirleme ölçütleri üzerinde incelemelerini sürdürmüşlerdir [10].

Cornfield (1962) lojistik modele ilişkin katsayı tahminlerinde diskriminant fonksiyonunu önermiştir. Bu çalışma, lojistik modelin kullanımı konusunda bir dönüm noktası olmuştur [11]. Daha sonraki yıllarda Breslow ve Day (1980) epidemioloji, Abbott (1985) yaşam analizi alanında çeşitli uygulamalı araştırmalar gerçekleştirmişlerdir. Gardside ve Glueck (1995) insanlarda beslenme şekli, sigara ve alkol kullanımı, fiziksel aktivite gibi risk faktörlerinin kalp hastalığı üzerindeki etkilerini incelemişlerdir [12].

Türkiye'de de son yıllarda lojistik model ile ilgili çalışmalar sıkça yapılmaktadır. Özellikle bankacılık ve tıp alanında gerçekleştirilen çalışmalarda lojistik model yaygın olarak kullanılmaktadır. Lojistik regresyon analizi, çok değişkenli diğer analizlere nispeten daha kullanışlı olmasının sağladığı avantaj sebebiyle, kategorik veri analizinde önemli bir yere sahip olmakla beraber son dönemlerde kullanımı yaygınlaşan bir yöntemdir. Doğrusal regresyon analizi ve diskriminant analizinde normallik varsayımı aranmaktadır ancak lojistik regresyon analizinde böyle bir koşul yoktur. Lojistik regresyon analizi için varyans kovaryans matrislerinin homojen olması varsayımı aranmamaktadır. Uygulamada bazı durumlarda bu varsayımlar sağlanamadığı için lojistik modelin kullanımı elverişli hale gelmiştir. Kullanım kolaylığı dışında sayısal verilerle rahat yorumlanabilir olması nedeniyle tercih edilen bir yöntem olarak karşımıza çıkmaktadır [4], [13].

# BÖLÜM 4

# LOJİSTİK REGRESYON MODELİNİN TAHMİNİ

## 4.1 Lojistik Modelin Doğrusal Modelle İlişkisi

Lojistik regresyon ile doğrusal regresyon analizi arasında bulunan varsayımlar, fonksiyonlar gibi bazı farklılıkların açıklanabilmesi için doğrusal regresyon modelini incelerken kullandığımız temel ilkeler, lojistik regresyon analizinde de kullanılacaktır.

Öncelikle doğrusal regresyon modeli ile bir başlangıç yapmak gerekirse bağımlı değişken $Y=E(Y/X) + u$ denklemi ile gösterilebilir. u, bu denklemde hata terimi olarak adlandırılır ve gözlemlerin koşullu ortalamadan sapmalarını gösterir.

Doğrusal regresyon modeline göre hata terimi 0 ortalama ve sabit varyans ile normal dağılıma uymaktadır. $u \sim N(0; \sigma^2)$. Ancak bağımlı değişkenin kesikli olduğu durumlarda hata terimine ilişkin bu varsayımlar sağlanamamaktadır.

Bağımlı değişken Y=0 ve Y=1 şeklinde dikotomik olarak gösterilsin. Bu durumda bağımsız değişken için gerçekleşme olasılığı $\Pi(x)$ ve gerçekleşmeme olasılığı $1 - \Pi(x)$ şeklinde ifade edilir. Koşullu ortalamayı lojistik regresyon modelini göz önüne alındığında $\Pi(x)=E(Y/X)$ olarak ifade edilir.

Bağımsız değişkenin aldığı değerlere göre hata terimleri iki farklı şekilde gösterilebilir:

Y=1 için $u=1-\pi(x)$ ( $\pi(x)$ olasılık ile) (1)

Y=0 için $u= -\pi(x)$ ( $1-\pi(x)$ olasılık ile) (2)

Y bağımlı değişkeni dikotomik olarak gösterildiğinde Y iki farklı değer alacağı için Bernoulli sürecine sahiptir. Bernoulli sürecinin varyansı olayın gerçekleşme ve gerçekleşmeme durumlarının çarpımı şeklinde edilmektedir. Yukarıdaki denklemlere göre hata terimlerinin varyansı şu şekilde elde edilir:

$$Var(u)= \pi(x).(1- \pi(x)) \quad (3)$$

Hata terimlerinin varyansı Bernoulli sürecine sahiptir ve Binom dağılımına uymaktadır. Bu nedenle lojistik regresyon modeline ait hata terimleri normal dağılıma uymadığı için normallik varsayımı ihlal edilmektedir. Bağımlı değişkenin kategorik olduğu durumda normallik varsayımı ihlal edilir ve hata terimleri de normal dağılıma uymaz. Bu sebeple lojistik regresyon analizi ile doğrusal regresyon analizi arasındaki en temel farklılık hata terimlerinin dağılımı noktasında oluşmaktadır. Ayrıca hata terimlerinin varyansı gözlemden gözleme değiştiği için hata terimlerinin varyansı sabit değildir. Hata terimleri varyansı sabit olmadığı için heteroskedastisite (değişen varyans) sorunu ortaya çıkmaktadır. Lojistik regresyon modeli ile doğrusal regresyon modeli arasında hata terimlerinin varyansları açısından farklılık bulunmaktadır [4], [14].

Genel olarak lojistik model ile doğrusal model arasında dört temel farklılığın olduğunu söyleyebiliriz:

1. Lojistik modelin bağımlı değişkeni kategorik biçimdeki nitel gözlemlerden, doğrusal modelin bağımlı değişkeni nicel gözlemlerden oluşmaktadır.
2. Doğrusal modelde hata terimleri normal dağılıma uygunken, lojistik modelde hata terimleri normal dağılıma uygun değildir.
3. Lojistik modelin hata terimleri arasında değişen varyans problemi vardır. Doğrusal modelin hata terimleri arasında değişen varyans problemi bulunmamaktadır.
4. Doğrusal model tahmin edilirken bağımlı değişkene ait bir değer tahmini yapılırken, lojistik modelde bağımlı değişkenin gerçekleşme olasılığı tahmin edilmektedir [15], [16], [25].

Doğrusal modeli kullanabilmemiz için varsayımların yerine getirilmesi zorunlu bir koşuldur. Ancak lojistik modelde varsayımların gerçekleşmesi koşulu aranmamaktadır. Doğrusal regresyon analizini kullanabilmek için çeşitli varsayımların geçerli olmadığı durumlarda çeşitli metotlar uygulanarak doğrusal regresyon analizi kullanılabilir.

Örneğin hata terimleri normal dağılıma uymadığı durumlarda örnek sayısı artırılır ve hataların asimptotik olarak normal dağılıma uyması sağlanabilir. Değişen varyans sorunu ortaya çıktığında çeşitli değişken dönüşümleri ile bu problem ortadan kaldırılabilir. Ancak bağımlı değişken kategorik olarak nitelendirildiği zaman, bağımlı değişken her değeri alamayacağı için doğrusal regresyon analizi kullanılamamaktadır.

Sonuç olarak varsayımların ihlal edildiği ve bağımlı değişkenin kategorik olduğu durumlarda lojistik regresyon analizi doğrusal regresyon analizine göre daha elverişli hale gelmektedir [3], [4].

### 4.2 Tek Değişkenli Lojistik Regresyon Modelinin Kurulması

Doğrusal regresyon modeli kurulurken, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_1 X_1 + \ldots\ldots\ldots\ldots + \beta_n X_n +$ u şeklinde doğrusal bir denklemden yararlanılmaktadır. Lojistik model kurulurken ise bağımlı değişkenin kategorik olması sebebiyle doğrusal bir denklem kullanılamamaktadır. Doğrusal bir denklem yerine lojistik denklem kurulmaktadır [9].

Varsayalım ki elimizde n tane bağımsız ($X_i$, $Y_i$) çifti şeklinde gözlem bulunsun (i=1,2,…….n). Y bağımlı değişkeni 0 ve 1 şeklinde dikotomik olduğunda doğrusal fonksiyon modeli kullanılamaz. Lojistik model kurulurken kullanılan fonksiyon

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)} \qquad (4)$$

şeklindedir.

Tanımlamış olduğumuz lojistik fonksiyonu bağımlı değişken dikotomik olduğu için 0 ile 1 arasında değer almaktadır. Bu durumda model 0 ile 1 arasında değer alabilen bir olasılık üzerine kurulmuştur. Kurulan lojistik model, 0 ile 1 arasında bir değer alabilen karşılaşabileceğimiz veya etkilenebileceğimiz riski tahmin etmek için kullanılır. Bağımlı değişkenin dikotomik olmadığı durumlarda lojistik fonksiyonun haricinde Gompertz, Normal ve Burr gibi çeşitli eğriler kullanılabilmektedir [3], [6], [9].

### 4.3 Lojit Dönüşümü ve Lojit Modelin Elde Edilmesi

Doğrusal olasılık modeli kullanılırken bağımsız değişkenin verilen bir değeri için bağımlı değişkenin ortalama değeri olarak şu denklem kullanılmaktadır:

$$E(Y/X)= \beta_0 + \beta_1 X \quad (5)$$

Bu doğrusal modelin sol tarafındaki Y değerleri dikotomik olduğu için 0 ile 1 arasında bir değer alması gerekmektedir. Bu modeldeli temel sorun, tanım aralığının 0 ile 1 arasında bir değer almasıdır. Dikotomik bağımlı değişken sınırlı olasılık değerleri alabilirken, sonsuz sayıda değer alabilecek bağımsız değişkenlerle ilişkilendirilmektedir. Bağımsız değişken sonsuz sayıda değer alabileceği için bu eşitlik her zaman sağlanamaz. Böyle bir durumda 0 ile 1 arasında değer alabilen olasılık değerleri (-∞, +∞) arasında tanımlı hale getirilerek ve bu sorun ortadan kaldırılmaktadır. Bu sorunu çözebilmek için lojit dönüşümü dediğimiz logaritmik dönüşüm kullanılmaktadır. Bu dönüşüm odds olarak adlandırılan bir kavram üzerinde uygulanmaktadır [7], [17].

Odds kavramı, lojistik regresyon analizinde kullanılan en temel kavramlardan biridir. Bu oran, bir olayın gerçekleşme sayısının gerçekleşmeme sayısına oranı olarak tanımlanır. Lojistik fonksiyonun gerçekleşme olasılığı olarak tanımladığımız Π(x), lojistik dağılım kullanıldığında verilen X değeri için Y' nin ortalaması olarak tanımlanmaktadır ve Π(x)=E(Y/X) olarak gösterilmektedir. Bu denklem lojistik regresyon denklemi olarak da adlandırılmaktadır. Bu lojistik denklem kullanılarak odds, şu şekilde gösterilmektedir [7], [17]:

$$\frac{\pi(X)}{1-\pi(X)} \quad (6)$$

Elde edilen odds üzerinde logaritmik dönüşüm kullanılarak logit fonksiyonu elde edilmektedir. Bu dönüşüme logit dönüşümü denir. Elde edilen logit fonksiyonu şu şekilde gösterilmektedir:

$$g(X) = \frac{In[\prod(x)]}{In[1-\prod(x)]} = \beta_0 + \beta_1 X \quad (7)$$

Bu dönüşüm sonucunda elde logit fonksiyonu istenilen birçok özelliğe sahiptir. Bu fonksiyon parametrelerinde doğrusaldır ve X' in alabileceği değerlere bağlı olarak (-∞,+∞) aralığında değerler alabilmektedir. Burada dikkat edilmesi gereken nokta hata terimlerinin bulunmayışıdır. Modele baktığımız zaman koşullu ortalama E(Y/X) yerine

dönüştürülmüş g(x) fonksiyonu kullanılmıştır. Bu sebeple bağımlı değişken Y modelin sol tarafında bulunmadığı için logit modelin içerisinde hata terimi bulunmamaktadır.

### 4.4 Odds Oranının Regresyon Katsayısı ile İlişkisi ve Yorumlanması

Odds oranı, lojistik fonksiyona ilişkin odds değerleri kullanılarak elde edilir ve şu şekilde tanımanmaktadır:

$$\text{Odds oranı} = \frac{e^{g(x)}}{1+e^{g(x)}} \tag{8}$$

Denkem (8)' de gösterilen odds oranı, regresyon katsayısı ile ilişki içerisindedir. Bu ilişki elde edilen lojit fonksiyonu kullanılarak tanımlanabilir. Varsayalım ki elimizdeki bağımsız X değişkeni 0 ve 1 gibi iki değer alsın. Lojit fonksiyonunda X yerine 0 ile 1 değerlerini yazdığımızda g(1) ile g(0) arasındaki fark şu şekildedir:

$$g(1)-g(0)=[\beta_0+\beta_1]-[\beta_0]=\beta_1 \tag{9}$$

Lojit fonksiyonunun X değerleri arasındaki farka bakılarak, bu farkın bağımsız X değişkenin eğim katsayısı olduğu görülmektedir.Bu lojit farkının βı parametresine eşit olması sonucunu yorumlayabilmemiz için odds oranı ölçüsünü belirtmemiz gerekir. Bağımsız X değişkeni 0 ve 1 değerlerini aldığında olasılık oranları şu şekilde gösterilmektedir:

$$\text{X=1 için} \qquad \frac{\pi(1)}{1-\pi(1)} \tag{10}$$

$$\text{X=0 için} \qquad \frac{\pi(0)}{1-\pi(0)} \tag{11}$$

X=0 ve X=1 değerleri için olasılıkların oranını şu şekilde tanımlayabiliriz:

$$\frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]} \tag{12}$$

π(x) fonksiyonu yerine 0 ve 1 değerlerini koyduğumuz zaman denklem (12) kullanılarak olasılıkların oranı olarak tanımladığımız odds oranı elde edilir:

$$O.O = e^{\beta_1} \tag{13}$$

Lojistik regresyon analizinde bağımsız değişken 0 ve 1 değerlerini aldığında odds oranı ile regresyon katsayısı arasındaki ilişki denklem (13)' teki şekilde tanımlanmaktadır. Odds oranı ile regresyon katsayısı arasındaki bu ilişki lojistik regresyon analizinin güçlü bir analiz tekniği olduğunu ortaya koymaktadır. Odds oranlarının yorumlanması oldukça kolaydır. Bağımlı değişkenin dikotomik olduğunda odds oranı 0,5' ten büyük ise gözlem 1 olarak nitelendirilen gruba, odds oranı 0,5'ten küçükse gözlem 0 olarak nitelendirilen gruba atanmaktadır. Bağımsız X değişkenin ve β parametrelerinin değeri ne olursa olsun, bağımlı fonksiyonun olasılık değeri 0 ile 1 arasında bir değer alacaktır [6].

# BÖLÜM 5

# LOJİSTİK MODELİN TAHMİNİ VE KARŞILAŞTIRILMASI

## 5.1 Lojistik Modelin Parametrelerinin Tahmini

Doğrusal regresyon analizinde hata terimleri normal dağıldığı ve dolayısıyla bağımlı değişkenin normal dağılıma uyduğu durumlarda en küçük kareler yöntemi kullanılarak regresyon katsayıları tahmin edilmektedir. En küçük kareler yöntemi verilen örneklem için hata kareleri toplamını minimum yapacak şekilde regresyon katsayılarını tahmin etmektedir. Tek bağımsız değişkene sahip bir doğrusal regresyon modeli için, hata kareleri toplamının birinci türevi alınarak normal denklemler elde edilmektedir. Normal denklemler şu şekilde gösterilmektedir [6]:

$$\sum Y_i = n\beta_0 + \beta_1 \sum X_i \tag{14}$$

$$\sum Y_i X_i = \beta_0 \sum X_i + \beta_1 \sum X_i^2 \tag{15}$$

Normal denklemler kullanılarak doğrusal regresyon modeli için regresyon katsayıları tahmin edilmektedir. Bu denklemleri kullanarak doğrusal modele ilişkin sabit terim ve eğim katsayısı şu şekilde hesaplanmaktadır:

$$\beta_0 = \frac{\sum X_i{}^2 \sum Y_i - \sum X_i \sum X_i Y_i}{n\sum X_i{}^2 - (\sum X_i)^2} \quad (16)$$

$$\beta_1 = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i{}^2 - (\sum X_i)^2} \quad (17)$$

Lojistik modele ilişkin parametre tahmini doğrusal modelin tahmininden farklıdır. Lojistik modelde bağımlı Y değişkeni kategorik değerler aldığı için hata terimleri normal dağılıma uymamaktadır. Bu varsayım sağlanamadığı için lojistik modele ilişkin regresyon katsayılarının tahmininde en çok olabilirlik yöntemi kullanılmaktadır. Bu yöntem, gözlenen veri kümesine ulaşılma olasılığını maksimize eden parametrelerin tahminini sağlamaktadır. En çok olabilirlik yöntemini kullanabilmek için en çok olabilirlik fonksiyonu olarak nitelendirilen bir fonksiyondan yararlanılmaktadır. Bu fonksiyon, bilinmeyen parametreler için gözlenen verilerin olasılıklarını göstermektedir.

İki farklı değer alabilen bağımlı Y değişkeni dikotomik olduğunda 0 ve 1 gibi iki farklı değer alabilir. Buna göre bağımlı 1 değerini aldığında koşullu olasılık değeri P(Y=1 /X) ve bağımlı değişken 0 değerini aldığında koşullu olasılık değeri P(Y=0/X) şeklinde gösterilmektedir. π(x) Y=1 için koşullu olasılık, 1- π(x) Y=0 için koşullu olasılık değeri olarak tanımlanabilir. Koşullu olasılık değerleri iki farklı değer alabildiği için Bernoulli dağılımına uymaktadır. Bağımsız $X_i$ ve $Y_i$ çiftleri için sahip olduğu Bernoulli dağılımının fonksiyonundan yararlanılarak olabilirlik fonksiyonu şu şekilde gösterilir:

$$\pi(X_i)^{y_i}.[1-\pi(X_i)]^{1-y_i} \quad (18)$$

Tüm gözlemler bağımsız olarak kabul edildiği için çarpım şeklinde gösterilebilir. En uygun olabilirlik fonksiyonu çarpım olarak şu şekilde tanımlanır:

$$L(Y/X) = \prod_{i=1}^{n} \pi(X_i)^{y_i}.[1-\pi(X_i)]^{1-y_i} \quad (19)$$

Olabilirlik fonksiyonu, $\pi(X_i)$ yerine açık ifadesi konularak şu şekilde elde edilir:

$$L(Y/X)=\prod_{i=1}^{n}\left[\frac{\exp(\sum_{k=0}^{p}\beta_k X_k)}{1+\exp(\sum_{k=0}^{p}\beta_k X_k))}\right]^{yi}.\left[\frac{1}{1+\exp(\sum_{k=0}^{p}\beta_k X_k))}\right]^{1-y_i} \quad (20)$$

İşlemlerde kolaylık sağlaması açısından olabilirlik fonksiyonun logaritması alınır ve lojistik modelin olabilirlik fonksiyonu elde edilir:

$$In\left[L(Y/X)\right]=\sum_{i=1}^{n}Y_i.\log\pi_i+(1-Y_i).\log(1-\pi_i) \quad (21)$$

En çok olabilirlik yönteminde, p tane bağımsız değişkene ait β parametrelerinin tahminleri, L(Y/X) fonksiyonunu maksimum yapacak şekilde seçilir. Olabilirlik fonksiyonunu maksimum yapan bu değer aynı zamanda logaritması alınmış olan fonksiyonu da maksimum yapmaktadır. Lojistik model kullanılarak β değerlerine göre türev alınarak olabilirlik eşitlikleri elde edilmektedir. Olabilirlik eşitliklerinden yararlanılarak β parametrelerinin tahminleri elde edilir [6].

Doğrusal regresyon modelinde hata kareleri toplamının fonksiyonunda bulunan β parametreleri için türev alınarak parametre tahminleri elde edilmektedir. Bu şekilde hesaplanan parametre değerleri kolaylıkla çözümlenebilmektedir. Ancak lojistik modelin parametreleri doğrusal olmadığı için özel çözümler gereklidir ve bunun için iteratif yöntemler kullanılmaktadır [17].

İteratif tahmin yöntemi kullanıldığında öncelikle olabilirlik eşitliklerinde yer alan doğrusal olmayan β değerine bir başlangıç değeri verilmektedir. Bu başlangıç değerlerinden itibaren olabilirlik eşitliklerinin türevleri alınarak en çok olabilirlik yöntemine göre tahminler iteratif bir şekilde elde edilmektedir. Bu iterasyon parametre tahminleri arasındaki fark belirlenen δ değeri kadar olana kadar devam eder. Parametreler arasındaki bu fark yeteri kadar azaldığında yakınsama sağlanır ve iterasyon işlemi durdurulur. Buradaki temel amaç, en az sayıda iterasyon ile parametreleri tahmin edebilmektir. Bu nedenle başlangıç değerleri uygun olarak seçilmelidir [6], [17].

Olabilirlik eşitlikleri kullanılarak seçilen tahminciler maksimum olabilirlik tahmincileridir ve $\hat{\beta}$ şeklinde gösterilir. Y=1 iken koşullu olasılık tahmini $\pi(X_i)$ olarak gösterilmektedir. Olabilirlik eşitliklerinden yararlanılarak gözlenen y değerlerinin toplamı tahmin edilen y değerlerinin toplamına eşittir ve şu şekilde gösterilir:

$$\sum Y_i = \pi(X_i) \tag{22}$$

En çok olabilirlik yönteminin dışında sıkça kullanılan bir diğer yöntem yeniden ağırlıklandırılmış iteratif en küçük kareler yöntemidir. Yeniden ağırlıklandırılmış iteratif en küçük kareler yöntemine göre, tahmin edilen Y değerlerinin gözlenen Y değerlerinden sapmalarının karesini minimum yapacak şekilde β parametreleri tahmin edilir. En küçük kareler yöntemi bağımlı değişken kesikli olduğu durumlarda varsayımlar sağlanamadığı için kullanılamaz. En küçük kareler yöntemine alternatif olarak bu yöntem kullanılabilir [15], [18].

Matris notasyonu ile gösterecek olursak, X bağımsız değişkene ait gözlem değerlerini, Y bağımlı değişkene ait gözlem değerlerini, W ise π(x).(1- π(xi)) elementlerinden oluşan diagonal matris olarak gösterilebilir. Z ise, yeniden ağırlıklandırılmış iteratif en küçük kareler yönteminde Y'nin yerine kullanılmaktadır ve şu şekilde gösterilir:

$$Z_i = \log\left[\hat{\pi}(x_i)/(1-\hat{\pi}(x_i))\right] + \frac{Yi - \hat{\pi}(x_i)}{\hat{\pi}(x_i).(1-\hat{\pi}(x_i))} \tag{23}$$

Matris notasyonu ile gösterdiğimiz iteratif β tahminleri için kullanılan denklem şu şekilde tanımlanır:

$$\hat{\beta} = \left(X'WX\right)^{-1} X'WZ \tag{24}$$

Z matrisi,matris notasyonu ile β tahminleri kullanarak şu şekilde gösterilir:

$$Z = X \overset{\wedge}{\beta}^{(i)} + W^{-1}e \tag{25}$$

Bu denklem kullanılarak iterasyon yapabilmek için hata terimlerini de denkleme eklemek gerekir. Hata terimini e=Yi-$\pi(X_i)$ şeklinde gösterecek olursak, iterasyon için kullanılacak denklem elde edilmektedir:

$$\beta^{(i+1)} = \beta^{i} + (X'WX)^{-1}X'e \tag{26}$$

Yukarıda tanımlanan denklem tahminler yakınsayana kadar iterasyona tabi tutulur. Yakınsama için genel olarak tahminler arasında 0,000000001 kadar farkın olması yeterli kabul edilmektedir [15], [19].

### 5.2 Katsayıların Anlamlılıklarının Test Edilmesi

Lojistik model kullanılarak elde edilen katsayı tahminlerinin ardından katsayıların anlamlı olup olmadığına bakılmalıdır. Katsayıların anlamlılığını test edebilmek için istatistiksel hipotezler kullanılır ve bağımsız değişkenin bağımlı değişkeni açıklamada ne derecede etkin olduğu incelenir. Yapılacak hipotez testleri modelden modele farklılık gösterir. Burada tek değişkenli model için kullanılacak yöntem açıklanacaktır.

Lojistik regresyon analizinde katsayıların anlamlılığının test edilmesi doğrusal regresyon analizi ile benzerdir. Doğrusal regresyon analizinde AKT (Açıklanan Kareler Toplamı) kullanılarak modelin karşılaştırması yapılmaktadır. AKT' nin yüksek olması bağımsız değişkenin önemli olmasını sağlar. Lojistik regresyon analizinde de doğrusal regresyon analizinde olduğu gibi, modelde yer alan bağımsız değişkenin modelde olduğu ve olmadığı durumlara göre karşılaştırma yapılmaktadır. İçinde bağımsız değişkenin bulunduğu modeli uygun model olarak tanımlayabiliriz. Uygun model bağımsız değişkenin olmadığı modele göre daha güvenilir ise, bağımsız değişkenin anlamlı olduğunu söylemek mümkümdür [20].

Lojistik regresyon analizinde bağımsız değişkenin modelde olduğu ve olmadığı durumlarda karşılaştırma yapabilmek için en çok olabilirlik fonksiyonu kullanılmaktadır. Bu fonksiyondan yararlanılarak doymuş model olarak adlandırabileceğimiz modelden yararlanılır. Doymuş model, içinde verileri taşıyan

parametreleri içerir. Gözlem değerleri ile tahmin edilen değerler arasında bir karşılaştırma yapabilmek için, logaritması alınmış en çok olabilirlik fonksiyonu kullanılarak D istatistiği kullanılır. D istatistiği şu şekilde tanımlanmaktadır:

$$D = -2In\left[\frac{\text{Tahmin edilen modelin olabilirliği}}{\text{Doymuş modelin olabilirliği}}\right] \tag{27}$$

Parantez içerisinde kalan kısma olabilirlik oranı denir. D istatistiğinin başında eksi bulunmasının nedeni elde edilen niceliğin dağılımının bilinmesi ve hipotez testi yapabilmek içindir. Bu teste olabilirlik oranı testi denir. D istatistiği ayrıca şu şekilde de gösterilebilir:

$$D = -2In\sum y_i.In\left(\frac{\pi_i}{y_i}\right) + (1-y_i).In\left(\frac{1-\pi_i)}{1-y_i}\right) \tag{28}$$

Bu eşitlik için kullanılan D istatistiği bazı yazarlar tarafından [McCullagh, Nelder(1983)] sapma olarak da nitelendirilir. Bu sapma uyum iyiliği testlerinde çok önemli rol oynamaktadır. D istatistiği doğrusal regresyon analizinde kullanılan AKT ile aynı rolü oynamaktadır. Doğrusal regresyon analizinde kullanılan AKT, lojistik regresyon analizinde D ile gösterilir.

Bağımlı Y değişkeni dikotomik olduğunda doymuş modelin olabilirliği 1'dir. Doymuş modelin olabilirliği $\hat{\pi}_i = y_i$ olduğu için, doymuş modelin olabilirliği şu şekilde gösterilebilir:

$$L(\text{Doymuş Model}) = \prod_{i=1}^{n} y_i^{\,y_i}(1-y_i)^{1-y_i} = 1 \tag{29}$$

Bağımsız değişkenin modele dahil olduğu ve dahil olmadığı durumlara göre D istatistiklerinin karşılaştırılması için G istatistiği kullanılmaktadır. G istatistiği şu şekilde ifade edilmektedir:

G=D(Modelde bağımsız değişkenin yok)- D(Modelde bağımsız değişken var) (30)

G istatistiği doğrusal regresyon analizinde kullanılan F testi ile aynı rolü üstlenmektedir.

G istatistiği, Ho:βı=0 hipotezi altında, yani tek değişkenli lojistik model için 1 serbestlik derecesi ile ki-kare dağılımına uygunluk göstermektedir. Çeşitli ek varsayımlar gerekli olsa dahi, örnek hacmi yeteri kadar büyük olduğunda varsayımların sağlanma zorunluluğu bulunmamaktadır [3], [6], [9].

### 5.3 Güven Aralıklarının Tahmin Edilmesi

Regresyon katsayılarının anlamlılıklarının test edilmesinin ardından katsayıların güven aralıklarının tespit edilir. Lojistik regresyon analizinde de doğrusal regresyon analizinde olduğu gibi sabit terim ve eğim katsayısı için ayrı şekilde güven aralıkları tahmin edilir. Doğrusal regresyon analizinde β parametreleri için güven aralıkları şu şekilde gösterilir [6]:

$$\beta_o \pm z_{1-\alpha/2.} \hat{S.E}(\hat{\beta}_o) \tag{31}$$

$$\beta_l \pm z_{1-\alpha/2.} \hat{S.E}(\hat{\beta}_l) \tag{32}$$

Tahmin edilen β parametreleri hesaplanan güven aralıkları içinde kalırsa parametrelerin anlamlı olduğu söylenebilir. Lojistik modelde ise elde edilen lojit denklemi modelin doğrusal kısmıdır. Bu kısım doğrusal modelde olduğu gibi değerlendirilebilir. Lojit denklemi şu şekildedir:

$$\hat{g(x)} = \hat{\beta}_o + \hat{\beta}_l \mathrm{X} \tag{33}$$

$\hat{g(x)}$ fonksiyonu, lojit denklemi olarak adlandırdığımız g(x) fonksiyonunun tahmini olarak nitelendirilmektedir. $\hat{g(x)}$ fonksiyonun varyansını bulabilmek için denklemin her iki tarafının da varyansını almamız gerekir. Bu şekilde elde edilen $\hat{g(x)}$ fonksiyonunun varyansı şu şekilde gösterilir:

$$\hat{Var}\left[\hat{g(x)}\right] = \hat{Var}(\hat{\beta}_o) + x^2 \hat{Var}(\hat{\beta}_l) + 2x \hat{\mathrm{cov}}(\hat{\beta}_o, \hat{\beta}_l) \tag{34}$$

Lojit fonksiyonunun güven aralığının yazılabilmesi için elde edilen varyansın karekökü alınarak standart hatalar hesaplanır. 100(1-α/2) güven ile lojit fonksiyonu için güven aralığı doğrusal regresyon analizine benzer şekilde, şu şekilde gösterilebilir [6], [8]:

$$\hat{g}(x) \pm z_{1-\alpha/2.} \hat{S.E}(\hat{g}(x)) \tag{35}$$

### 5.4 Uyum İyiliği Ölçütleri

Lojistik modele ilişkin tahminler yapıldıktan ve modelleme aşaması tamamlandıktan sonra modelin ne derecede başarılı olduğu ölçülebilir. Modelde yer alan bağımsız değişkenlerin bağımlı değişkeni açıklamada ne derece başarılı olduğu konusunda değerlendirme yapabilmemiz uyum iyiliği ölçütleri kullanılır [20].

Lojistik modelin oluşturulması sırasında çeşitli hatalar gözlenebilmektedir. Bu hatalar şu şekilde sıralanabilir:

- Modelin kurulması sırasında logaritmik dönüşüm uygulanmıştır ancak bu dönüşüm model için uygun değildir.
- Modelin kurulum aşamasında gerekli değişkenler modele dahil edilmemiş olabilir veya kullanımı gereksiz değişkenler modele dahil edilmiştir. Bu durumda model hatalı kurulmuştur.
- Modelde bağımlı Y değişkeni yanlış ölçek biçiminde kullanılmış olabilir. Ölçeğin sıralı, çok seviyeli ve dikotomik olmasına göre farklı ölçekler kullanılabilir. Ölçeğin kullanımına göre kurulacak model de değişkenlik göstereceği için, tahminler yanlış sonuç verebilir.

Lojistik regresyon analizinde kullanılan başlıca uyum iyiliği ölçütleri şunlardır:

- Pearson Ki-Kare İstatistiği
- Sapma (Deviance) İstatistiği
- C İstatistiği
- Belirlilik Katsayısı ( $R^2$)
- Sahte (Pseudo) Belirlilik Katsayısı

- Hosmer-Lemeshow İstatistiği

### 5.5 Pearson Ki-Kare İstatistiği

Doğrusal regresyon analizinde uyum iyiliği testleri hata terimlerine bakılarak yapılmaktadır. Lojistik regresyon analizinde de aynı yaklaşım geçerlidir. Lojistik modelde tahminciler değişkenin gözlenme olasılıkları kullanılarak şu şekilde hesaplanır [6]:

$$\hat{y}_i = m_j \cdot \frac{e^{\hat{g}(x_j)}}{1+e^{\hat{g}(x_j)}} \tag{36}$$

Burada kullanılan $\hat{g}(xj)$, g(xj) fonksiyonunun tahmini olarak gösterilmektedir. Lojistik modelin hata terimleri şu şekilde tanımlanır:

$$r(y_i, \hat{\pi}_j) = \sqrt{\frac{(y_i - m_j \hat{\pi}_j)}{m_j \hat{\pi}_j (1-\hat{\pi}_j)}} \tag{37}$$

Buradan elde elde edilen hata terimlerinde yola çıkarak Pearson ki-kare istatistiği şu şekilde hesaplanır:

$$\chi^2 = \sum_{j=1}^{J} r(y_i, \hat{\pi}_j)^2 \tag{38}$$

Denklem (38)' den elde edilen formüle göre ki-kare istatistiği J-(p+1) serbestlik derecesi ile ki-kare dağılımına uymaktadır [1].

### 5.6 Sapma İstatistiği

Sapma istatistiğini hesaplayabilmemiz için, sapma kalıntılarını kullanmamız gerekir. Sapma kalıntıları şu şekilde gösterilmektedir:

$$d(y_i,\hat{\pi}_i)=\pm\left\{2.\left[y_j In\left(\frac{y_j}{m_j\hat{\pi}_j}\right)+(m_j-y_j)In\left(\frac{m_j-y_j}{m_j(1-\hat{\pi}_j)}\right)\right]\right\}^{1/2} \quad (39)$$

Sapma kalıntısının başındaki işaretin + veya – olmasına göre ( $y_j - m_j \pi_j$ )' nin işareti de aynı olmaktadır.

Denklem (39)'daki yj değeri 0 olduğu durumda sapma kalıntısı şu şekilde yazılabilir:

$$d\left(y_j,\hat{\pi}_j\right)=-\sqrt{2m_j\left|In\left(1-\hat{\pi}_j\right)\right|} \quad (40)$$

Denklem (39)'daki yj değeri 1 olduğu durumda sapma kalıntısı şu şekilde yazılabilir:

$$d\left(y_j,\hat{\pi}_j\right)=-\sqrt{2m_j\left|In\left(\hat{\pi}_j\right)\right|} \quad (41)$$

Sapma istatistiği olarak tanımlayabileceğimiz D istatistiği, sapma kalıntıları ile hesaplandığında şu şekilde gösterilmektedir:

$$D=\sum_{j=1}^{J} d\left(y_i,\hat{\pi}_j\right)^2 \quad (42)$$

Sapma istatistiği J-(p+1) serbestlik derecesi ile ki kare dağılımına uygunluk göstermektedir [3].

### 5.7 C İstatistiği

C istatistiği, doğrusal regresyon analizinde bağımsız değişkenlere ait β katsayılarının anlamlılıklarını ölçmek için kullandığımız F testi ile aynı işlevi görmektedir. C istatistiği, sabit terim dışındaki tüm katsayıların sıfırdan farklı olup olmadığını test etmek için kullanılmaktadır. Bu hipotez şu şekilde kurulabilir:

$Ho: \beta_1=\beta_2=\beta_3=.....=\beta_k=0$

$Hı:$ Katsayıların en az biri sıfırdan farklıdır.

C istatistiği, esas olarak benzerlik oranından yola çıkılarak şu şekilde gösterilmektedir:

$$C = -2\log\left(\frac{L_o}{L_\iota}\right) \tag{43}$$

Lo, sabit terim dışındaki tüm katsayıların 0 olması durumunda hesaplanan olabilirlik değeridir. Lı ise, katsayıların anlamlı olduğu durumdaki tüm modelin olabilirlik değeridir.

Denklem (43)' te hesaplanan C istatistiği (k-1) serbestlik derecesi ile ki-kare dağılımına uygunluk göstermektedir. Serbestlik derecesinin (k-1) olmasının sebebi, sabit terimin hipoteze dahil olmamasıdır [3].

### 5.8 Belirlilik Katsayısı

Doğrusal regresyon analizinde sıkça kullanılan belirlilik katsayısı yani $R^2$ ölçütü, lojistik regresyon analizinde kullanılamaz. Doğrusal regresyon modelinde belirlilik katsayısı bağımsız değişkenler tarafından açıklanan varyans oranını verir. Ancak bağımlı değişkenin kategorik değerler aldığı modellerde hata varyansının minimize edilmesi gibi bir durum söz konusu değildir [3].

Bağımlı değişkenin kategorik olduğu durumlarda ortalama ve varyans gibi parametreler birbirinden farklı değildir. Örneğin Poisson dağılımında ortalama ve varyans birbirine eşittir. Aynı şekilde Bernoulli dağılımına göre varyans, bağımlı değişken dikotomik iken beklenen değerin 0 veya 1' e yakın olduğu durumlarda minimum değerini alır. Beklenen değerin alacağı değere göre varyansı minimize etmek rasyonel bir yaklaşım değildir. Ayrıca lojistik modele göre tahmin edilemeyen her bir değer için belirlilik katsayısı büyük oranda düşüş gösterebilmektedir. Bu durumda mükemmele yakın tahmin edilmiş bir modelin belirlilik katsayısı 0,9' dan küçük çıkabilir. Bu durum, doğrusal modelde kullanılan belirlilik katsayısının lojistik model için uygun olmadığını göstermektedir.

Cox ve Wermuth (1992) bağımlı değişkenin 0 ve 1 değerini aldığında, iyi tahmin edilmiş modellerin belirlilik katsayısının sıkça 0,1 gibi düşük bir değer aldığı sonucuna varmıştır. Doğrusal modelde kullandığımız belirlilik katsayısı yerine alternatif belirlilik katsayısı ölçütleri geliştirilmiştir. Örneğin Madalla (1983) ve Magee (1990) şu ölçütü geliştirmiştir [10], [11], [23]:

$$R^2 = 1 - \left\{ L(0) / L(\hat{\beta}) \right\}^{2/n} \quad (44)$$

L(0), modelin içinde regresyon katsayıları olmadığı durumda modelin olabilirliğini göstermektedir. $L(\hat{\beta})$ ise modelin içinde regresyon katsayılarının bulunduğu durumdaki modelin olabilirliğini göstermektedir. Cox ve Snell (1989) geliştirdikleri belirlilik katsayısında 2/n yerine 1/n değerini kullanmışlardır. Denklem (44)'te gösterilen ölçüt, doğrusal modeldeki belirlilik katsayısı ile aynı işlevi görmektedir.

Doğrusal modelde belirlilik katsayısını maksimum yapan değer bağımlı değişken Y tahmincilerinin Y' nin ortalamasına eşit olduğu değerdir. Aynı şekilde lojistik model için πi değerlerinin yüzdesinin maksimum olduğu durumda belirlilik katsayısı maksimum olacaktır. Doğrusal modelde kullanılan düzeltilmiş belirlilik katsayısı, lojistik modelde doğrudan kullanılamaz. Lojistik model için Nagelkerke (1991) düzeltilmiş belirlilik katsayısı için şu formülü önermiştir [21]:

$$\bar{R}^2 = R^2 / \max(R^2) \quad (45)$$

### 5.9 Sahte (Pseudo) Belirlilik Katsayısı

Lojistik model için kullanılabilecek bir diğer belirlilik katsayısı ölçütü de sahte (pseudo) belirlilik katsayısıdır. Bu ölçüte doğru sınıflandırma yüzdesi ve gölge belirlilik katsayısı da denilmektedir.

Bu ölçüt kolayca hesaplandığı için kullanımı avantajlıdır. Bağımlı değişken kategorik olduğu durumda alt sınırı 0 ve üst sınırı da 1 değerini almaktadır. Bu ölçüt bağımsız değişken sayısından etkilendiği için serbestlik derecesi ile düzeltilmelidir [22], [23].

#### 5.9.1 Hosmer-Lemeshow İstatistiği

Hosmer-Lemeshow istatistiği, Hosmer ve Lemeshow' un 1980 yılında tahmin edilen olasılıkların gruplandırılması ile ilgili gerçekleştirdikleri çalışma sonucunda elde edilmiştir. Varsayalım ki J=n olsun ve bu durumda n tane sütuna karşılık n tane tahmin

edilmiş olasılık değeri küçükten büyüğe doğru sıralanmış şekilde gösterilsin. Bu olasılıklardan faydalanarak iki farklı şekilde gruplama yapılabilir [3]:

1. Tahmin edilen olasılık değerlerinin yüzdelerine göre elde edilen olasılık tablosu gruplandırılır. (Örneğin en küçük %10'luk en büyük % 10'luk grup gibi)
2. Tahmin edilen olasılık değerlerinin belirlenmiş eşik değerlerine göre olasılık tablosu gruplandırılır.

Örneğin ilk yönteme göre grup sayısı 10 olduğunda ilk grup nı=n/10 birim en küçük olasılıkları içerem değerleri içermektedir. İkinci kullanılan yöntemde 10 tane gruba karşılık k tane eşik değeri olmak üzere k/10 değerine göre düzenlenmiş tahmin edilen olasılıklar bulunmaktadır. Y bağımlı değişkeninin alacağı değerlere göre y=1 satırı için ortalamaların tahmini grup içinde yer alan tahmin edilen olasılıkların toplamı ile hesaplanmaktadır. Y=0 satırı için de Y=1 satırı için bulunan olasılıkların toplamı 1' den çıkarılır.

Yukarıda gösterilen her iki gruplama yöntemi için de Hosmer-Lemeshow istatistiği, Pearson ki kare istatistiği ve g*2'lik tabloda gözlenen değerler ile beklenen değerler frekanslar ile şu şekilde hesaplanır:

$$\hat{C} = \sum_{k=1}^{g} \frac{(O_k - n'_k \bar{\pi}_k)^2}{n'_k \bar{\pi}_k (1 - \bar{\pi}_k)} \tag{46}$$

Denklem (46)' da yer $n'_k$ değerleri k. grupta bulunan birim sayısını göstermektedir. $O_k$ değeri ise her bir değişkene karşılık gelen y sayısını göstermektedir. $\bar{\pi}_k$ değeri tahmin edilen olasılığın beklenen değerini belirtmektedir.

Hosmer ve Lemeshow' un türetmiş oldukları bu test istatistiği, J=n olduğu durumda kurulan lojistik model başarılı ise (g-2) serbestlik derecesi ile Ki-kare dağılımına uygunluk göstermektedir. J=n olmadığı durumlarda dahi bu istatistik yine ( g-2) serbestlik derecesi ile ki kare dağılımına yakınsamaktadır [6].

# BÖLÜM 6

# ÇOK TERİMLİ (MULTİNOMİNAL) LOJİSTİK REGRESYON ANALİZİ

## 6.1 Çok Terimli (Multinominal) Lojistik Modelin Varsayımları

Çok terimli lojistik regresyon modellerinde bağımlı değişken sınıflayıcı ölçme düzeyinde ölüçülmüş ve en az üç kategoriye sahip olmalıdır. Çok terimli lojistik model, esas olarak bağımlı değişkenin iki değer alabildiği dikotomik durumun bir uzantısı olarak ifade edilebilir. Modelin ikili lojistik modele benzerliğinden dolayı varsayımları da benzerdir. Genel olarak çoklu lojistik modelin varsayımlarını şu şekilde sıralayabiliriz:

- Y bağımlı değişkenine ait tekrarlı değerler birbirinden bağımsızdır. Aynı değerler ile kodlanmış değerlerin birbirleri ile ilişkisi yoktur.
- Bağımsız X değişkenleri arasında çoklu doğrusal bağlantı yoktur.
- X' in verilen herhangi bir gözlem değerine ait Y sonucunun olasılığı $P_{ji}$ için verilen fonksiyon kullanılarak bulunmaktadır.
- Çok terimli lojistik regresyon analizinde temel bir grup seçilerek diğer gruplar arasında karşılaştırma yapılabilir. İkili lojistik regresyon modelinden ayrıştığı en temel nokta bu karşılaştırma metodudur [1], [7], [16], [24], [25].

## 6.2 Çok Terimli (Multinominal) Lojistik Modelin Kurulması

Bağımlı Y değişkeninin ikiden fazla değer aldığı durumlarda çok terimli lojistik regresyon analizi kullanılmaktadır. Örneğin bağımlı değişken iyi, orta, kötü gibi üç

farklı düzey aldığı durumlarda ikili (binary) lojistik regresyon analizi yerine çok terimli lojistik regresyon analizi kullanılır [18].

Çok terimli lojistik modeller kurulurken bağımlı değişkenin en üç değer alması gerekir. Genel olarak bağımlı değişken dikotomik olduğunda da bu yöntem kullanılabilir. Bu yönteme göre gözlemleri sınıflandırmamız için gruplar arasında ayrı ayrı karşılaştırma yapılmaktadır. Örneğin bağımlı Y değişkeni 0,1,2 gibi üç farklı değer alsın. Bu durumda Y=0 düzeyini temel grup olarak belirlediğimiz takdirde, Y=1 için Y=0 ve Y=2 için Y=0 karşılaştırması yapılması gerekmektedir. Bu karşılaştırmalar p değişken ve bir sabit terim içeren lojistik fonksiyonlar kullanılarak şu şekilde gösterilmektedir:

$$g_1(X) = In\left[\frac{P(y=1/X)}{P(Y=0/X)}\right] = \beta_{10} + \beta_{11}X_1 + \beta_{12}X_2 + \ldots\ldots\ldots\beta_{1p}X_p \quad (47)$$

$$g_2(X) = In\left[\frac{P(y=2/X)}{P(Y=0/X)}\right] = \beta_{20} + \beta_{21}X_1 + \beta_{22}X_2 + \ldots\ldots\ldots\beta_{2p}X_p \quad (48)$$

Bağımlı değişkenim üç seviyeli olduğu durumda lojistik fonksiyonun genel gösterimi şu şekildedir:

$$P(Y=0/X) = \frac{e^{g_j(x)}}{\sum_{k=0}^{2} e^{g_k(x)}} \quad (49)$$

Üç seviyeli lojistik modelin olabilirlik fonksiyonunu kurabilmemiz için grup üyeliğini belirlemede iki değer alabilen üç farklı değişken kullanılır. Bu değişkenler bağımlı değişken Y=0 için $y_0 = 0, y_1 = 1, y_2 = 0$, Y=1 için $y_0 = 0, y_1 = 1, y_2 = 0$ ve Y=2 için $y_0 = 0, y_1 = 1, y_2 = 0$ olmak üzere kukla değişkenler olarak tanımlanmaktadır. Kukla y değişkenlerine göre y' nin tüm değerleri için $\sum_{j=0}^{2} y_{j=1}$ koşulu sağlandığı takdirde n tane bağımsız gözlem değeri için olabilirlik fonksiyonu şu şekilde tanımlanmaktadır:

$$l(\beta)=\prod_{i=1}^{n}\left[\pi_o(x_i)^{y0i}.\pi_1(x_i)^{y1i}.\pi_2(x_i)^{y2i}\right] \tag{50}$$

Bağımlı değişken 3 değer alabildiğinde temel bir grup seçerek ikişerli karşılaştırmalar yapılmaktadır. Buna göre, bağımlı değişken g farklı değer aldığında (g-1) tane lojistik model kurulacaktır. Bu durumda denklem (50)'de bulunan olasılıkları j=1,2,3……..g için yazdığımız zaman, g grup için şu eşitlik elde edilir [9]:

$$P_j(x)=\frac{e^{\beta' j(x)}}{\sum_{k=0}^{g-1} e^{\beta' j(x)}} \tag{51}$$

Bağımlı değişken g farklı değer aldığında j. grupta tekrarlı gözlem olduğu takdirde j=1,2,3……g ve i=1,2,3…….j için aşağıdaki eşitlik elde edilir:

$$P_j(x)=\frac{e^{\beta'_j(x)}}{\sum_{k=0}^{g-1} e^{\beta'_j(x)}} \tag{52}$$

Bağımlı değişken ikiden fazla değer aldığı durumda denklem (52) kullanılarak gözlemlerin sınıflaması yapılmaktadır. Buna göre, diğer grupların temel grup ile karşılaştırılması şu şekilde gerçekleştirilir:

$$\frac{F_s(x)}{F_g(x)}=e^{\beta'_s(x)} \tag{53}$$

Denklem (53)' ün doğal logaritması alınarak, karşılaştırma yapabilmek için aşağıda gösterilen lojistik model kullanılmaktadır:

$$In\left(\frac{F_s(x)}{F_g(x)}\right)=\beta_s^{'}(x) \tag{54}$$

$F_s(x)$ : s. grup için olasılık yoğunluk fonksiyonu

$F_g(x)$ : Temel grup için olasılık yoğunluk fonksiyonu

Hs, grup üyeliği göstergesi olarak tanımlansın. $H_1, H_2 H_3 H_4 \ldots\ldots\ldots H_g$ şeklinde g farklı grup ve her gruba ait $n_1, n_2, n_3, n_4 \ldots\ldots\ldots n_g$ şeklinde gözlemler gösterilsin. Bu durumda her bir gruba ilişkin öncelikle β katsayı terimleri elde edilmelidir. G farklı grup için (g-1) katsayı elde edilir. Bağımlı değişkenin alacağı g farklı değer için (g-1) tane lojistik fonksiyon gereklidir ve p değişken için için (g-1).(p+1) tane tahmin elde edilir. Bu tahminler, olabilirlik fonksiyonları kullanılarak elde edilir.

Olabilirlik fonksiyonlarını tanımlayabilmek için sonsal olasılıkların hesaplanması gerekir. Koşullu ve rasgele örnekleme için kullanılan sonsal olasılıklar şu şekilde hesaplanmaktadır [9]:

$$P_s = P\left(H_s / x\right) = \frac{e^{\beta_s^{'} X}}{\sum_{i=1}^{g} e^{\beta_t^{'} X}} \tag{55}$$

Sonsal olasılıklar kullanılarak olabilirlik olasılıkları tanımlanmaktadır. Böylece iki gruba ait bir lojistik model, g grup için genellenebilmektedir. Örneğin n gözlemli örnekleme ait s. gruptaki i. gözlemi gösteren olabilirlik fonksiyonu şu şekilde tanımlanmaktadır:

$$L(X,\beta) = \prod_{i=1}^{n} P(H_{s(i)} / X_i) = \frac{\prod_{i=1}^{n} e^{\beta_{s(i)}^{'} X_i}}{\sum_{i=1}^{g} e^{\beta_t^{'} X}} \tag{56}$$

Bu fonksiyon ayrıca şu şekilde de gösterilebilir:

$$L(X,\beta)=\prod_{i=1}^{n} e^{\sum_{t=1}^{g} y_{ti}.In(P(H_t / X_i)} \qquad (57)$$

Bu fonksiyona bakılarak i. gözlemin H grubuna ait olup olmamasına göre, fonksiyonun değerinin 0 veya 1 olduğu tespit edilir.

Yukarıda tanımlanan denklem (56) ve (57) kullanılarak en çok olabilirlik yöntemi ile β katsayıları hesaplanmaktadır. Hesaplanan katsayıların tahmincileri hesaplandığında, çok terimli lojistik modeli kullanarak gözlemleri ayırabilmek için iki ölçüt kullanılır:

- $P(H_s / x) = P(H_t / x)$ $1 \le t \le g$ için
- $(\beta_s - \beta_t)'$ xi>0 için

Bu iki ölçütten herhangi biri gözlemlerin gruplara atanmasında kullanılabilir. Bu iki ölçütten biri dahi sağlanıyorsa, i. gözlem Hs grubuna atanmaktadır.

### 6.3 Çok Terimli (Multinominal) Lojistik Modelin Parametrelerinin Tahmini

Çok terimli lojistik modelin parametre tahminlerini yapabilmek için en çok olabilirlik metodu kullanılmaktadır. Bu metodu uygulayabilmek için öncelikle en çok olabilirlik fonksiyonunun tanımlanması gerekir. Logaritmik dönüşüm uygulanmış log olabilirlik fonksiyonu şu şekilde gösterilmektedir [6]:

$$L(\beta)=\sum_{i=1}^{n} y_{1i}.g_1(x_i)+y_{2i}g_2(x_i)-In(1+e^{g_1(x_i)}+e^{g_2(x_i)} \qquad (58)$$

Bu olabilirlik fonksiyonundan yararlanılarak olabilirlik eşitlikleri bulunmaktadır. Olabilirlik eşitliklerini bulabilmek için L(β) fonksiyonunun sırasıyla 2(p+1) parametre için kısmi türevi alınır. Bu eşitliklerin genel biçimi şu şekilde tanımlanır:

$$\frac{\partial L(\beta)}{\partial \beta_{jk}} = \sum_{i=1}^{n} x_{ki}(y_{ji} - \pi_{ji}) \qquad \text{k=0,1,2……p} \qquad \text{j=1,2} \tag{59}$$

Denklem (59)' daki olabilirlik eşitlikleri çözüldüğünde $\hat{\beta}$ tahminci değerleri her bir parametre için hesaplanır. Bu denklemlerin çözümü, ikili lojistik regresyon modelinde olduğu gibi iteratif yöntemlerle gerçekleştirilir [19].

L(β) fonksiyonunun ikinci kısmi türevleri alındığı zaman bilgi matrisi olarak tanımlayabileceğimiz bir matris elde ederiz. Bu matrisi kullanarak $\hat{\beta}$ tahmincilerinin varyans-kovaryans matrisi elde edilir. İkinci kısmi türevleri alınmış fonksiyonun genel çözümü şu şekilde gösterilmektedir [6]:

$$\frac{\partial^2 L(\beta)}{\partial \beta_{jk} \partial \beta_{jk'}} = -\sum_{i=1}^{n} x_{k'i} x_{ki} \pi_{ji}(1 - \pi_{ji}) \tag{60}$$

$$\frac{\partial^2 L(\beta)}{\partial \beta_{jk} \partial \beta_{jk'}} = -\sum_{i=1}^{n} x_{k'i} x_{ki} \pi_{ji} \pi_{j'i} \tag{61}$$

Denklem (60) ve (61)' de ifade edilen bileşenlerin negatif değerleri kullanılarak, 2(p+1)*2(p+1) satır ve sütun boyutuna sahip bilgi matrisi kullanılır. Bu matrisin bileşenleri $\hat{\beta}$ değerlerinin kullanılmasıyla oluşmaktadır. Bilgi matrisi I($\hat{\beta}$) şeklinde gösterilmektedir. Bilgi matrisinin tersi alınarak $\hat{\beta}$ tahmincilerinin varyans-kovaryans matrisleri elde edilmektedir [6], [15].

$$Var(\hat{\beta}) = I(\hat{\beta})^{-1} \tag{62}$$

Bilgi matrisi, ayrıca ikili lojistik regresyon analizine benzer biçimde daha kısa bir yöntemle hesaplanabilir. Varsayalım, X matrisi n*(p+1) boyutlu her bir durum için

kovariyetleri içeren bir matris olsun. *Vj* matrisi de n*n boyutlu, $\hat{\pi}_{ji}(1-\hat{\pi}_{ji})$ elemanlarını içeren diagonal bir matris olsun. Vj matrisi için i=1,2,3,……n şeklinde tanımlansın. Bu durumda bağımlı değişken üç değer alabildiğinde $V_1, V_2, V_3$ şeklinde üç farklı matris tanımlanacaktır. Bilgi matrisinin elemanlarını şu şekilde gösterelim [6]:

$$\hat{I}(\hat{\beta}) = \begin{bmatrix} \hat{I}(\hat{\beta})_{11} & \hat{I}(\hat{\beta})_{12} \\ \hat{I}(\hat{\beta})_{21} & \hat{I}(\hat{\beta})_{22} \end{bmatrix} \tag{63}$$

$\hat{I}(\hat{\beta})_{11}$ , $\hat{I}(\hat{\beta})_{12}$ ve $\hat{I}(\hat{\beta})_{22}$ ifadeleri için matris notasyonları gösterildiğinde, bilgi matrisine kısmi türev alınmadan ulaşılır. Bilgi matrisinin bileşenlerini matris notasyonu ile şu şekilde gösterebiliriz:

$$\hat{I}(\hat{\beta})_{11} = (X'V_1X) \tag{64}$$

$$\hat{I}(\hat{\beta})_{22} = (X'V_2X) \tag{65}$$

$$\hat{I}(\hat{\beta})_{21} = \hat{I}(\hat{\beta})_{12} = -(X'V_3X) \tag{66}$$

### 6.4 Çok Terimli (Multinominal) Lojistik Modelde Değişken Seçimi

Çok terimli lojistik model için değişken seçiminde başlıca şu yöntemler kullanılmaktadır [14], [15], [19]:

- Wald istatistiği
- Olabilirlik oran testi
- Skor testi

- Koşullu skor testi

Doğrusal modelin regresyon katsayılarının anlamlılığını test edebilmek için t istatistiği kullanılır. Çok terimli lojistik modelin katsayılarının anlamlılığını sınamak için doğrusal regresyonda kullanılan t istatistiği ile aynı işlevi gören Wald istatistiği kullanılmaktadır. Wald istatistiği şu şekilde gösterilebilir:

$$W = \frac{\hat{\beta}_i}{\hat{SE}(\hat{\beta}_i)} \tag{67}$$

W istatistiği, t istatistiğine benzer şekilde formülize edilir ancak t dağılımına uymamaktadır. W istatistiği, örnek sayısı yeterince büyük olduğu durumda asimptotik olarak normal dağılıma uymaktadır. W istatistiğinin karesi alındığında 1 serbestlik derecesi ile ki kare dağılımına uygunluk göstermektedir. Wald istatistiği ile değişkenler arası karşılaştırma yapılarak anlamlı olan değişkenler lojistik modele dahil edilir. Aşağıdaki hipotez yardımıyla modele giren katkıyaların anlamlılığı sınanır [6]:

$H_0 : \beta = 0.$

$H_1 : \beta \neq 0.$

Olabilirlik oranı testi , temel olarak değişkenin modele dahil olduğu ve değişkenin modele dahil olmadığı durumlar arasındaki farka bakarak karşılaştırma gerçekleştirir. Bu test, ikili lojistik modelde kullanılan D istatistiği ile aynı işleve sahiptir. Test istatistiği 1 serbestlik derecesi ile ki kare dağılımına uygunluk gösterir. Moulton (1993) olabilirlik oranı testini geliştirebilmek için düzeltme faktörü önermektedir. Olabilirlik oranı testi, Wald testine göre daha kullanışlı bir karşılaştırma yöntemidir [18].

Skor testi, log olabilirlik yöntemine göre kısmi türevleri alınmış olabilirlik eşitliklerinden yola çıkarak skor istatistiklerinin hesaplanması ile gerçekleştirilir. Skor testi istatistiği 1 serbestlik derecesi ile ki kare dağılımına uymaktadır. Koşullu skor testi ise tüm sonuçlarının kombinasyonlarının anlamlılıklarının kontenjans tabloları kullanılarak gerçekleştirilir. Koşullu skor testi hem ikili, hem de çok terimli lojistik modeller için kullanılabilmektedir [2], [10].

## 6.5 Çok Terimli (Multinominal) Lojistik Modelde Uyum İyiliği Ölçüleri

İkili lojistik modelde kullanılan uyum iyiliği ölçülerine benzer şekilde, ki kare ölçütü ve sapma ölçütü de çok terimli lojistik modeller için kullanılabilmektedir [15].

Ki kare ölçütü şu şekilde gösterilir:

$$\chi^2 = \sum_{i=1}^{n} \chi^2 = \frac{\sum_{i=1}^{n} \sum_{t=1}^{g} (y_{ti} - \hat{\pi}_{ti})^2}{\hat{\pi}_{ti}(1 - \hat{\pi}_{ti})} \tag{68}$$

Sapma ölçütü şu şekilde gösterilir:

$$D = -2\sum_{i=1}^{n} d^2 = -2\sum_{i=1}^{n} \sum_{t=1}^{g} y_{ti}.In\,\hat{\pi}_{ti} \tag{69}$$

Çok terimli için geliştirilen uyum iyiliği ölçütleri iki grup için geliştirilen ölçütlerle benzerdir. Ancak denklem (69)' daki D istatistiği ikili lojistik modelden farklı olarak n-p(g-1) serbestlik dereceli ki kare dağılımına uymaktadır. Bunun yanısıra lojistik modelin uyum iyiliğinin değerlendirilmesinde Hosmer-Lemeshow(H-L) testi de kullanılabilir. Bu testin amacı tahmin edilen olasılık değerlerini gruplandırmaktır. Hosmer-Lemeshow test istatistiği, (g-2) serbestlik derecesi ile Ki-kare dağılımına yaklaşmaktadır [3].

# BÖLÜM 7

# UYGULAMA

Bu tez çalışmasında kullanılan veriler, Alman kredi verisinden alınmıştır [26].

Uygulama yapılan veri setinde 1000 kişiye ait çeşitli nicel ölçüm değerleri ve kategorik özellikler bulunmaktadır. Uygulama, SPSS paket programı kullanılarak gerçekleştirilmiştir.

Örnek veri setindeki denekler üzerinde yapılan ölçüm değerlerine dayanarak, banka müşterilerilerinin kredi ödeme durumlarını etkileyen faktörleri belirlemek ve gelecekteki kredi başvurularını sınıflandırabilmek amacıyla model kurulmuştur. Bu çalışmada 7 bağımsız 1 bağımlı değişken kullanılmıştır. Bağımlı değişken olarak kredi durumu kullanılmıştır. Bağımsız değişken olarak hesap dengesi, kredi amacı, tasarruf, yaş, medeni durum, cinsiyet ve daire tipi değişkenleri kullanılmıştır. Bağımlı ve bağımsız değişkenleri kodlanması şu şekildedir:

Kredi Durumu: Müşteriler, geçmiş ödeme performansları, güncel gecikme durumları, kredi kayıt bürosu bilgileri incelenerek 0: "İyi" ve 1: "Kötü" şeklinde kodlanmıştır. 90 gün ve üzeri gecikmesi olan müşteriler Bankacılık Düzenleme ve Denetleme Kurumu'nun da önerdiği gibi "Kötü" olarak işaretlenmiştir.

Hesap Dengesi: Müşterilerin hesap bakiyesini gösteren nicel değişkendir. Değişken 1: ">300", 2: "<=300", 3: "Aktif hesabı yok", 4: "Hesap dengesi yok" şeklinde kodlanmıştır.

Yaş: Kredi başvurusunda bulunan müşterilerin yaşını belirten nicel değişkendir.

Kredi Amacı: Kişilerin krediye başvurma amacını gösteren kategorik değişkendir. Değişken kişilerin amacına göre 1: "Diğer", 2: "Eşya Alımı", 3: "Hanehalkı uygulamaları", 4: "İkinci el taşıt", 5: "İş kredisi", 6:" Mesleki eğitim", 7: "Onarım", 8: "Tatil", 9: "Televizyon alımı", 10: "Yeni taşıt" şeklinde kodlanmıştır.

Tasarruf Düzeyi: Kredi başvurusunda bulunan müşterilere ait aylık tasarruf değerlerini gösteren nicel değişkendir.

Medeni Durum: Kredi başvurusunda bulunan müşterilerin medeni durumunu gösteren 1: "Evli", 2: "Bekar", 3: "Dul" olarak kodlanmış nitel değişkendir.

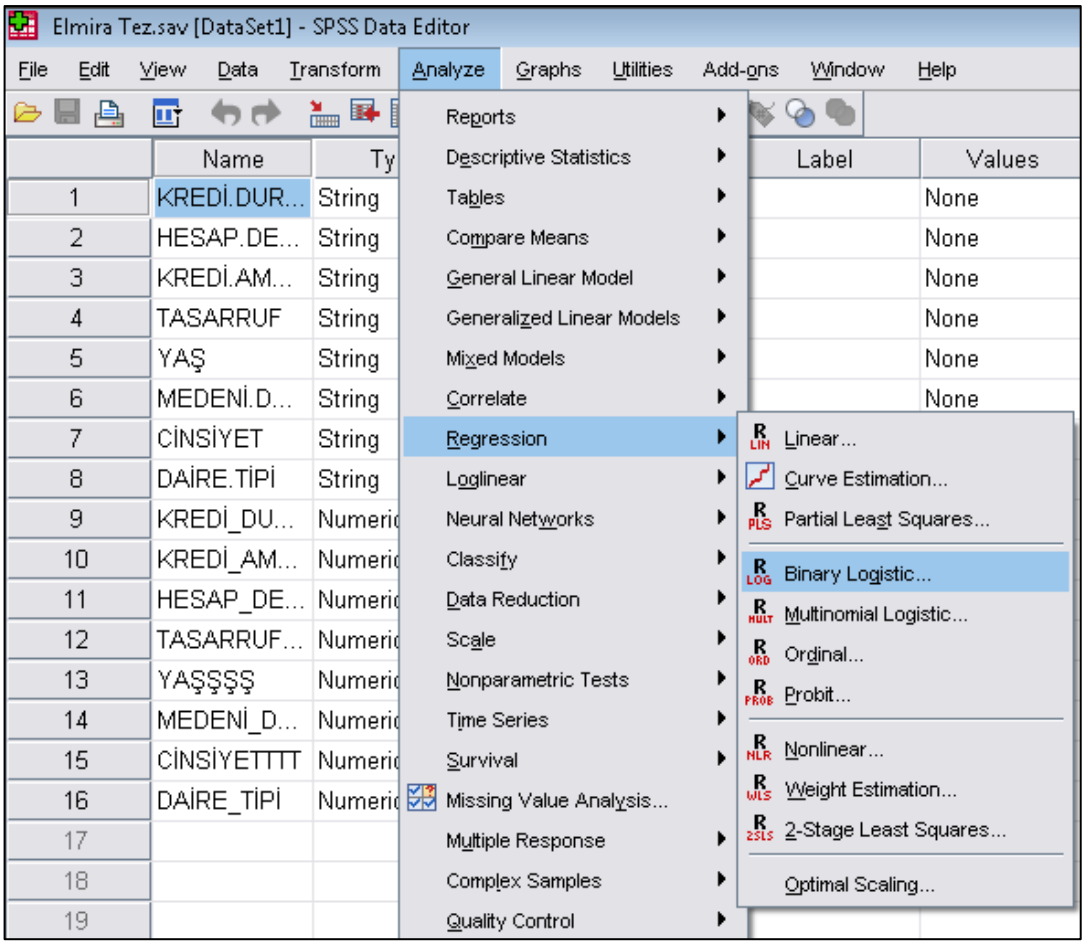
Cinsiyet: Kredi başvurusunda bulunan müşterilerin cinsiyetini gösteren 1: "Kadın", 2: "Erkek" olarak kodlanmış nitel değişkendir.

Daire Tipi: Kredi başvurusunda bulunan müşterilerin ev durumlarını gösteren 1: "Ev Sahibi", 2: "Kiralık", 3: "Ücretsiz" şeklinde kodlanmış nitel değişkendir.

SPSS paket programı kullanılarak lojistik regresyon analizi çıktıları aşağıda üretilmiştir.

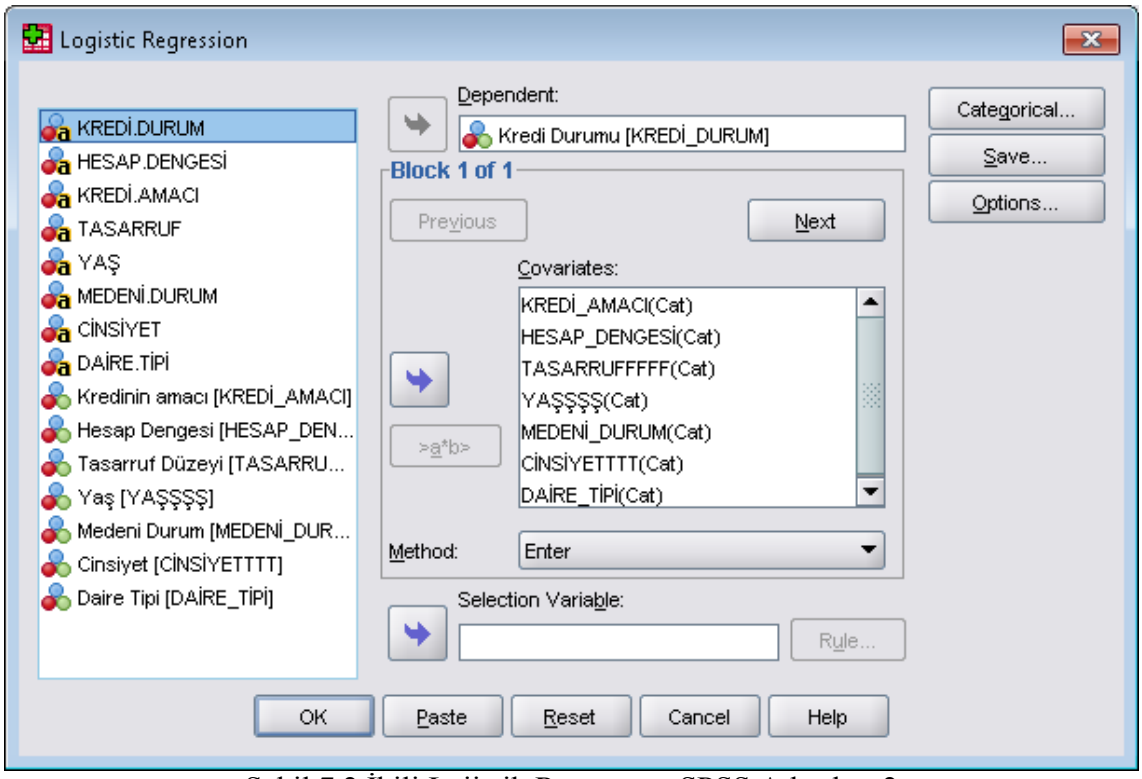
Lojistik regresyon analizine ait SPSS ekran görüntüleri ve çıktıları aşağıdaki tablolarda gösterilmiştir.

Bu analizde bağımlı değişken olarak ele alacağımız kredi durumu değişkeni, müşterilerin geçmiş ve güncel ödeme performansları baz alınarak belirlenen "İyi" ve "Kötü" şeklinde iki gruptan oluşmaktadır. Bağımlı değişkenin iki kategorili olduğu durumlarda ikili lojistik regresyon yöntemi kullanılmaktadır. SPSS'te aşağıdaki ekran görüntüsünden de görülebileceği gibi Analyze sekmesinden sırasıyla Regression ve Binary Logistic sekmeleri seçilir.



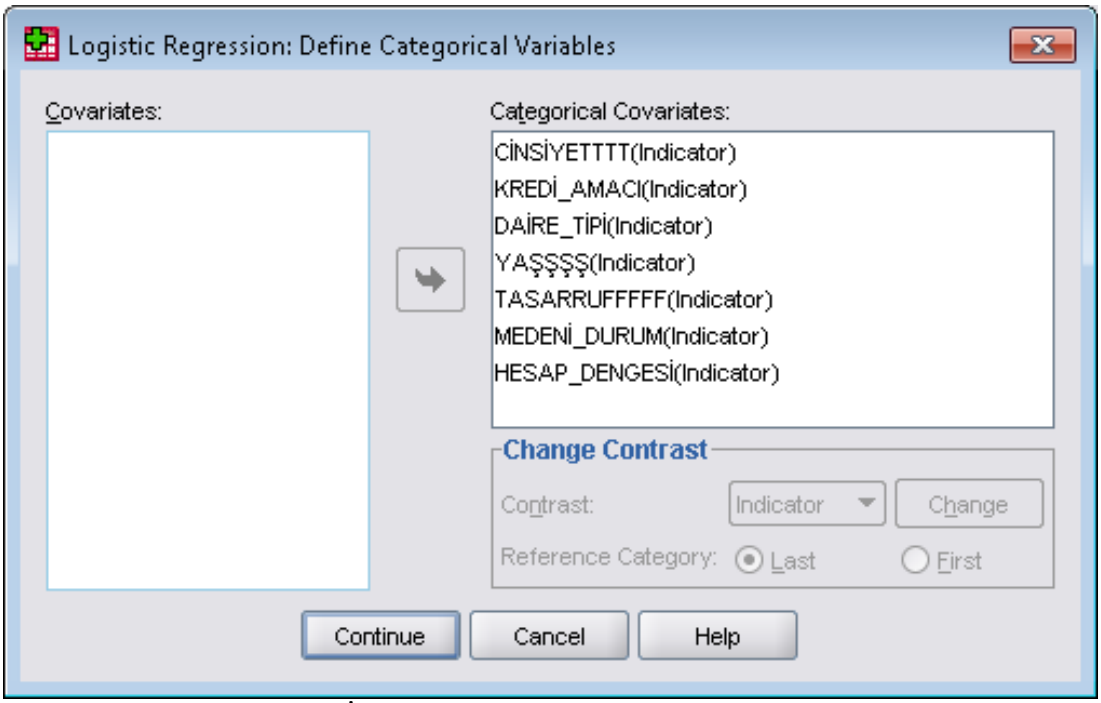

Şekil 7.1 İkili Lojistik Regresyon SPSS Adımları 1

Dependent kısmına Kredi Durumu bağımlı değişkeni sürüklenerek bırakılır. Modele girmesini istediğimiz bağımsız değişkenlerin tümü ise covariates kısmına atılır.



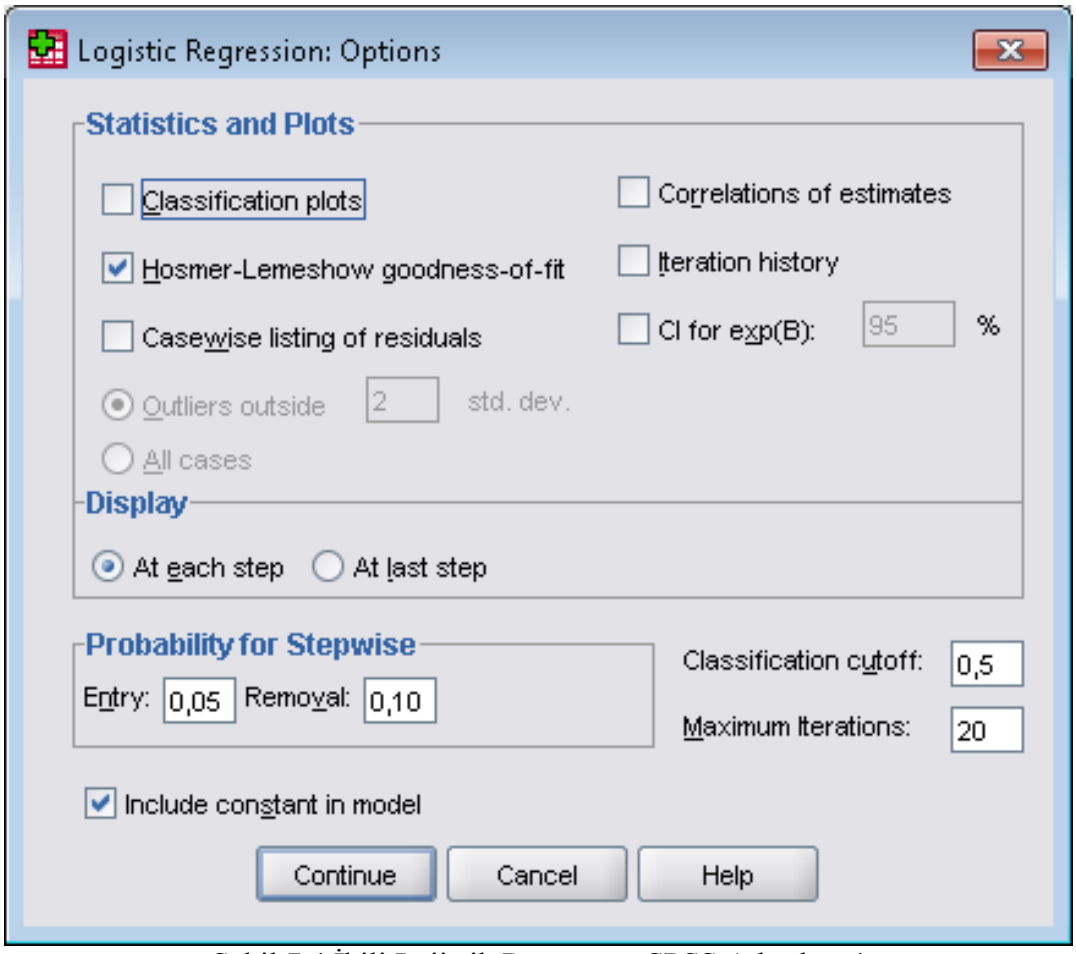

Şekil 7.2 İkili Lojistik Regresyon SPSS Adımları 2

Categorical sekmesi açılarak bağımsız değişkenler Categorical Covariates kısmına atıldıktan sonra Continue ile devam edilir.




Şekil 7.3 İkili Lojistik Regresyon SPSS Adımları 3

Options sekmesi açılarak aşağıdaki seçimler işaretlenerek Continue ile devam edilir.




Şekil 7.4 İkili Lojistik Regresyon SPSS Adımları 4

Çizelge 7.1 Kategorik Değişkenlerin Kodlanması

| | | | Parametre Kodlaması | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Sıklık | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| Kredinin amacı | Diğer | 234 | 1,000 | ,000 | ,000 | ,000 | ,000 | ,000 | ,000 | ,000 | ,000 |
| | Eşya Alımı | 280 | ,000 | 1,000 | ,000 | ,000 | ,000 | ,000 | ,000 | ,000 | ,000 |
| | Hanehalkı uygulamaları | 22 | ,000 | ,000 | 1,000 | ,000 | ,000 | ,000 | ,000 | ,000 | ,000 |
| | İkinci el taşıt | 181 | ,000 | ,000 | ,000 | 1,000 | ,000 | ,000 | ,000 | ,000 | ,000 |
| | İş kredisi | 12 | ,000 | ,000 | ,000 | ,000 | 1,000 | ,000 | ,000 | ,000 | ,000 |
| | Mesleki eğitim | 97 | ,000 | ,000 | ,000 | ,000 | ,000 | 1,000 | ,000 | ,000 | ,000 |
| | Onarım | 50 | ,000 | ,000 | ,000 | ,000 | ,000 | ,000 | 1,000 | ,000 | ,000 |
| | Tatil | 9 | ,000 | ,000 | ,000 | ,000 | ,000 | ,000 | ,000 | 1,000 | ,000 |
| | Televizyon alımı | 12 | ,000 | ,000 | ,000 | ,000 | ,000 | ,000 | ,000 | ,000 | 1,000 |
| | Yeni taşıt | 103 | ,000 | ,000 | ,000 | ,000 | ,000 | ,000 | ,000 | ,000 | ,000 |
| Tasarruf Düzeyi | 1400> | 183 | 1,000 | ,000 | ,000 | ,000 | | | | | |
| | <140 | 103 | ,000 | 1,000 | ,000 | ,000 | | | | | |
| | 140-700 | 63 | ,000 | ,000 | 1,000 | ,000 | | | | | |
| | 700-1400 | 48 | ,000 | ,000 | ,000 | 1,000 | | | | | |
| | Tasarruf yok | 603 | ,000 | ,000 | ,000 | ,000 | | | | | |
| Hesap Dengesi | 300> | 394 | 1,000 | ,000 | ,000 | | | | | | |
| | <=300 | 63 | ,000 | 1,000 | ,000 | | | | | | |
| | Aktif hesabı yok | 274 | ,000 | ,000 | 1,000 | | | | | | |
| | Hesap dengesi yok | 269 | ,000 | ,000 | ,000 | | | | | | |
| Yaş | 35> | 136 | 1,000 | ,000 | ,000 | | | | | | |
| | <15 | 476 | ,000 | 1,000 | ,000 | | | | | | |
| | 15- 25 | 157 | ,000 | ,000 | 1,000 | | | | | | |
| | 25-35 | 231 | ,000 | ,000 | ,000 | | | | | | |
| Medeni Durum | Evli | 548 | 1,000 | ,000 | | | | | | | |
| | Bekar | 360 | ,000 | 1,000 | | | | | | | |
| | Dul | 92 | ,000 | ,000 | | | | | | | |
| Daire Tipi | Ev sahibi | 107 | 1,000 | ,000 | | | | | | | |
| | Kiralık | 714 | ,000 | 1,000 | | | | | | | |
| | Ücretsiz | 179 | ,000 | ,000 | | | | | | | |
| Cinsiyet | Bayan | 310 | 1,000 | | | | | | | | |
| | Erkek | 690 | ,000 | | | | | | | | |

Çizelge 7.2 Lojistik regresyon modeli için katsayı tahmin sonuçları

| | | B | S.E. | Wald | df | Sig. | Exp(B) | 95% C.I.for EXP(B) | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | Lower | Upper |
| Step 1[a] | KREDİ_AMACI | | | 23,828 | 9 | ,005 | | | |
| | KREDİ_AMACI(1) | 1,099 | ,334 | 10,819 | 1 | ,001 | 3,001 | 1,559 | 5,775 |
| | KREDİ_AMACI(2) | ,405 | ,346 | 1,369 | 1 | ,242 | 1,499 | ,761 | 2,955 |
| | KREDİ_AMACI(3) | ,982 | ,567 | 2,993 | 1 | ,084 | 2,669 | ,878 | 8,116 |
| | KREDİ_AMACI(4) | ,656 | ,350 | 3,511 | 1 | ,061 | 1,927 | ,970 | 3,825 |
| | KREDİ_AMACI(5) | ,874 | ,683 | 1,639 | 1 | ,200 | 2,397 | ,629 | 9,140 |
| | KREDİ_AMACI(6) | 1,109 | ,387 | 8,223 | 1 | ,004 | 3,030 | 1,420 | 6,465 |
| | KREDİ_AMACI(7) | 1,320 | ,435 | 9,211 | 1 | ,002 | 3,743 | 1,596 | 8,779 |
| | KREDİ_AMACI(8) | -,483 | 1,165 | ,172 | 1 | ,679 | ,617 | ,063 | 6,057 |
| | KREDİ_AMACI(9) | ,769 | ,741 | 1,077 | 1 | ,299 | 2,157 | ,505 | 9,219 |
| | HESAP_DENGESİ | | | 81,194 | 3 | ,000 | | | |
| | HESAP_DENGESİ(1) | -1,491 | ,211 | 50,092 | 1 | ,000 | ,225 | ,149 | ,340 |
| | HESAP_DENGESİ(2) | -,869 | ,346 | 6,307 | 1 | ,012 | ,420 | ,213 | ,826 |
| | HESAP_DENGESİ(3) | ,283 | ,191 | 2,193 | 1 | ,139 | 1,327 | ,912 | 1,931 |
| | TASARRUF | | | 15,442 | 4 | ,004 | | | |
| | TASARRUF(1) | -,691 | ,235 | 8,627 | 1 | ,003 | ,501 | ,316 | ,795 |
| | TASARRUF(2) | -,084 | ,254 | ,110 | 1 | ,740 | ,919 | ,558 | 1,513 |
| | TASARRUF(3) | -,398 | ,366 | 1,188 | 1 | ,276 | ,671 | ,328 | 1,375 |
| | TASARRUF(4) | -1,301 | ,476 | 7,486 | 1 | ,006 | ,272 | ,107 | ,691 |
| | YAŞ | | | 9,132 | 3 | ,028 | | | |
| | YAŞ(1) | -,221 | ,274 | ,649 | 1 | ,420 | ,802 | ,469 | 1,372 |
| | YAŞ(2) | ,430 | ,201 | 4,589 | 1 | ,032 | 1,538 | 1,037 | 2,280 |
| | YAŞ(3) | ,153 | ,257 | ,356 | 1 | ,551 | 1,166 | ,704 | 1,930 |
| | MEDENİ DURUM | | | 7,879 | 3 | ,041 | | | |

Çizelge 7.2 Lojistik regresyon modeli için katsayı tahmin sonuçları(devam)

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| MEDENİ DURUM(1) | -,122 | ,283 | ,185 | 1 | ,668 | ,886 | ,508 | 1,542 |
| MEDENİ_DURUM(2) | ,322 | ,291 | 1,222 | 1 | ,269 | 1,380 | ,780 | 2,443 |
| MEDENİ DURUM(3) | ,491 | ,416 | 1,391 | 1 | ,238 | 1,634 | ,723 | 3,694 |
| DAİRE TİPİ | | | 9,560 | 2 | ,008 | | | |
| DAİRE TİPİ(1) | ,185 | ,296 | ,391 | 1 | ,532 | 1,203 | ,674 | 2,148 |
| DAİRE TİPİ(2) | -,448 | ,203 | 4,879 | 1 | ,027 | ,639 | ,429 | ,951 |
| Sabit | -,964 | ,452 | 4,537 | 1 | ,033 | ,381 | | |

Yukarıdaki tabloda bağımsız değişkenler için katsayılara ait parametre tahmin sonuçları gösterilmektedir. Elde edilen sonuçlara göre kredinin amacı, hesap dengesi, tasarruf, yaş, medeni durum ve daire tipi değişkenlerinin bireylerin kredi durumu üzerinde anlamlı bir etkiye sahip olduğu görülmektedir. Tüm bağımsız değişkenler için son kategori değeri referans kategorisi olarak alınmıştır. Referans kategorisine göre üsteli alınmış beta katsayılarını kullanarak çıkarsamalar yapılabilir.

Lojistik model kullanılarak seçilen bir kişi için bağımsız değişken değerleri modelde yerine konulduğu zaman 0 ile 1 arasında bir değer elde edilir. Bu şekilde kişinin hangi gruba atanacağı belirlenir. Elde edilen olasılık değeri 0,5' ten küçük olduğu durumda kişi 0 numaralı gruba, olasılık değeri 0,5' ten büyük veya eşit olduğu durumda 1 numaralı gruba atanacağı tahmin edilir.

Tablo 1' e bakıldığında her bağımsız değişken için β katsayıları, standart hata değerleri, Wald istatistikleri ve anlamlılık değerleri görülmektedir. Bağımsız değişkenlere ait β katsayılarının anlamlılıklarını sınayabilmek için Wald testi kullanılmaktadır.

$H_0 : \beta = 0.$

$H_1 : \beta \neq 0.$

Wald test istatistiği örneklem mevcudu yeteri kadar büyük olduğu zaman ki kare dağılımına uygunluk göstermektedir. Modele giren anlamlı 6 değişken kendi içinde kategorileri bazında da incelenmiş ve modele katkı sağlayan seviyeleri belirlenmiştir.

Bir değişkenin her kategorisi anlamlı olmayabilir, sadece tek bir kategorisi anlamlı olduğu için modele katkı sağlamış olarak kabul edilmektedir.

Modele giren değişkenlerden en etkili olanını bulmak için değişkenlerin B değerleri S.E. değerlerine bölünerek ortaya çıkan Wald istatistiği sonuçları incelendiğinde Hesap Dengesi değişkeninin en büyük değere sahip olup en baskın değişken olarak modele girdiği görülmektedir.

Çizelge 7.3 Lojistik regresyon modeli için çok maddeli test sonuçları

**Omnibus Tests of Model Coefficients**

| | | Chi-square | df | Sig. |
|---|---|---|---|---|
| Step 1 | Step | 198,686 | 24 | ,000 |
| | Block | 198,686 | 24 | ,000 |
| | Model | 198,686 | 24 | ,000 |

Yukarıdaki tabloda beta katsayıları için modelin anlamlılığına ilişkin gerçekleştirilen çok maddeli test sonuçları gösterilmektedir. Test istatistiklerinin anlamlılık değerleri 0.00<0.05 olduğundan, en az bir beta katsayısının anlamlı olduğu sonucuna varılmaktadır. Sonuç olarak oluşturulan lojistik regresyon modelinin anlamlı olduğu söylenebilir.

Çizelge 7.4 Hosmer Lemeshow testi sonuçları

| | | Kredi Durumu = İyi | | Kredi Durumu = Kötü | | |
|---|---|---|---|---|---|---|
| | | Observed | Expected | Observed | Expected | Total |
| Step 1 | 1 | 94 | 95,573 | 6 | 4,427 | 100 |
| | 2 | 94 | 91,560 | 6 | 8,440 | 100 |
| | 3 | 86 | 87,827 | 14 | 12,173 | 100 |
| | 4 | 78 | 82,457 | 21 | 16,543 | 99 |
| | 5 | 78 | 77,586 | 22 | 22,414 | 100 |
| | 6 | 76 | 68,458 | 23 | 30,542 | 99 |
| | 7 | 55 | 60,821 | 43 | 37,179 | 98 |
| | 8 | 57 | 53,714 | 43 | 46,286 | 100 |
| | 9 | 48 | 46,264 | 52 | 53,736 | 100 |
| | 10 | 34 | 35,739 | 70 | 68,261 | 104 |

Yukarıdaki tabloda Hosmer-Lemeshow uyum iyiliği testi için uygulanan her adıma ilişkin kredi durumu değişkenine göre beklenen ve gözlenen değerler gösterilmektedir.

Çizelge 7.5: Hosmer Lemeshow testi sonuçları

**Hosmer and Lemeshow Test**

| Step | Chi-square | Df | Sig. |
|---|---|---|---|
| 1 | 7,966 | 8 | ,437 |

$H_0$ : Kurulan model ile veri arasında fark yoktur.

$H_1$ : Kurulan model ile veri arasında fark vardır. Model veriyi temsil etmez.

Yukarıdaki tabloda Hosmer-Lemeshow uyum iyiliği testi sonuçları gösterilmektedir. Hosmer Lemeshow test istatistiğinin anlamlılık değeri 0.437>0.05 olduğundan $H_0$ hipotezi kabul edilir ve modelin uyum iyiliğine sahip olduğu belirlenmiştir.

Çizelge 7.6 Lojistik regresyon modeli için özet sonuçlar

**Model Summary**

| Step | -2 Log likelihood | Cox & Snell R Square | Nagelkerke R Square |
|---|---|---|---|
| 1 | 1023,043[a] | ,180 | ,255 |

Yukarıdaki tabloda lojistik regresyon modeline ilişkin log-olabilirlik değeri ve belirtme katsayıları gösterilmektedir. Cox Snell ve Nagelkerke belirtme katsayılarına bakıldığında kredi durumu değişkeninin bağımsız değişkenler tarafından orta düzeyde açıklanabildiği sonucuna varılabilir.

$H_0$ : Sabit terim dışında bütün katsayılar sıfıra eşittir.

$H_1$ : Katsayılardan en az biri sıfırdan farklıdır.

-2LL değeri 1023,043'tür. Modelde 6 değişken anlamlı bulunduğundan serbestlik derecesi 6 ve anlamlılık düzeyi 0,05 olan Ki-kare tablo değerinden büyük olduğundan $H_0$ reddedilir. Katsayılardan en az biri istatistiki olarak anlamlı bir şekilde sıfırdan farklıdır. Model anlamlıdır.

Tablo 5'te gözlenen değerler ile tahmin edilen değerlerin karşılaştırılmasını yapabilmek için kullanılan -2LL ölçüsü ve uyum istatistikleri verilmiştir. Bu ölçü, modelin veriye uygunluğunu ölçmek için kullanılır. İyi bir modelin olması için -2 LL istatistiğinin de

küçük olması gerekir. Bu tabloda Cox ve Snell belirlilik katsayısına baktığımızda bağımlı değişken ile bağımsız değişkenler arasında %18'lik bir ilişki olduğunu söyleyebiliriz. Ancak bu ölçütün maksimum değeri 1 olmadığı için yorum yapabilmek güçleşmektedir. Nagelkerke belirlilik katsayısı tabloya göre % 25,5 olarak hesaplanmıştır. Bağımlı değişken ile bağımsız değişkenler arasında %25,5' lik bir ilişkinin olduğunu söyleyebiliriz.

Çizelge 7.7 Lojistik regresyon modeli için sınıflama oranı

**Classification Table[a,b]**

| Observed | | | Predicted | | |
|---|---|---|---|---|---|
| | | | Kredi Durumu | | |
| | | | İyi | Kötü | Percentage Correct |
| Step 0 | Kredi Durumu | İyi | 700 | 0 | 100,0 |
| | | Kötü | 300 | 0 | ,0 |
| | Overall Percentage | | | | 70,0 |

Çizelge 7.8 Lojistik regresyon modeli için sınıflama oranı tablosu

**Classification Table[a]**

| Observed | | | Predicted | | |
|---|---|---|---|---|---|
| | | | Kredi Durumu | | Percentage |
| | | | İyi | Kötü | Correct |
| Step 1 | Kredi Durumu | İyi | 616 | 84 | 88,0 |
| | | Kötü | 175 | 125 | 41,7 |
| | Overall Percentage | | | | 74,1 |

Yukarıdaki tabloda lojistik regresyon modeline ilişkin sınıflama oranları gösterilmektedir. Gözlenen değerlere karşılık tahmin edilen değerler matrisinden modelin genel başarı oranına ulaşılmaktadır. Elde edilen sonuçlara göre, modelin doğru sınıflama oranı % 74,1 olarak hesaplanmıştır. İkinci aşamadaki sınıflama oranının ilk aşamadakinden büyük olması beklenir. Çünkü ilk yapılan hesaplamada her değişkeni tek bir kategori olarak ele alır ve tahmin eder. Bu değere göre modelin sınıflama gücünün yüksek olduğu belirlenmiştir. Sınır değeri %50 olarak düşünülmelidir.

Bu çalışmada referans kategorisi olarak her değişkenin son kategorik değeri ele alınmıştır. Buna göre istatistiksel açıdan anlamlı bulunan kategoriler için şu çıkarsamalar yapılabilir:

Kredi Amacı Değişkeni:

- Kredi amacı diğer olanların kredi amacı yeni taşıt almak olanlara göre kredi durumu yaklaşık 3 kat daha iyi düzeydedir.
- Kredi amacı hane halkı uygulamaları olanların kredi amacı yeni taşıt almak olanlara göre kredi durumu yaklaşık 2,67 kat daha iyi düzeydedir.
- Kredi amacı mesleki eğitim olanların kredi amacı yeni taşıt almak olanlara göre kredi durumu yaklaşık 3 kat daha iyi düzeydedir.
- Kredi amacı onarım olanların kredi amacı yeni taşıt almak olanlara göre kredi durumu yaklaşık 3,74 kat daha iyi düzeydedir.

Hesap Dengesi Değişkeni:

- Hesap dengesi düzeyi 300 doların üzerinde olanların hesap dengesi olmayanlara göre kredi durumu yaklaşık 0,2 kat daha iyi düzeydedir.
- Hesap dengesi düzeyi 300 doların altında olanların hesap dengesi olmayanlara göre kredi durumu yaklaşık 0,4 kat daha iyi düzeydedir.

Tasarruflar:

- Tasarruf düzeyi 1400 doların üzerinde olanların tasarrufu olmayanlara göre kredi durumu yaklaşık 0,5 oranında daha iyi düzeydedir.
- Tasarruf düzeyi 700-1400 dolar arasında olanların tasarrufu olmayanlara göre kredi durumu yaklaşık 0,27 oranında daha iyi düzeydedir.

Yaş:

- Yaşları 15'in altında olanların 25-35 yaş aralığında olanlara göre kredi durumu yaklaşık 1,538 kat daha iyi düzeydedir.

Daire Tipi:

- Daire tipine göre kirada oturanların ücretsiz oturanlara göre kredi durumu yaklaşık 0,64 oranında daha iyi düzeydedir.

# SONUÇLAR

Bu çalışmada genel olarak lojistik regresyon analizi teorik olarak özetlenmiş, sonrasında bir uygulama sonucunda lojistik modelin sonuçları incelenmiştir. Lojistik regresyon bankacılıkta skorkart modellemelerinde en yaygın kullanılan istatistiksel tekniktir. Varsayımlarının az olması, bağımlı değişkenin sürekli değişken olmadığı durumlarda rahatlıkla kullanılabilmesi ve kolay yorumlanabilir olması önemli tercih nedenleridir.

Kurmuş olduğumuz lojistik model ile banka müşterilerinin kredi durumu bağımlı değişken olarak ele alınmıştır. Müşterilerin kredi ödeme performansını etkileyen etkenler incelenmiş ve bu etkenlerin etkileme seviyeleri tespit edilmiştir.

İncelemiş olduğumuz veri setinin yapısına göre hangi analiz türünün kullanılacağı teorik yapısıyla karşılaştırmalı olarak sunulmuştur. Bağımlı değişkenin kategorik veya nicel olmasına göre hangi modelin kullanılacağı belirlenmiştir. Bağımlı değişkenin yapısına göre kişileri gruplandırmak ve önsel bilgi ile seçilen değişkenlerin etki düzeylerini saptayabilmek için lojistik regresyon analizi tekniği kullanılmıştır.

Gerçekleştirdiğimiz araştırma sonucunda cinsiyet değişkeninin bağımlı değişken üzerinde etkisi olmadığı gözlenmiştir. Hesap Dengesinin bağımlı değişkeni etkileyen en önemli değişken olduğu görülmektedir. Lojistik regresyon analizi sonucunda kredi durumu iyi olan kişilerin %88'i, kredi durumu kötü olan kişilerin %41,7'si doğru tahmin edilmiştir. Lojistik modelin, kredi durumu iyi ve kötü olan tüm kişileri ayırt edici değişkenlere göre doğru gruplandırma olasılığı %74,1'dir ve başarılıdır.

# KAYNAKLAR


[1] Agresti, A., (1996). An Introduction to Categorical Data Analysis, New York, John Wiley & Sons

[2] Işığıçok, E., (2003). "Bebeklerin Doğum Ağırlıklarını ve Boylarını Etkileyen Faktörlerin Lojistik Regresyon Analizi İle Araştırılması". Ankara, VI. Ulusal Ekonometri ve İstatistik Sempozyumu Bildiri Kitabı, Gazi Üniversitesi İ.İ.B.F. Ekonometri Bölümü.

[3] Hosmer D. ve Lemeshow S., (2000). Applied Logistic Regression

[4] Tatlıdil, H., (1992). Uygulamalı Çok Değişkenli Istatistiksel Analiz, Ankara:Engin Yayınları

[5] Çolak, E., (2002). Koşullu ve sınırlandırılmış lojistik regresyon yöntemlerinin karşılaştırılması ve bir uygulama. (Basılmamış Yüksek Lisans Tezi) Osmangazi Üniversitesi, Eskişehir.

[6] Gujarati, D., (2006). Temel Ekonometri (Çeviren: Ümit Şenesen, Gülay Günlük Şensesen), 4. Baskı, İstanbul, Çevik

[7] Korkmaz, T., (2004). "Bankalarda Kredi Risk Ölçümünde Alternatif Yöntemler", Active Dergisi, Temmuz-Ağustos

[8] Ryan, T., (1997). Modern Regression Methods

[9] Kalaycı, Ş., (2009). SPSS Uygulamalı Çok Değişkenli İstatistik Teknikleri, 4. Baskı, Ankara, Asil Yayın Dağıtım

[10] Anderson, J. A., (1983). Robust İnference Using Logistic Models, Bulletion of ınternational Statistical Institute

[11] Cornfield, J., (1962). Joint Dependence Of The Risk Of Coronary Heart Disease On Serum Cholesterol And Sistolic Blood Pressure: A Diskrimant Function Analysis, Federation Proceedings

[12] Gardside, P.S., (1995). The Important Role of Modifiable Dietary and Behaviour Characteristic in the Causation and Prevention of Coronary Hearth Disease Hospitalization and Mortality. Journal of American College of Nutrition

[13] Oğuzlar, A., (2005). "Lojistik Regresyon Analizi Yardımıyla Suçlu Profilinin Belirlenmesi", Atatürk Üniversitesi İİBF Dergisi

[14] Aktaş, C. ve Erkun, O., (2009). "Lojistik Regresyon Analizi ile Eskişehir Sis Kestiriminin İncelenmesi", Eskişehir

[15] Kutlar, A., (2009). Uygulamalı Ekonometri, 3. Baskı, Ankara, Nobel Yayın Dağıtım

[16] Özdamar, K., (2004). Paket Programları ile İstatistiksel Veri Analizi 2, Eskişehir, Kaan Kitabevi

[17] Siddiqi, N., (2006). Credit Risk Scorecards, New York, John Wiley & Sons

[18] Şahin, E., (2007). Lojistik Regresyon Analizi ve Bir Uygulama, İstanbul, Yıldız Teknik Üniversitesi

[19] Landau S. ve Everitt B., (2004). A Handbook of Statistical Analyses Using SPSS

[20] Kocaeli Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, (2004). Sayı 2

[21] S.P.S.S., (2001). SPSS Regression Models 11.0, Chicago, S.P.S.S. Inc.

[22] Ordu Üniversitesi Sosyal Bilimler Enstitüsü Dergisi (SOBIAD), (2013). Sayı 7

[23] Abbott, R.D., (1985). Logistic Regression In Survival Analysis, American Journal of Epidemiology

[24] Turangil N., (2006). Kredi Skorlamada Kullanılan Yöntemler ve Uygulamaları, Fen Bilimleri Enstitüsü, Yıldız Teknik Üniversitesi, Yüksek Lisans Tezi, İstanbul

[25] Özdamar, K., (2004). Paket Programları ile İstatistiksel Veri Analizi 1, Eskişehir, Kaan Kitabevi

[26] German Credit Data, ftp://ftp.ics.uci.edu/pub/machine-learning-databases/statlog/german/, 01 Mayıs 2013

# ÖZGEÇMİŞ

## KİŞİSEL BİLGİLER

**Adı Soyadı** : Elmira KOCABAŞ

**Doğum Tarihi ve Yeri :** 12.01.1988, Kırcali

**Yabancı Dili** : İngilizce

**E-posta** : elmirakocabas@gmail.com

## ÖĞRENİM DURUMU

| Derece | Alan | Okul/Üniversite | Mezuniyet Yılı |
|---|---|---|---|
| Y. Lisans | İstatistik | Yıldız Teknik Üniversitesi | 2014 |
| Lisans | İstatistik | Yıldız Teknik Üniversitesi | 2011 |
| Lise | Sayısal | İst. Köy Hiz.Anadolu Lisesi | 2006 |

## İŞ TECRÜBESİ

| Yıl | Firma/Kurum | Görevi |
|---|---|---|
| 2011 | IPSOS KMG Araştırma ve Dan. Ştd. | İstatistikçi |
| 2012 | ING Bank A.Ş. | Kredi Risk Yönetimi Yetkilisi |
| 2013 | Türkiye Finans Katılım Bankası A.Ş. | Rating ve Skorlama Modelleri Geliştirme Yetkilisi |