T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

# SAĞKALIM ANALİZİ MODEL VE YÖNTEMLERİNİN TELEKOM SEKTÖRÜ MÜŞTERİ KAYIP VERİLERİ ÜZERİNE UYGULAMALARI

**Melik MASARİFOĞLU**

DOKTORA TEZİ

İstatistik Anabilim Dalı

İstatistik Programı

Danışman

Prof. Dr. Ali Hakan BÜYÜKLÜ

Şubat, 2023

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

# SAĞKALIM ANALİZİ MODEL VE YÖNTEMLERİNİN TELEKOM SEKTÖRÜ MÜŞTERİ KAYIP VERİLERİ ÜZERİNE UYGULAMALARI

Melik MASARİFOĞLU tarafından hazırlanan tez çalışması 13.02.2023 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı, İstatistik Programı **DOKTORA TEZİ** olarak kabul edilmiştir.

Prof. Dr. Ali Hakan BÜYÜKLÜ
Yıldız Teknik Üniversitesi
Danışman

**Jüri Üyeleri**

Prof. Dr. Ali Hakan BÜYÜKLÜ, Danışman
Yıldız Teknik Üniversitesi _____

Prof. Dr. Oya KALIPSIZ, Üye
Yıldız Teknik Üniversitesi _____

Prof. Dr. Gülay BAŞARIR, Üye
Mimar Sinan Güzel Sanatlar Üniversitesi _____

Prof. Dr. Semra ERPOLAT TAŞABAT, Üye
Mimar Sinan Güzel Sanatlar Üniversitesi _____

Dr. Öğr. Üyesi Doğan YILDIZ, Üye
Yıldız Teknik Üniversitesi _____

Danışmanım Prof. Dr. Ali Hakan BÜYÜKLÜ sorumluluğunda tarafımca hazırlanan Sağkalım Analizi Model ve Yöntemlerinin Telekom Sektörü Müşteri Kayıp Verileri Üzerine Uygulamaları başlıklı çalışmada veri toplama ve veri kullanımında gerekli yasal izinleri aldığımı, diğer kaynaklardan aldığım bilgileri ana metin ve referanslarda eksiksiz gösterdiğimi, araştırma verilerine ve sonuçlarına ilişkin çarpıtma ve/veya sahtecilik yapmadığımı, çalışmam süresince bilimsel araştırma ve etik ilkelerine uygun davrandığımı beyan ederim. Beyanımın aksinin ispatı halinde her türlü yasal sonucu kabul ederim.

Melik MASARİFOĞLU

İmza

*Aileme*

*ve*

*biricik eşime*

# TEŞEKKÜR

Tez çalışmamın gerçekleşmesinde değerli katkı ve yönlendirmelerinden dolayı tez danışmanım Sayın Prof. Dr. Ali Hakan BÜYÜKLÜ'ye değerli görüşleri ile tezin bu noktaya gelmesinde emeği olan tez izleme komitesindeki değerli hocalarıma içtenlikle teşekkür ederim.

Son olarak, desteğini her zaman yanımda hissettiğim canım eşim Sümeyra ve çocuklarım Tarık, Meva ve Serra'ya, annem Nesrin ve babam Faruk'a, istatistik bölümündeki değerli hocalarıma teşekkür ederim.

Melik MASARİFOĞLU

# İÇİNDEKİLER


**SİMGE LİSTESİ** vii

**KISALTMA LİSTESİ** viii

**ŞEKİL LİSTESİ** ix

**TABLO LİSTESİ** x

**ÖZET** xi

**ABSTRACT** xiii

**1 GİRİŞ** 1

1.1 Literatür Özeti ... 1

1.2 Tezin Amacı ... 2

1.3 Hipotez ... 3

1.4 Tezin İçeriği ... 4

**2 SAĞKALIM ANALİZİ** 5

2.1 Tanım ... 5

2.2 Model ... 6

2.2.1 Olasılık Yoğunluk Fonksiyonu ... 6

2.2.2 Yaşam Fonksiyonu ... 6

2.2.3 Tehlike Fonksiyonu ... 7

2.2.4 Birikimli Tehlike Fonksiyonu ... 8

2.2.5 Yaşam Fonksiyonu ile Tehlike Fonskiyonu Arasındaki İlişki ... 8

2.3 Durdurma ... 9

2.3.1 Sağdan Durdurma ... 10

2.4 Parametrik Olmayan Yöntemler (Kaplan-Meier) ... 11

**3 YARI PARAMETRİK COX REGRESYON MODELİ** 16

3.1 Model Tanımı ... 16

3.2 Tehlike Oranı ... 17

3.3 CRMO Kısmi Olabilirlik Fonksiyonu ... 18

3.3.1 Kesin Metodu ... 21

3.3.2 Breslow Metodu ... 23

3.3.3 Efron Metodu ... 24

3.4 Oransal Tehlike Varsayımı ... 26

3.5 Tabakalandırılmış (Stratified) CRMO ... 27

3.6 Genişletilmiş (Extended) CRMO ..... 28
3.7 Parçalı ve Kübik Spline CRMO ..... 29
**4 PARAMETRİK REGRESYON MODELİ 31**
4.1 Hızlandırılmış Başarısızlık Süresi Modeli ..... 31
4.1.1 Model ..... 31
4.1.2 Metot ..... 31
4.1.3 Weibull HBSM ..... 33
4.1.4 Log-Lojistik HBSM ..... 36
4.1.5 Log-Normal HBSM ..... 38
4.2 HBSM Olabilirlik Fonksiyonu ..... 39
4.3 Model Seçimi ..... 41
4.3.1 Akaike Information Criterion (AIC) ..... 41
**5 TELEKOM SEKTÖRÜ MÜŞTERİ KAYIP VERİLERİNE UYGULANIŞI 42**
5.1 Müşteri Kayıp Verileri ..... 42
5.2 Kaplan Meier ..... 46
5.3 Ön Hazırlık ..... 50
5.3.1 Kukla Değişken Ataması ..... 50
5.3.2 Çoklu Bağımlılık ve VIF (Variance Inflation Factor) ..... 50
5.4 Yarı Parametrik Cox Regresyon Modeli ..... 52
5.4.1 Panelizer Katsayısı ..... 52
5.4.2 Oransal Tehlike Varsayımı Kontrolü ..... 56
5.4.3 Tabakalandırılmış CRMO ..... 60
5.4.4 Genişletilmiş CRMO ..... 61
5.5 PRCRMO ve KSCRMO ..... 62
5.6 HBSM ..... 63
**6 SONUÇ VE ÖNERİLER 71**
**KAYNAKÇA 75**
**TEZDEN ÜRETİLMİŞ YAYINLAR 79**

# SİMGE LİSTESİ

| | |
|---|---|
| $\Lambda(t)$ | Birikimli Tehlike Fonskiyonu |
| $\delta_i$ | Durdurma Göstergesi |
| F(t) | Kümüle Dağılım Fonksiyonu |
| $\lambda(t)$ | Tehlike Fonskiyonu |
| S(t) | Yaşam Fonksiyonu |

# KISALTMA LİSTESİ

| | |
|---|---|
| CRMO | Yarı Parametrik Cox Regresyon Modeli |
| HBSM | Hızlandırılmış Başarısızlık Süresi Modeli |
| KM | Kaplan Meier |
| KOF | Kısmi Olabilirlik Fonksiyonu |
| KSCRMO | Kübik Spline Cox Regresyon Modeli |
| LRT | Log Rank Testi |
| OTV | Orantılı Tehlikeler Varsayımı |
| PRCRMO | Parçalı Regresyon Cox Regresyon Modeli |
| SKA | Sağkalım Analizi |
| VIF | Çoklu Bağımlılık Katsayısı |
| YF | Yaşam Fonksiyonu |

# ŞEKİL LİSTESİ

**Şekil 2.1** Sağdan durdurma örnek .....11
**Şekil 5.1** Çalışma takvimi .....42
**Şekil 5.2** Çalışma süresi .....43
**Şekil 5.3** KM sağkalım eğrisi .....47
**Şekil 5.4** KM tarife sağkalım eğrisi.....48
**Şekil 5.5** KM teknik destek sağkalım eğrisi .....49
**Şekil 5.6** Penalizer seçim grafiği .....52
**Şekil 5.7** OTV sağlamayan değişkenlerin Schoenfeld artıkları grafikleri.....58
**Şekil 5.8** OTV sağlamayan değişkenlerin log t ye göre (ln-lnS(t)) grafikleri .....59
**Şekil 5.9** HBSM qq grafikleri.....64
**Şekil 5.10** Weibull için açıklayıcı değişkenlerin %95 güven aralığı .....69
**Şekil 5.11** 1-5 müşteriler için Log-Lojistik sağkalım eğrisi tahmini.....69
**Şekil 5.12** 1000-1004 müşteriler için Log-Lojistik sağkalım eğrisi tahmini.....70

# TABLO LİSTESİ

**Tablo 1.1** Müşteri kazanım maliyet tablosu .....2
**Tablo 2.1** KM örnek .....12
**Tablo 2.2** Hata terimi ve güven aralığı .....13
**Tablo 2.3** LRT bir örnek .....14
**Tablo 3.1** Tehlike oranı değeri ve etkisi .....18
**Tablo 3.2** KOF için örnek sağkalım tablosu .....20
**Tablo 3.3** KOF tablosu .....20
**Tablo 3.4** KOF örnek .....22
**Tablo 4.1** Log T dağılımı .....33
**Tablo 5.1** Değişken tanımları ve düzeyleri .....44
**Tablo 5.2** Sağkalım olasılık tablosu .....46
**Tablo 5.3** LRT tarife .....49
**Tablo 5.4** LRT teknik destek .....49
**Tablo 5.5** VIF bağlılık seviyesi .....50
**Tablo 5.6** Değişken VIF değerleri .....51
**Tablo 5.7** CRMO sonuçları .....53
**Tablo 5.8** Log-rank testi sonuçları .....56
**Tablo 5.9** Düğüm sayısı ve düğüm yüzdeleri .....62
**Tablo 5.10** Yarı parametrik modellerin performansları .....65
**Tablo 5.11** Parametrik modellerin performansları .....66
**Tablo 5.12** Log-Lojistik HBSM çıktısı .....66

# ÖZET

## Sağkalım Analizi Model ve Yöntemlerinin Telekom Sektörü Müşteri Kayıp Verileri Üzerine Uygulamaları

Melik MASARİFOĞLU

İstatistik Anabilim Dalı

Doktora Tezi

Danışman: Prof. Dr. Ali Hakan BÜYÜKLÜ

Rekabetin yoğun yaşandığı Telekom sektöründe kaybedilen herbir müşteri, yerini doldurmanın maliyeti, azalan nakit akışı ve kaybedilen cirodan dolayı mobil operatör için finansal olarak büyük bir zarardır. Müşteriyi elde tutmak daha kolay, daha başarılabilir ve daha az maliyetlidir. Mobil operatörlerin elde tutulan müşteri ile uzun dönemli, kazan-kazan üzerine kurduğu ilişkiler, günümüzün rekabetçi ortamında varlıklarını sürdürebilmeleri için hayati önemdedir. Müşteriyi elde tutma boyutunda geliştirilecek stratejiler için önceden kayıp yaşanacak müşteriyi tahminlemek ve profillemek büyük avantajdır. Bu çalışmada müşterileri doğru ve etkin bir şekilde profilleyerek, müşteri kaybını minimize etmek amaçlanmıştır. Çalışmamızda mobil operatör müşteri kayıp verilerine, başta tıp olmak üzere farklı disiplinlerde sıkça kullanılan Sağ Kalım Analizi metotları ve yöntemleri başvurulmuştur. Parametrik Olmayan, Yarı Parametrik ve Parametrik Modeller olmak üzere üç ana başlıkta toplam ondokuz adet modele başvurulmuştur. Bu modeller sırasıyla Parametrik Olmayan Kaplan-Meier Metodu, Yarı Parametrik Modellerden Cox Regresyon Modeli (CRMO) ve CRMO için ön şart olan Oransal

Tehlike Varsayımı (OTV) kontrol edilmiş, OTV sağlanmadığından dolayı Tabakalandırılmış CRMO ve Genişletilmiş CRMO modeli, Parçalı Regresyon CRMO, Kübik Spline CRMO uygulanmıştır. Yine OTV sağlanmadığından dolayı Parametrik Modellerden Hızlandırılmış Başarısızlık Süresi Modellerine (HBSM) başvurulmuştur. Model performanslarını karşılaştırmak için AIC metriği kullanılmıştır. Yarı Parametik Modellerden en iyi performans sergileyen model Tabakalandırılmış CRMO ve Parametrik Modellerden en iyi sonuç Log-Lojistik Modeli olarak karşımıza çıkmıştır. Kullanılan bu modeller ve yöntemler sayesinde müşteri ne zaman kayıp yaşar tahminlenmiş olup müşterilerin, müşterilik sürelerine etki eden faktörler belirlenmiştir.

**Anahtar Kelimeler:** Müşteri kayıp analizi, sağkalım analizi, telekom

# ABSTRACT

## Applying Survival Analysis Models and Methods to Telecom Churn Data

Melik MASARİFOĞLU

Department of Statistics

Doctor of Philosophy Thesis

Supervisor: Prof. Dr. Ali Hakan BÜYÜKLÜ

For the customer-centric businesses such as telecom, where competetion is fierce, customer churn is a disaster due to acquisition cost, reduced cash flow and payments. In this market, customer retention is easier, more achievable and less costly. Long-term, win-win relationships with customers are vital for mobile operators to survive in today's stiff competitive environment. Profiling the customer that could be churned in advance is a great advantage in terms of retaining the customer. In this study, it is aimed to minimize customer loss by profiling customers accurately and effectively. Survival Analysis methods that is used in different disciplines, especially in medicines, are used in this study. A total of nineteen models were used under three main topics: Non-Parametric, Semi-Parametric and Parametric Models. These models are: Non Parametric Models Kaplan-Meier, Semi-Parametric Cox Regression Model (CRMO) and since proportional hazard assumpiton (PHA) is not satisfied, Stratified CRMO, Extended CRMO, Piecewise Regression CRMO and Cubic Spline CRMO are applied. In Parametric Models: Accelerated Failure Time Model (AFTM) are also applied. In order to compare models, AIC metric is used. Among CRMO models, Stratified CRMO and among Parametric Models, Log-Logistic AFTM performs the best. The

aim is to answer the question of when the customer will churn and to identify the factors of affecting the customer's survival time in the mobile operator.

**Keywords:** Churn analysis, survival analysis, telecom

# 1
# GİRİŞ

## 1.1 Literatür Özeti

Liberal marketlerde bir mobil operatörün müşteri kaybının yıllık ortalamada yaklaşık %30 civarında oluşu ve bu kaybın minimize edilme hedefi, mobil operatörlerde proaktif yaklaşımları geliştirme isteğini artırmıştır[1]. Bunun sonucu akedemik çalışmalara yansımış olup, müşteri kayıp analizi ile ilgili birçok çalışma yapılmış ve hala yapılmaktadır. Yüksek kayıp oranını düşürmekte bir hedef haline gelmiştir. Bu çalışmaların çoğunda mobil operatör kayıp veri setlerine geleneksel sınıflandırma modelleri uygulanmıştır. Tahmin edilen değişken olan sonuç değişkeni iki sınıflı olup, müşterinin kayıp yaşayıp yaşamayacağıdır. Geleneksel makine öğrenimi metotlarından derin öğrenmeye kadar birçok değişik sınıflandırma modelleri ve yöntemleri kullanılmıştır [2,3,4,5,6,7]. Bu çalışmalar Sağ Kalım Analizi(SKA)'nin aksine müşterinin ne zaman kayıp yaşayacağı ve müşterilik ömrüne etki eden faktörlerle ilgili bir tahmin vermemektedir.

Telekom sektöründe SKA'ya başvurulan sınırlı sayıda çalışmalar mevcuttur. [8] ABD'de aylık müşteri kaybının %5'in olduğu bir mobil operatör verilerine Parametrik Olmayan Kaplan-Meier Modeli (KM) ve Yarı Parametrik Cox Regresyon Modeli (CRMO) uygulamıştır. Abonelerin ne zaman kayıp yaşayacağı tahmin edilmiş olup, abone kayıp süresi bilindiği için çok önceden kayıp yaşanacak müşteri için o müşteriye uygun pazarlama teşvikleri kullanılarak müşterinin operatörde tutulmasını hedeflemişlerdir.

[9] Hızlandırılmış Başarısızlık Süresi Modeli (HBSM) için SAS(PROC LIFEREG) kullanılarak müşteri sağ kalım olasılığı hesaplanmıştır. Bu hesaplanan değerler, müşteri değeri metriği hesaplamak için kullanılmıştır. CRMO, müşteri kaybı değişkenine etki eden açıklayıcı değişken sayısını azaltmak için kullanılmıştır. [10] Venezuella'dan bir operatör verilerine KM metodunu uygulamış ve müşteri değer metriğini hesaplamıştır. [11] Endonezya'da mobil operatör veri setine CRMO model varsayımı kontrolü yapılarak CRMO uygulanmıştır. [12]

çalışmasında, çalışmamızda uygulanan kayıp [13] veri setini uygulamış olup CRMO ve HBSM(Weibull, Log-Normal) dağılımları modellenmiştir. Çalışmamız Log-Lojistik dağılımınıda içermekte olup, veri setini modele hazırlarken yapılan ön hazırlık (korelasyon elemesi-Variance Inflation Factor) ve parametre stabilizasyonu için kullanılan cezalandırıcı Panelizer Katsayısı, çalışmamızı bu çalışmadan ayırmış ve çalışmamızda uygulanan Tabakalandırılmış CRMO, Genişletilmiş CRMO, CRMO Parçalı Regresyon ve CRMO Kübik Spline modellerine bu çalışmada başvurulmamıştır. Çalışmamızdaki model performansları, etkin ön hazırlık aşamaları ve başvurulan yeni modeller sayesinde bu çalışmadan çok daha iyi sonuçlar vermiştir.

## 1.2 Tezin Amacı

Telekom sektöründe müşteri kaybı, müşterinin mevcut operatöründe yürürlükte olan sözleşmesini feshedip, rakip operatöre geçmesidir. Çok fazla operatörün oluşu, müşteri merkezli ürünlerin, kampanyaların ve tarifelerin fazlalığı ve çeşitliliği rekabeti yoğunlaştırmış ve kayıp yaşanan müşterilerin sayısını artırmıştır. Bu rekabetin bir yansıması olarak, kayıp müşterinin yerini doldurmak, mevcut müşteriyi elde tutmaktan ortalamada beş-altı kat daha maliyetli bir hale gelmiştir.Bu maliyeti artıran ana sebep, giden müşterinin yerini doldurmak için rakip operatörlerdeki müşteriye verilen maddi teşvikler ve reklam maliyetleridir[14,15].

Telekom sektöründeki rekabet ortamı diğer sektörlerden daha fazladır. Tablo 1.1 ‘de 2011 yılında Amerikan firmaları için değişik endüstrilerde dolar bazında ortalamada yeni müşteri kazanım maliyeti verilmiştir[16]. Telekom sektöründe bir firma olan Sprint mobil operatörü diğer sektörlerdeki firmalara göre en yüksek maliyete sahiptir.

**Tablo 1.1** Müşteri kazanım maliyet tablosu

| Sektör | Firma | Maliyet |
|---|---|---|
| Seyahat | Priceline.com | $7 |

**Tablo 1.1** Müşteri kazanım maliyet tablosu (devamı)

| | | |
|---|---|---|
| Telekom | Sprint | $315 |
| Perakende | Barnesandnoble.com | $10 |
| Finans | TD Waterhouse | $175 |

Bu yüksek maliyetten dolayı, mobil operatörler yeni müşteri bulmaktan ziyade, mevcut müşterileri elde tutmaya odaklanmışlardır.

Mevcut müşteriyi elde tutmanın avantajları, sadece maliyetle sınırlı kalmayıp, müşteri ile operatör arasında uzun dönemli bir bağ kurulmasına ve müşteriye daha fazla ürün ve servis satışı sağlanmasına olanak tanır. Bu durum mobil operatöre nakit akışı ve karlılık olarak döner. Mobil operatörle kurulan bu bağ sayesinde müşteriler uzun süreler elde tutulur ve elde tutulan bu müşterilerin sadakati artar. Sadık müşteriler, mobil operatörün yeni fiyatlandırmalarına karşı daha az hassas davranırlar ve mobil operatörle alakalı pozitif konuşarak yeni müşteriler kazandırırlar. Sadık olmayan, kısa sürede mobil operatörü terk eden kayıp müşterilerle, potansiyel ürün satışlarından gelecek nakit akışı kaybedilir ve kayıp müşteri, memnuniyetisizliğini yayarak mobil operatörün imajını zedeleyebilir[17,18,19].

Müşteri kaybının telafisi zor ve müşteriyi elde tutmanın daha değerli olduğu doymuş markette, müşteri kayıp modelleri işte bu noktada devreye girer. Kayıp yaşanacak riskli müşterilerin önceden tahminlenmesi ve kayba etki eden faktörlerin bilinmesi mobil operatörler için mevcut müşteriyi elde tutma adına büyük avantajdır. Doğruluğu yüksek modeller sayesinde, müşteriyi elde tutmakta %5'lik bir gelişim, %25 ile %85 arasında mobil operatörlere karlılık olarak dönmektedir[20].

## 1.3 Hipotez

Mobil operatörlerin efektif müşteri elde tutma stratejileri geliştirebilmeleri ve müşteri kaybını minimize edebilmeleri için önceden hangi müşteri ne zaman ve

ne için kayıp yaşar sorularına cevap bulmaları gerekir. Bu sayede kampanyalarla etkili bir şekilde reaksiyon verip müşteri kaybının önüne geçebilirler. Çalışmamızda SKA ile bu sorulara cevap aranmış, müşteri ne zaman kayıp yaşar tahmin edilmiş ve müşterilik yaşam süresine hangi faktörlerin etkisi olduğu araştırılmıştır.

SKA'da Kaplan Meier Analizi(KM), CRMO ve HBSM uygulanmıştır. Bildiğimiz kadarıyla Telekom sektöründe bu üç modelin aynı anda uygulanması ve CRMO varsayımlarının sağlanmadığı durumlarda Tabakalandırılmış CRMO, Genişletilmiş CRMO, Parçalı CRMO, Kübik Spline CRMO ve HBSM modellerine başvurulması ve performanslarının araştırılması bir ilktir[21].

## 1.4 Tezin İçeriği

Çalışmamız şu şekilde organize edilmiştir. Bölüm 2'de SKA tanımı ve teorik bilgileri verilmiş olup, bölüm 3'de CRMO ve CRMO'nun ön şartı olan OTV tanıtılmış olup OTV sağlanmadığı durumlarda literatürde başvurulan modeller tanıtılmıştır. Bölüm 4'de Parametrik Regresyon Modeli olan HBSM tanıtılmış ve yapıları anlatılmıştır. Bölüm 5'de, SKA'da teorik bilgileri verilen modeller, Telekom veri setine uygulanmıştır.

Bölüm 6'de, ugulanan yeni yöntemlerin ve yaklaşımların bir değerlendirilmesi yapılmış, model performansları üzerine analizler yapılmıştır.

# 2
# SAĞKALIM ANALİZİ

## 2.1 Tanım

SKA, zaman-başarısızlık verilerini analiz eden, niceliğini belirten metotlar, modeller kümesidir. Zaman, başarısızlığa yani çalışmada tanımlanan olaya kadar ki geçen süredir. Araştırmacılar, veri bilimciler ve aktüerler, bir çok akademik çalışmada, araştırmada ve projede başarısızlık (olay) ne zaman gerçekleşir?" sorusuna cevap aramaktadırlar. Bu soruya cevap verecek teknikler bütünü SKA'dır.

SKA konseptinde olay, çalışmaya giren bireyin ölmesi, kanserin tekrarı, sporcunun ameliyattan sonra spora dönmesi, işsiz bir insanın iş bulması, elektronik parçaların bozulması, kredi başvurusunun onaylanmaması, şirketlerin iflas etmesi gibi örneklerle tanımlanabilir. Sosyal bilimlerde olay; boşanma, akademik promosyon gibi bireyin, hal ve statü değiştirmesidir

SKA hedefleri:

- Yaşam fonksiyonunu tahmin etmek ve yorumlamak
- Yaşam fonksiyonlarını karşılaştırmak
- Yaşam süresi ile bir veya birden fazla açıklayıcı değişken arasında ki ilişkiyi belirlemek

SKA karakteristiği:

- Sonuç değişkeni olan olayın oluşuna kadar ki geçen süre
- Gözlemi tamamlamamış yani çalışma süresince olayı yaşamamış bireylerin verilerinin analizi

SKA'nın çerçevesi:

- Olayın tanımlanması
- Çalışmanın zaman ölçeğinin tanımlanması (yıl, ay, hafta, gün…)

- Çalışma başlangıç ve bitiş zamanının çalışmadan önce tanımlanması

SKA, geçmiş başarısızlık analizi, tehlike analizi, geçiş analizi veya durağanlık analizi olacak şekilde değişik tanımları literatürde mevcuttur.

**SKA'nin geçmişi:**

SKA'nın geçmişte kökleri ölümlülük araştırmaları yapan John Garunt'a kadar dayanır. 1962'de yazdığı kitap "Natural and Political Observations on the Bills of Mortality" de Yaşam Tabloları konsepti tanıtılmıştır. SKA'de modern yaklaşım, II. Dünya Savaşı sırasında, elektronik askeri ekipmanların güvenilirliği testinde kullanılmıştır. Savaştan sonrada diğer disiplinlerde de kullanımı hızlı bir şekilde artmıştır.

SKA'nın en prestijli iki makalesi Kaplan Meier [22] ve Cox [23]'dir. Makalelerde, SKA'de en popüler modellerinden olan KM ve CRMO tanıtılmıştır.

## 2.2 Model

SKA'da kullanılan bazı temel kavramların ve fonksiyonların tanımları aşağıdaki şekildedir.

### 2.2.1 Olasılık Yoğunluk Fonksiyonu

Olasılık yoğunluk fonksiyonu f(t) şeklinde gösterilir ve T yaşam süresi sürekli rastlantı değişkeni için aşağıdaki şekildedir.

$$f(t) = \lim_{\Delta \to 0} \frac{P(t < T \leq t + \Delta)}{\Delta} \tag{2.1}$$

### 2.2.2 Yaşam Fonksiyonu

Yaşam fonksiyonu S(t), T yaşam süresi sürekli olan rastlantı değişkeni için, t süresinden sonra hayatta kalma olasılığını verir ve aşağıdaki şekilde tanımlanır.

$$S(t) = \Pr\{T \geq t\} \tag{2.2}$$

$$= 1 - F(t) \tag{2.3}$$

$$= \int_t^{\infty} f(x)dx \tag{2.4}$$

Çalışma başlangıç ve bitiş süreleri önceden belirlendiğinden çalışma başında bütün bireyler hayattadır. T=0'da $S(0) = 1$ şeklinde ifade edilir ve ispatı;

$$S(0) = \Pr\{T \geq 0\} \tag{2.5}$$

$$= 1 - F(0) \tag{2.6}$$

$$= \int_0^\infty f(t)dt \tag{2.7}$$

$$= 1 \tag{2.8}$$

Çalışma süresi sonsuza kadar gittiğinde yani t=∞ olduğunda bütün bireyler olayı yaşar. t=∞'da $S(\infty) = 1$ şeklinde ifade edilir ve ispatı;

$$S(\infty) = \Pr\{T \geq \infty\} \tag{2.9}$$

$$= 1 - F(\infty) \tag{2.10}$$

$$= 1 - 1 \tag{2.11}$$

$$= 0 \tag{2.12}$$

S(t) monoton azalan bir fonksiyondur ve S(t) aşağıdaki değerleri alır.

$$0 \leqslant S(t) \leqslant 1 \tag{2.13}$$

### 2.2.3 Tehlike Fonksiyonu

Tehlike fonksiyonu $\lambda(t)$, T yaşam süresi sürekli rastlantı değişkeni olmak üzere, bireyin t süresine kadar olayı yaşamadığı varsayılarak t anında tehlikeyi verir.

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\Pr\{t \leq T < t + \Delta t | T \geq t\}}{\Delta t} \quad , t \geq 0 \tag{2.14}$$

λ(t) bir olasılık değeri değildir. 2.14'de paydaki olasılık değeri, paydadaki $\Delta t$ ifadeye bölündüğü zaman, çıkan sonuç 1'i geçebilir ve üst sınırı yoktur. Bu durum olasılık aksiyomlarından birininin reddi olur.

2.14 denkleminin paydasında ki şartlı olasılık fonksiyonu, başarısızlığın olma ihtimali, [t; t+$\Delta t$) aralığında o ana kadar ki başarısızlığın olmama ihtimaline

oranıdır. Paydada ki ifade ise bu zaman aralığının genişliğidir. Buradaki birim aralık 0'a yaklaştıkça bu oran ani tehlike oranını verir.

### 2.2.4 Birikimli Tehlike Fonksiyonu

Birikimli tehlike fonksiyonu $\Lambda(t)$, t süresine kadar tehlike fonksiyonunun birikimli fonksiyonudur.

$$\Lambda(t) = \int_0^t \lambda\,(x)dx \tag{2.15}$$

### 2.2.5 Yaşam Fonksiyonu ile Tehlike Fonskiyonu Arasındaki İlişki

S(t) biliniyorsa, $\lambda(t)$ bulanabilir veya $\lambda(t)$ biliniyorsa S(t) bulunabilir. Verilen bir t değeri için S(t) yüksekse, $\lambda(t)$ düşüktür. Aralarındaki ilişki aşağıdaki şekilde bulunur. 2.14 denklemi şartlı olasılık formulü kullanılarak yeniden yazılırsa;

$$= \lim_{\Delta t\to 0}\{\frac{Pr(t<T\leq t+\Delta t, T>t)}{\Delta t}\cdot\frac{1}{P(T>t)}\} \tag{2.16}$$

$$= \lim_{\Delta t\to 0}\frac{Pr(t<T\leq t+\Delta t)}{\Delta t}\cdot\frac{1}{P(T>t)} \tag{2.17}$$

$$= \frac{1}{P(T>t)}\cdot\lim_{\Delta t\to 0}\frac{P(t<T\leq t+\Delta t)}{\Delta t} \tag{2.18}$$

$$= \frac{1}{P(T>t)}\cdot\lim_{\Delta t\to 0}\frac{F(t+\Delta t)-F(t)}{\Delta t} \tag{2.19}$$

$$= \frac{f(t)}{S(t)} \tag{2.20}$$

$\lambda(t)$, f(t) ve S(t) cinsinden yazılmış olur. 2.20 denkleminde f(t)'yi, S(t) cinsinden yazmak için aşağıdaki adımlar uygulansın:

$$S(t) = 1 - F(t) \tag{2.21}$$

Her iki tarafın türevi alınırsa:

$$S'(t) = -F'(t) \tag{2.22}$$

$$= -\mathrm{f(t)} \tag{2.23}$$

$$\mathrm{F(t)} = -S'(t) \tag{2.24}$$

$$\lambda(t) = -\frac{S'(t)}{S(t)} \tag{2.25}$$

$\lambda(t)$, S(t) cinsinden yazılır. 2.25 denklemi aşağıdaki şekilde ifade edilebilir:

$$\lambda(t) = -\frac{d}{dt} log S(t) \tag{2.26}$$

2.26 denkleminde her iki tarafın integrali 0 dan t 'ye kadar alınıp ve S(0)=1 denklemi kullanılırsa , 2.26 denkleminin çözümü aşağıdaki ifadeyi verir.

$$S(t) = \exp\left\{-\int_0^t \lambda(x)dx\right\} \tag{2.27}$$

Parantez içindeki ifade birikimli tehlike fonksiyonudur. Denklem 2.24 kullanılıp, f(t), λ(t) cinsinden yazılabilir,

$$f(t) = -S'(t) \tag{2.28}$$

$$= \lambda(t) \exp\left\{-\int_0^t \lambda(x)dx\right\} \tag{2.29}$$

örnek olarak, verilen bir t için;

$$\lambda(t) = \lambda \tag{2.30}$$

sabit bir değer olsun. denklem 2.27'den

$$S(t) = \exp\{-\lambda t\} \tag{2.31}$$

olacak şekilde ifade edilir. Denklem 2.24 kullanılırsa;

$$f(t) = \lambda \exp\{-\lambda t\} \tag{2.32}$$

Üstel dağılımın f(t)'sini verir.

## 2.3 Durdurma

Teoride, en doğru model tahminlemesi için çalışmaya giren her bireyin olaya kadar ki geçen süresini bilmek isteriz. Pratikte bu mümkün değildir. Bazı yaşam süreleri bir sınırın altında veya üstünde kalabilir. Çalışma süresi belirli bir zaman diliminde sınırlandırıldığı için çalışmadaki bütün bireyler olayı yaşayamaz. Bazı bireyler olayı yaşamadan çalışmayı tamamlayabilirler. Bazı bireylerde çalışmayı olayla ilişkili olmayan nedenlerden dolayı terk edebilirler, kaybolabilirler veya ayrılabilirler. Bu bireylerin gerçekte ne zaman olayı yaşayacakları bilinmemektedir. Bu tip durdurmaya bilgilendirici-olmayan durdurma denir.

Yanlılıktan kaçınmak için durdurma varsayımı rastgele ve bilgilendirici-olmayan olmasıdır[24].

Çalışmayı olayla alakalı bir sebepten dolayı terk eden, ayrılan veya kaybolan bireylerin verileri (bilgilendirici durdurma) çalışmada tutulup tutulmayacağı dikkatle incelenmelidir. Çalışmada tutulacaksa eksik veri analizi yapılabilir veya çalışmadan tamamen çıkarılabilir [25].

Sağdan, Soldan ve Aralık durdurma olarak üç çeşittir. Çalışmamızda veri seti sağdan durdurulmuş bireyler içerdiğinden, sağdan durdurmanın tanımı ve örneği aşağıdaki şekildedir.

### 2.3.1 Sağdan Durdurma

Matematiksel olarak tanımını;

i=1,2,…n birey sayısı olamk üzere:

- i bireyi için $T_i$ yaşam süresi
- i bireyi için $C_i$ durdurma zamanı
- $\delta_i$, i bireyin durdurma göstergesi olmak üzere aşağıdaki şekilde tanımlanır

$$\delta_i = \begin{cases} 1, \text{olay} & gerçekleşmiş\ ise\ (\mathrm{T}_i \leq C_i) \\ 0, \text{olay} & gerçekleşmemiş\ ise (\mathrm{T}_i > C_i) \end{cases} \quad (2.33)$$

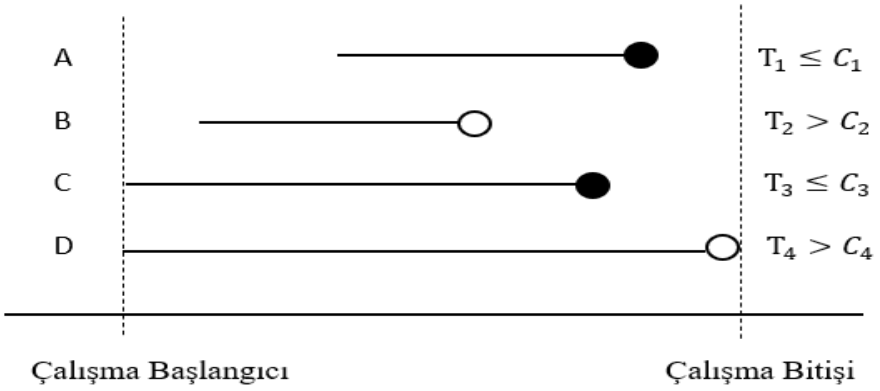
- Gözlenen yaşam süresi herbir i bireyi için $Y_i = \min(T_i, C_i)$ olacaktır

Başka bir deyişle; sağdan durdurmada, gözlenen yaşam süreleri her zaman gerçek yaşam sürelerinden küçük veya eşittir.

Sağdan durdurma, Tip 1 ve Tip 2 olacak şekilde ikiye ayrılır. Tip 1 durdurma, çalışma takvimi önceden planlanmış ve belirlenmiştir. Klinik araştırmalar bu tipin en güzel örneğidir. Klinik çalışmalarda önceden belirlenen başlangıç ve bitiş tarihleri arasında çalışmaya giren bireyler gözlenir. Bazı bireyler;

- Olayı belirlenen zaman diliminde yaşamamışlarsa,
- Çalışma bitmeden önce farklı sebeplerden dolayı çalışmadan ayrılmışlar, kaybolmuşlarsa,

- Başka bir olayın oluşundan dolayı çalışmada tanımlanan olayı yaşamamışlarsa bu bireyler sağdan durdurulmuş olur.




**Şekil 2.1** Sağdan durdurma örnek

Şekil 2.1'de sağdan durdurma örnek verileri mevcuttur. A bireyinin durdurma göstergesi 1 olup olayı yaşamış, durdurma zamanı en az yaşam süresi kadar veya daha büyüktür. B bireyinin durdurma göstergesi 0 olup durdurulmuş ve yaşam süresi durdurma zamanından büyüktür. C bireyinin durdurma göstergesi 1 olup olayı yaşamış, durdurma zamanı yaşam zamanından büyüktür.

Tip II durdurma, daha çok endüstriyel çalışmalarda karşımıza çıkar. Çalışmaya bütün bireyler (elektronik parçalar vb.) aynı anda başlar, önceden belirlenen belirli sayıda birey başarısızlığa ulaştığı anda çalışma bitirilir. Örnek olarak, 100 transistor çalışmaya giriyor ve 50 tanesi başarısız olduğunda çalışma bitiriliyorsa, bu çalışma Tip II durdurmaya bir örnektir.

## 2.4 Parametrik Olmayan Yöntemler (Kaplan-Meier)

Kaplan Meier(KM) yöntemi, herbir başarısızlık zaman aralıkları t'yi dikkate alarak, t ye kadar olan sağ kalım olasılığını veririr. KM, çarpım-limit olarak da adlandırılır ve herhangi bir parametre içermediğinden parametrik olmayan bir metottur. Formulü aşağıdaki şekildedir;

$$\widehat{S}(t) = \prod_{t_{(i)<t}} \left\{1 - \frac{m_{(i)}}{r_{(i)}}\right\} \qquad (2.34)$$

Denklemde sırasıyla başarısızlık zamanları sıralı olarak $t_{(1)} < t_{(2)} < t_{(3)} < \cdots < t_{(k)}$ , $t_{(i)}$ zamanına karşılık gelen toplam başarısız birey sayısı $m_{(i)}$ olmak üzere, zaman t'de toplam birey sayısı ise risk seti $R_{(i)}$ olarak adlandırılır.

Varyansıda Greenwood formulü ile hesplanır;

$$\hat{\sigma}^2(t) = \widehat{S(t)}^2 \sum_{t_{(i)}<t} \frac{m_{(i)}}{r_i(r_i - m_i)} \tag{2.35}$$

Büyük örneklemlerde t başarısızlık zamanı standart normal dağılım ile dağıldığı varsayımı kabul edildiğinden, 100(1-α) güven aralığı aşağıdaki şekilde hesaplanır.

$$\widehat{S}(t) \pm z_{1-\frac{\alpha}{2}}\hat{\sigma}(t) \tag{2.36}$$

Tablo 2.1, KM için bir örnektir. 100 bireyle çalışmaya başlanmış olup, yirminci ayda 10 birey çalışmada tanımlanan olayı yaşamış, 5 bireyde aşağıdaki ihtimallerden birini yaşamış olabilir;

- Mevcut çalışma için tanımlanan olaydan bağımsız olayı yaşamışlar
- Çalışmayı terk etmişler
- Taşınmış, telefon numarası veya adresi değişmiş ulaşılamaz hale gelmişlerdir

**Tablo 2.1** KM örnek

| Zaman (Ay) | Risk Adeti | Başarısızlık Adeti | Sağdan Durdurulmuş Adeti | Tehlike Fonksiyonu | Kümüle Tehlike Fonksiyonu | Sağ Kalım Olasılığı |
|---|---|---|---|---|---|---|
| 3 | 100 | 5 | 2 | $\frac{5}{100} = 0{,}05$ | 0,05 | 1-0,05 =0,95 |
| 10 | 93 | 3 | 1 | $\frac{3}{93} = 0{,}03$ | 0,08 | 0,95(1-0,03) =0,92 |
| 15 | 90 | 6 | 0 | $\frac{6}{90} = 0{,}06$ | 0,14 | 0,92(1-0,06) =0,86 |

**Tablo 2.1** KM örnek (devamı)

| | | | | | | |
|---|---|---|---|---|---|---|
| 20 | 84 | 10 | 5 | $\frac{10}{84} = 0{,}03$ | 0,26 | 0,86(1-0,12) =0,75 |

Tablo 2.1 için denklem 2.7'de kullanılacak olursa, çalışmaya giren herhangi bir bireyin 3 ay veya daha fazla sağ kalım olasılığı:

$$S(3) = P(T \geq 3) = \frac{95}{100} \tag{2.37}$$

10 ay veya daha fazla sağ kalım olasılığı:

$$S(10) = P(T \geq 10) = \frac{92}{100} \tag{2.38}$$

20 ay veya daha fazla sağ kalım olasılığı:

$$S(20) = P(T \geq 20) = \frac{75}{100} \tag{2.39}$$

15. ay ve 20. ay arasındaki kümüle fonksiyonundaki azami artışın fazla oluşu, çok fazla başarısızlığın bu aylar arasında gerçekleştiğini gösterir. Tablo 2.2 'de standart hata ve %95 güven aralığı hesaplamaları yapılmıştır.

**Tablo 2.2** Hata terimi ve güven aralığı

| **Zaman (Ay)** | $\frac{m_i}{r_i(r_i - m_i)}$ | $\sum \frac{m_i}{r_i(r_i - m_i)}$ | $\widehat{S(t)} \sqrt{\sum \frac{m_i}{r_i(r_i - m_i)}}$ | **%95 Güven Aralığı** | |
|---|---|---|---|---|---|
| 0 | | | 0 | | |
| 3 | 0,00052 | 0,00052 | 0,0218 | 0,9 | 0,992 |
| 10 | 0,00035 | 0,00087 | 0,0283 | 0,89 | 1,005 |
| 15 | 0,0008 | 0,00167 | 0,0392 | 0,87 | 1,026 |

**Tablo 2.2** Hata terimi ve güven aralığı (devamı)

| | | | | | |
|---|---|---|---|---|---|
| 20 | 0,0016 | 0,00327 | 0,0548 | 0,84 | 1,057 |

KM'de birden fazla grup karşılaştırılabilir, grafikleri çizilebilir. Log Rank Testi (LRT) uygulanarak eğriler arasında istatistiksel fark olup olmadığı anlaşılabilir. LRT parametrik olmayan bir testtir ve aşağıdaki özelliklere sahiptir.

- Ki Kare testidir
- Kaplan-Meier sağ kalım eğrilerinin karşılaştırılmasında kullanılır
- Beklenen rakamlarla gözlenen rakamların karşılaştırılmasıdır
- Kategoriler, başarısızlık zaman aralıklarının küçükten büyüğe sıralanmasıyla oluşturulur

Log Rank test hipotezi aşağıdaki gibidir.

$$H_O: S^1(t) = S^0(t) \quad \text{bütün } t \leq \tau$$
$$H_1: S^1(t) \neq S^0(t) \quad \text{bazı } t \leq \tau \tag{2.40}$$

τ iki grupta en az bir deneğin risk altında olduğu en geniş zaman aralığıdır. Test istatistiğininin bir çok denklemsel formu mevcuttur. [26] göre Ki-Kare testi aşağıdaki şekilde hesaplanır.

$$\chi^2 = \sum \frac{\left(\sum O_{jt} - \sum E_{jt}\right)^2}{\sum E_{jt}} \tag{2.41}$$

Denklemde sırasıyla, başarısızlık süreleri t, j grubunda $N_{jt}$ toplam birey sayısı, $\sum O_{jt}$ olayı yaşayan toplam birey sayısı, $\sum E_{jt}$ olayı yaşayan birey sayısının beklenen değeridir. Tablo 2.3'de LRT'nin nasıl hesaplandığına dair bir örnek vardır.

**Tablo 2.3** LRT bir örnek

| Ay | $N_{1t}$ | $N_{2t}$ | $N_t$ | $O_{1t}$ | $O_{2t}$ | $O_t$ | $E_{1t} = N_{1t} \times \frac{O_t}{N_t}$ | $E_{2t} = N_{2t} \times \frac{O_t}{N_t}$ |
|---|---|---|---|---|---|---|---|---|
| 8 | 40 | 30 | 70 | 2 | 1 | 3 | 1,71 | 1,29 |

**Tablo 2.3** LRT bir örnek (devamı)

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 12 | 37 | 28 | 65 | 1 | 0 | 1 | 0,57 | 0,43 |
| 14 | 33 | 23 | 56 | 0 | 2 | 2 | 1,18 | 0,82 |
| 21 | 32 | 22 | 54 | 1 | 1 | 2 | 1,19 | 0,81 |
| 26 | 32 | 21 | 53 | 0 | 0 | 0 | 0,0 | 0,0 |
| 27 | 27 | 19 | 46 | 2 | 1 | 3 | 1,76 | 1,24 |
| 28 | 22 | 18 | 40 | 3 | 2 | 5 | 2,75 | 2,25 |
| 33 | 21 | 15 | 36 | 1 | 4 | 5 | 2,92 | 2,08 |
| 41 | 20 | 11 | 31 | 0 | 0 | 0 | 0,0 | 0,0 |
| | | | Toplam | 10 | 11 | | 12,07 | 8,93 |

Tablo 2.3 ‘den test istatistiği

$$\chi^2 = \frac{(10 - 12{,}07)^2}{12{,}07} + \frac{(11 - 8{,}93)^2}{8{,}93} \tag{2.42}$$

$$= 0{,}83 \tag{2.43}$$

şeklinde hesaplanır.

# 3
# YARI PARAMETRİK COX REGRESYON MODELİ

## 3.1 Model Tanımı

Bazı değişkenlerin sonuç değişkeni olan yaşam süresine veya yaşam fonksiyonuna etkisi ölçülmek isteniyorsa Yarı Parametrik Cox Regresyon Modeli (CRMO)'ne başvurulabilir. CRMO, sonuç değişkeni olan yaşam süresi ile açıklayıcı değişkenlerin arasındaki etki düzeyini belirleyebilecek ve ölçebilecek kabiliyette olan etkili, popüler ve karmaşık bir modeldir. CRMO çalışma süresi boyunca bireylerin başarısızlık riskini ölçer.

CRMO'da, tehlike fonksiyonu $\lambda(t|X)$ şeklinde gösterilir ve aşağıdaki şekilde formülize edilir.

$$\lambda(t|X) = \lambda_0(t)c(\beta X) \tag{3.1}$$

X açıklayıcı değişken ve β regresyon parametreleri vektörüdür. 3.1 denkleminde gereklilik $\lambda(t|X) > 0$ olmasıdır. Bu durumda $c(\beta^t X)$ fonksiyonu için Cox, üstel fonksiyonunu seçer[23].

$$c(\beta X) = e^{\beta X} \tag{3.2}$$

Açıklayıcı değişken sayısı p olmak üzere, denklem 3.2'de parametre katsayı vektörü aşağıdaki şekilde gösterilebilir.

$$\beta = (\beta_1 \beta_2 \beta_3 \dots \beta_p) \tag{3.3}$$

Açıklayıcı değişken vektörü:

$$X = (X_1 X_2 X_3 \dots X_p) \tag{3.4}$$

Bu durumda $\beta X$

$$\beta X = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_p X_p = \sum_{i=1}^{p} \beta_i X_i \tag{3.5}$$

şeklinde ifade edilir. Denklem 3.5, denklem 3.1 'de yerine yazılacak olursa:

$$\lambda(t|X) = \lambda_0(t)e^{\beta X} = \lambda_0(t)e^{\sum_{i=1}^{p}\beta_i X_i} \quad (3.6)$$

tehlike fonksiyonunun formülü ortaya çıkar. $e^{\sum_{i=1}^{p}\beta_i X_i}$ fonksiyonu risk skoru olarak da adlandırılır ve $\lambda_0(t)$ temel tehlike fonksiyonudur. Temel tehlike fonksiyonu

$$\lambda(t|X = 0) = \lambda_0(t)e^{\beta_1 * 0 + \beta_2 * 0 + \cdots + \beta_p * 0} \quad (3.7)$$

$$= \lambda_0(t) \quad (3.8)$$

bütün değişken değerlerinin 0 olduğundaki tehlike değeridir. Temel tehlike fonksiyonu üzerinde herhangi bir dağılım varsayımı bulunmadığı için tahminlenmez ve hesaplanmaz , bu yüzden CRMO yarı parametrik bir modeldir.

Sir David Cox'ın literatüre kazandırdığı bu model, SKA'da çok başarılı, yaygın ve popülerdir [23]. 1972 yılında yazdığı makalesi 2009'a kadar 1000'den fazla kez referans gösterilerek, istatistikte en fazla referans gösterilen makale olmuştur [27].

## 3.2 Tehlike Oranı

CRMO neden oransal tehlikeler modeli dendiği bu bölümde daha net anlaşılacaktır. Sadece herhangi bir X değişkeninde farklılaşan iki farklı birey için sırasıyla değişken vektörleri $X$, $X^*$ olmak üzere tehlike oranı aşağıdaki şekilde tanımlanır.

$$\text{Tehlike oranı} = \frac{\lambda_0(t)e^{\beta X}}{\lambda_0(t)e^{\beta X^*}} \quad (3.9)$$

$$= \frac{\lambda_0(t)e^{\sum_{i=1}^{p}\beta_i X_i}}{\lambda_0(t)e^{\sum_{i=1}^{p}\beta_i X_i^*}} \quad (3.10)$$

$$= e^{\sum_{i=1}^{p}\beta_i (X_i - X_i^*)} \quad (3.11)$$

P adet açıklayıcı değişken için, örnek olarak, sadece k değişkeninde farklılaşan iki birey için tehlike oranı aşağıda ki şekildedir.

$$\frac{\lambda(t|X)}{\lambda(t|X^*)} = e^{\beta_k} = \varphi_k \text{ , tüm } t \geq 0 \quad (3.12)$$

Tehlike oranı değerine göre değişkenin tehlikeye etkisi olup olmadığı tablo 3.1'deki gibi yorumlanabilir.

**Tablo 3.1** Tehlike oranı değeri ve etkisi

| Tehlike Oranı Değeri | Tehlike Etkisi |
|---|---|
| $\varphi_k < 1$ | Tehlike Azalır |
| $\varphi_k = 1$ | Tehlike Etkisi Yoktur |
| $\varphi_k > 1$ | Tehlike Artar |

## 3.3 CRMO Kısmi Olabilirlik Fonksiyonu

Bilinmeyen açıklayıcı değişken katsayıları vektörü $\beta$'nın tahmin edilebilmesi için Kısmi Olabilirlik Fonksiyonu (KOF) kullanılır. Sırasıyla KOF'ta kullanılan terimler tanımlanacak olursa; n tane sıralı başarısızlık zamanı $t_1 < t_2 < \cdots < t_k$, $t_i$ zamanında henüz olayı yaşamayan bireylerin oluşturduğu küme $R_{t_i}$(risk kümesi), denklem 2.33'de tanımladığımız başarısızlık gösterge fonksiyonu $\delta_i$ olacak şekilde, durdurulmuş veri içeren bir örneklem için olabilirlik fonksiyonu aşağıdaki şekildedir.

$$L_i(\beta) = f_i(t_i) = S_i(t_i)\lambda_i(t_i) \tag{3.13}$$

Durdurulmuş bireyler olabilirlik fonksiyonuna katılmadıkları için denkleme $\delta_i$ eklenir. Birey sayısı i=1,...n olacak şekilde çarpılırsa

$$L(\beta) = \prod_{i=1}^{n} \lambda_i(t_i)^{\delta_i} S_i(t_i) \tag{3.14}$$

çıkan denklemi $\left[\sum_{j\epsilon R(t_i)} \lambda_J(t_i)\right]^{\delta_i}$ terimi ile çarpılıp bölünürse

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{\lambda_i(t_i)}{\sum_{j\epsilon R(t_i)} \lambda_J(t_i)}\right]^{\delta_i} \left[\sum_{j\epsilon R(t_i)} \lambda_J(t_i)\right]^{\delta_i} S_i(t_i) \tag{3.15}$$

elde edilir.[25] önerisi ile 3.15 denklemindeki çarpanlardan ilk terim, katsayılar vektörü $\beta$ ile ilgili hemen hemen bütün bilgiyi taşırken, ikinci ve üçüncü terim temel tehlike fonksiyonu ile ilgili bilgiler taşır.

Bu bilgiler ışığında sadece ihtiyacımız olan ilk terimi denklemde tutacak olursak, $L(\beta)$ aşağıdaki denkleme dönüşür.

$$L(\beta) = \prod_{i=1}^{n} \left[ \frac{\lambda_i(t_i)}{\sum_{j \epsilon R(t_i)} \lambda_j(t_i)} \right]^{\delta_i} \tag{3.16}$$

Aşağıdaki ifade 3.16 dekleminde yerine yazılırsa;

$$\lambda(t|X) = \lambda_0(t) e^{\beta^t X} \tag{3.17}$$

3.17 denklemi aşağıdaki şekle dönüşür.

$$= \prod_{i=1}^{n} \left[ \frac{\lambda_0(t_i) e^{\beta^t X_i}}{\sum_{j \epsilon R(t_i)} \lambda_0(t_i) e^{\beta^t X_j}} \right]^{\delta_i} \tag{3.18}$$

$\lambda_0(t_i)$ sadeleşmesi ile

$$= \prod_{i=1}^{n} \left[ \frac{e^{\beta^t X_i}}{\sum_{j \epsilon R(t_i)} e^{\beta^t X_j}} \right]^{\delta_i} \tag{3.19}$$

elde edilir. 3.19 denklemine durdurulmuş bireyler dahil edilmediğinden 3.19 denklemine Kısmi Olabilirlik Fonskiyonu (KOF) denir.Temel tehlike fonksiyonuna bağlı olmayan bir deklemdir. KOF'un logaritmik formu aşağıdaki şekli alır.

$$l(\beta) = \log \prod_{i=1}^{n} \left[ \frac{e^{\beta^t X_i}}{\sum_{j \epsilon R(t_i)} e^{\beta^t X_j}} \right]^{\delta_i} \tag{3.20}$$

$$= \sum_{i=1}^{n} \delta_i \left[ \beta^t X_i - \log \sum_{j \epsilon R(t_i)} e^{\beta^t X_j} \right] \tag{3.21}$$

$$= \sum_{i=1}^{n} l_i(\beta) \tag{3.22}$$

Denklemin payı durdurulmuş bireyleri içermezken, denklemin paydası durdurulmuş ve durdurulmamış bireyleri içerir. $L_i$, i bireyinin KOF'a katkısıdır. 3.22 denkleminin çözülmesi için maksimize edilmesi gerekmektedir. Maksimize edilmesi için $\beta$ 'ya göre türevi alındığında aşağıdaki KOF skor fonksiyonunu elde edilir.

$$U(\beta) = \frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n} \delta_i \left[ X_i - \frac{\sum_{j \epsilon R(t_i)} X_j e^{\beta^t X_j}}{\sum_{j \epsilon R(t_i)} e^{\beta^t X_j}} \right] \tag{3.23}$$

Değişken katsayıları vektörü $\beta$'nın tahminleyici vektörü, aşağıdaki sistemin nümerik bir metot olan Newton-Raphson yöntemi ile çözülmesiyle ortaya çıkar.

$$U(\beta) = 0 \tag{3.24}$$

Değişkenler arası yüksek korelasyon yoksa metot yakınsar ve çözüme ulaşır [28].

KOF örneği tablo 3.2'da verilmiştir. T sağ kalım süresi, X'de sağ kalım süresine etkisi olan bir değişken için,

**Tablo 3.2** KOF için örnek sağkalım tablosu

| Birey No | T | Olay | X |
|---|---|---|---|
| 1 | 2 | 1 | 12 |
| 2 | 5 | 1 | 6 |
| 3 | 4 | 0 | 8 |
| 4 | 1 | 1 | 9 |

KOF'ta her bir bireyin KOF'a katkısı, tablo 3.3'deki gibi bulunur.

**Tablo 3.3** KOF tablosu

| Birey No | Ay | Risk Kümesi | KOF Katkısı $R(t_i)$ |
|---|---|---|---|
| 4 | 1 | {1,2,3,4} | $\frac{e^{9\beta}}{e^{9\beta}+e^{12\beta}+e^{8\beta}+e^{6\beta}}$ |
| 1 | 2 | {1,2,3} | $\frac{e^{12\beta}}{e^{12\beta}+e^{8\beta}+e^{6\beta}}$ |
| 3 | 4 | {2,3} | 1 |

**Tablo 3.3** KOF tablosu (devamı)

| | | | |
|---|---|---|---|
| 2 | 5 | {2} | $\frac{e^{6\beta}}{e^{6\beta}}$ |

KOF aşağıda ki şekilde bütün parçaların çarpılması ile elde edilir.

$$L(\beta) = \frac{e^{9\beta}}{e^{9\beta}+e^{12\beta}+e^{8\beta}+e^{6\beta}} \times \frac{e^{12\beta}}{e^{12\beta}+e^{8\beta}+e^{6\beta}} \times 1 \times 1 \qquad (3.25)$$

3.22 denkleminde varsayım; eş zamanlı yaşam süreleri (tied events) olmamasıdır. Birden fazla birey aynı anda olayı yaşıyorsa bu durumda veri seti eş zamanlı yaşam sürelerine sahiptir. Mükemmel sürekli bir ölçekte iki birey aynı anda olayı yaşamaz. Her bir olayın gerçekleşme zamanı farklı olmalıdır. Bu pratikte gerçekçi değildir. Çalışmalarda olayın gerçekleşme zamanı, kesikli bir şekilde ölçüldüğü için bu varsayım çiğnenir. Yaşam süresi genel olarak yıl, ay, hafta vb. ölçüldüğü için birden fazla birey aynı yıl, ay veya hafta içinde olayı yaşayabilir. Bu durumda birden fazla bireyin aynı başarısızlık zamanı t'de olayı yaşaması kaçınılmazdır. Sonuç olarak 3.23 denklemini, bağlı olaylar verilerine göre ayarlamak gerekir.

Aşağıda ki üç metot veride bağlı olaylar varsa kullanılabilir.

- Cox'ın Kesin Metodu
- Breslow Metodu
- Efron Metodu

Bu metotların tanımı ve KOF fonksiyonu aşağıdaki şekildedir.

### 3.3.1 Kesin Metodu

Olay zamanının sürekli oluşu, eş zamanlı olayların oluşu ve olay zamanının doğru ölçülememesinden kaynaklandığı varsayımı, bu metot için geçerlidir. Bu varsayımdan dolayı sıralanan başarısızlık zamanlarında, eş zamanların hangisinin önce olduğu kestirilemez. Bu durumda bütün ihtimaller hesaba katılmalıdır. KOF uygun bir şekilde revize edilerek yazılır [23].

$t_1 < t_2 < \cdots < t_n$ sıralı başarısızlık zamanları ve i=1,2,…,m birey sayısı için, $t_i$'da bütün olayların düzenlemeleri olan $d_i$ ve permütasyonu $d_i!$ olmak üzere, $Q_i$'ya

atanacak olursa. P, $Q_i$'nın bir elemanı ve $R_i$, $t_i$ zamanında hayatta olan veya risk altında olan bireylerin risk kümesi olsun ve

$$\mathrm{D}(t_i) = \{i_1, \dots, i_{d_i}\} \tag{3.26}$$

$$\mathrm{P} = \{p_1, \dots, p_{d_i}\} \tag{3.27}$$

$$\mathrm{R}\ (t_i;\mathrm{P};\ \mathrm{k}) = \mathrm{R}\ (t_i) - \{p_1, \dots, p_{\mathrm{k}-1}\} \tag{3.28}$$

$$x_{i+} = \sum_{j=1}^{d_i} x_j \tag{3.29}$$

şeklinde tanımlansınlar. Bu durumda KOF denklemi aşağıda ki şekilde revize edilir.

$$L(\beta) = \frac{1}{d_i!} e^{x_{i+}\beta} \sum_{P \in Q_i} \prod_{k=1}^{d_i} \left[ \sum_{j \in R(t_i;P;\ k)} e^{x_j \beta} \right]^{-1} \tag{3.30}$$

**Tablo 3.4** KOF örnek

| **Birey No** | **T** | **Olay** | **X** |
|---|---|---|---|
| 1 | 3 | 1 | 12 |
| 2 | 3 | 1 | 6 |
| 3 | 4 | 0 | 8 |
| 4 | 6 | 1 | 9 |
| 5 | 10 | 1 | 7 |
| 6 | 10 | 1 | 11 |

Tablo 3.4'de, birey no; çalışmaya katılan her bir bireyin id'si, T; ay-zaman ölçeğinde yaşam zamanı, olay değişkeni; olayın olup olmadığını belirten değişken ve X; yaşam zamanına etki eden bir değişken olmak üzere, Kesin Metodu

uygulanacak olursa, üç ve onuncu ayda iki birey eş zamanlı olayı yaşamışlardır. Hangisinin önce başarısız olduğu bilinmemektedir. Bu durumda birinin başarısız olma ihtimali:

$$\text{P (Birey No 1'in başarısız olması)} = \frac{1}{Permütasyon(2,1)} = \frac{1}{2} \tag{3.31}$$

$$\text{P (Birey No 2'in başarısız olması)} = \frac{1}{Permütasyon(2,1)} = \frac{1}{2} \tag{3.32}$$

şeklinde olasılıklarla tanımlanabilir. Üçüncü birey durdurulmuş olduğundan KOF'a etkisi yoktur. KOF aşağıdaki şekilde yazılır.

$$L(\beta) = L_{12} \times L_4 \times L_{56} \tag{3.33}$$

3.33 denklemi her bir terim için şu şekildedir.

$$L_{12} = \frac{1}{2}\left(\frac{e^{12\beta}}{e^{12\beta}+e^{6\beta}+e^{8\beta}+e^{9\beta}+e^{7\beta}+e^{11\beta}}\right)\left(\frac{e^{6\beta}}{e^{6\beta}+e^{8\beta}+e^{9\beta}+e^{7\beta}+e^{11\beta}}\right) + \frac{1}{2}\left(\frac{e^{6\beta}}{e^{12\beta}+e^{6\beta}+e^{8\beta}+e^{9\beta}+e^{7\beta}+e^{11\beta}}\right)\left(\frac{e^{12\beta}}{e^{12\beta}+e^{8\beta}+e^{9\beta}+e^{7\beta}+e^{11\beta}}\right) \tag{3.34}$$

$$L_4 = \left(\frac{e^{9\beta}}{e^{9\beta}+e^{7\beta}+e^{11\beta}}\right) \tag{3.35}$$

$$L_{56} = \frac{1}{2}\left(\frac{e^{7\beta}}{e^{7\beta}+e^{11\beta}}\right)\left(\frac{e^{11\beta}}{e^{11\beta}}\right) + \frac{1}{2}\left(\frac{e^{11\beta}}{e^{7\beta}+e^{11\beta}}\right)\left(\frac{e^{7\beta}}{e^{7\beta}}\right) \tag{3.36}$$

### 3.3.2 Breslow Metodu

Olay zamanlarının sürekli oluşu ve $(t_i, t_{i+1})$ aralığında tehlike oranının sabit oluşu Breslow metodunun varsayımlarındandır. Ayrıca herhangi bir bireyin durdurma zamanı $(t_i, t_{i+1})$ aralığında ise, bu bireyin durdurma zamanı aralığın başlangıcı olan $t_i$'dır.

$X_j$; j bireyi için değişken değerleri, $D_i$ kümesi; k zamanında olayı yaşayan $d_i$ bireylerini
içersin, $R_i$; k zamanında hayatta olan veya risk altında olan bireylerin risk kümesi olsun. KOF Breslow formulü aşağıdaki şekildedir [29].

$$L(\beta) = \prod_{i=1}^{k} \frac{e^{\left[\left(\sum_{j\in D_j} X_j\right)\beta\right]}}{\left[\sum_{j\in D_i} e^{\left(X_j\beta\right)}\right]^{d_i}} \tag{3.37}$$

Tablo 3.4'e göre KOF şu şekilde ifade edilir;

$$L(\beta) = L_{12} \times L_4 \times L_{56} \tag{3.38}$$

terimler:

$$L_{12} = \frac{e^{\beta X_1} e^{\beta X_2}}{\left[e^{\beta X_1} + e^{\beta X_2} + e^{\beta X_3} + e^{\beta X_4} + e^{\beta X_5} + e^{\beta X_6}\right]^2} \tag{3.39}$$

$$= \frac{e^{12\beta} e^{6\beta}}{\left[e^{12\beta} + e^{6\beta} + e^{8\beta} + e^{9\beta} + e^{7\beta} + e^{11\beta}\right]^2} \tag{3.40}$$

$$L_4 = \frac{e^{\beta X_4}}{e^{\beta X_4} + e^{\beta X_5} + e^{\beta X_6}} \tag{3.41}$$

$$= \frac{e^{9\beta}}{e^{9\beta} + e^{7\beta} + e^{11\beta}} \tag{3.42}$$

$$L_{56} = \frac{e^{\beta X_5} e^{\beta X_6}}{\left[e^{\beta X_5} + e^{\beta X_6}\right]^2} \tag{3.43}$$

$$= \frac{e^{7\beta} e^{11\beta}}{\left[e^{7\beta} + e^{11\beta}\right]^2} \tag{3.44}$$

olacak şekildedir. Görüldüğü gibi hesaplamalar permütasyon içermediği için Kesin Metoduna göre daha pratiktir. Bu methot SAS ve Python'da varsayılan metottur[30,31].

### 3.3.3 Efron Metodu

Efron metodu R'da varsayılan metottur[32]. Efron metodu daha iyi sonuçlar vermesine rağmen Breslow metodu SAS ve Python'da varsayılan metottur. Efron KOF denklemi aşağıdaki şekilde ifade edilir.

$X_j$; j deneği için açıklayıcı değişken değerleri, $D_i$ kümesi; $t_i$ zamanında başarısız olan $d_i$ bireylerini içermek üzere, $R_i$; $t_i$ zamanında hayatta olan veya risk altında olan bireylerin risk kümesi olsun. KOF Breslow methodu için aşağıdaki şekildedir[33].

$$L(\beta) = \prod_{i=1}^{k} \frac{e^{\left[\left(\sum_{j \in D_j} X_j\right)\beta\right]}}{\prod_{l=1}^{d_i}\left[\sum_{j \in R_i} e^{X_j\beta} - \frac{l-1}{d_i}\sum_{j \in D_i} e^{X_j\beta}\right]} \tag{3.45}$$

Tablo 3.4'de örneğe göre;

$$L(\beta) = L_{12} \times L_4 \times L_{56} \tag{3.46}$$

$L_{12}$ 'yi yazmadan önce

$$M_X = e^{\beta X_1} + e^{\beta X_2} + e^{e^{\beta X_3}} + e^{\beta X_4} + e^{\beta X_5} + e^{\beta X_6} \tag{3.47}$$

şeklinde tanımlanacak olursa

$$L_{12} = \frac{e^{\beta X_1} e^{\beta X_2}}{M_X\left[M_X - \frac{1}{2}\left(e^{\beta X_1} + e^{\beta X_2}\right)\right]} \tag{3.48}$$

buradan

$$M_\beta = e^{12\beta} + e^{6\beta} + e^{8\beta} + e^{9\beta} + e^{7\beta} + e^{11\beta} \tag{3.49}$$

tanımlanırsa

$$L_{12} = \frac{e^{\beta X_1} e^{\beta X_2}}{M_\beta\left[M_\beta - \frac{1}{2}\left(e^{\beta X_1} + e^{\beta X_2}\right)\right]} \tag{3.50}$$

şeklinde olur.

$$L_4 = \frac{e^{\beta X_4}}{e^{\beta X_4} + e^{\beta X_5} + e^{\beta X_6}} \tag{3.51}$$

$$= \frac{e^{9\beta}}{e^{9\beta} + e^{7\beta} + e^{11\beta}} \tag{3.52}$$

$$L_{56} = \frac{e^{\beta X_5} e^{\beta X_6}}{[e^{\beta X_5} + e^{\beta X_6}][e^{\beta X_5} + e^{\beta X_6} - \frac{1}{2}(e^{\beta X_5} + e^{\beta X_6})]} \tag{3.53}$$

$$= \frac{e^{7\beta} e^{11\beta}}{[e^{7\beta} + e^{11\beta}][e^{7\beta} + e^{11\beta} - \frac{1}{2}(e^{7\beta} + e^{11\beta})]} \tag{3.54}$$

denklemin son halidir. Üç metotla alakalı hangisi seçilmelidir konusunda aşağıdaki maddeler dikkate alınabilir[34].

- Eş zamanlı olaylar yoksa üç metotta aynı sonucu verir
- Az sayıda eş zamanlı olaylar varsa herhangi bir metot kullanılabilir
- Eş zamanlı olayların yüzde olarak artması durumunda, en kötü tahmin Breslow metodundadır
- Efron Metodu'nun tahminlemesi genel olarak Breslow'dan daha doğrudur. Kesin metodunun kullanıldığı durumlarda fazla zaman alacak ise Efron metodu seçilebilir

Çalışmamızda kullanılan Python lifeline paketinde varsayılan metot Breslow olmasına rağmen eş zamanlı olayların çok fazla olduğu durumlarda Python Lifeline paketi Efron metodunu tercih edebilmektedir.

## 3.4 Oransal Tehlike Varsayımı

CRMO yaşam süresi dağılımının varsayımı olmamasına rağmen, oransal tehlike varsayımına(OTV) sahiptir. OTV, CRMO'ya başvurmak için ön şarttır. Çalışma süresi boyunca, tehlike fonksiyonu iki farklı grup arasında oransal bir şekilde değişmelidir ve bu oran zamandan bağımsız olarak sabit kalmalıdır. Tehlike oranı gruplar arasında zamanla değişiyorsa bu varsayımın sağlanamaması demektir.

OTV'nin kontrolü için Schoenfeld Artıklarının zamana karşı grafiği çizilebilir. Çizilen grafik, rastgele dağınık gibi görünüyorsa OTV sağlanıyor demektir.

Schoenfeld Artıkları(SA), 1982 yılında Schoenfeld tarafından CRMO'de kullanılmak için önerilmiştir [35]. Değişken değerinden beklenen değer çıkarılarak bulunur. 3.23 denklemi olan skor fonksiyonu kullanılarak hesaplanır. 3.23 denkleminde $a_i$ aşağıdaki şekilde tanımlanırsa:

$$a_i = \frac{\sum_{j \epsilon R(t_i)} X_j e^{\beta^t X_j}}{\sum_{j \epsilon R(t_i)} e^{\beta^t X_j}} \tag{3.55}$$

gösterge fonksiyonu;

$$Y_i(\mathrm{t}) = \begin{cases} 1, \text{eğer birey durdurulmamış} \\ 0\ , \text{eğer birey durdurulmuş} \end{cases} \tag{3.56}$$

Bu durumda k bireyinin herhangi i açıklayıcı değişkeni için aşağıdaki şekilde hesaplanır

$$s_k = Y_i(\mathrm{t})(X_i - a_i) \tag{3.57}$$

Grafiklerle gözlem insanı yanıltabilir ve özneldir. Sonuç odaklı bir tanısal test sağlamaz. Nesnel bir yaklaşım olan Ki-Kare uyum iyiliği testi OTV için uygulanabilir. Testin hipotezi aşağıdaki gibidir.

$$\begin{aligned} H_0 &: \text{OTV sağlanır} \\ H_1 &: \text{OTV sağlanmaz.} \end{aligned} \tag{3.58}$$

Eğer değişken için üretilen p-değeri >0.05 ise OTV sağlanır. OTV'nin sağlanmasını kontrol etmek için uyulanabilecek diğer bir yöntem ise kategorik değişkenler için log(-log(S(t)) grafiğini log(t)'ye karşı çizmektir. Çizilen eğriler kesişmiyor ve parallel ise OTV sağlanır.

OTV kontrolünden sonra CRMO'da OTV sağlamayan değişkenler mevcut ise bu değişkenlerin tehlikeye olan etkisi net değildir ve CRMO geçersizdir. [37]. OTV sağlanmadığı durumlarda literatürde önerilen, çalışmamızda da başvurulan aşağıdaki modeller uygulanabilir[31,36,39,40];

- Tabakalandırılmış CRMO kullanılması
- Genişletilmiş CRMO kullanılması
- Parçalı Regresyon CRMO (PRCRMO) ve Kübik Spline CRMO (KSCRMO) kullanılması
- Hızlandırılmış Başarısızlık Süresi Modeli (HBSM) kullanılması

## 3.5 Tabakalandırılmış (Stratified) CRMO

Tabakalandırılmış CRMO, OTV sağlamayan kategorik değişkenler tabakalandırılmış, sürekli değişkenlerde ayrıklaştırma(binning) yapılarak kategorilere bölünüp, tabakalandırılıp modele dahil edilmesidir.

Tabakalandırılan değişkenlerin benzersiz değerleri üzerinden ver seti M parçaya bölünür. Her bir parçanın temel tehlike fonksiyonu ayrı olmasına rağmen regresyon katsayıları ortak değişkenler için aynıdır. Bu durumda tabakalandırılmış değişkenler için regresyon katsayıları hesaplanmaz ve tehlikeye olan etkisi ölçülmez. Denklem 3.6'da tehlike fonksiyonu aşağıdaki formu alır.

$$\lambda_s(t) = \lambda_{0s}(t) e^{\sum_{i=1}^{p_1} \beta_i X_i + \sum_{j=1}^{p_2} \alpha_j X_j} \tag{3.59}$$

Denklemde, $\lambda_{0s}(t)$; herbir parçanın temel tehlike fonksiyonu (s=1, 2,...,m), $\beta_i$ zamandan bağımsız değişkenlerin, $\alpha_j$ zamana bağlı değişkenlerin katsayı vektörüdür.

## 3.6 Genişletilmiş (Extended) CRMO

OTV sağlamayan değişkenler, zamana bağlı değişkenlerdir. CRMO zamana bağlı aynı zamanda zamandan bağımsız değişkenleri aynı anda analiz edebilecek kabiliyette bir modeldir. Zamandan bağımsız değişkenler kümesi $(X_1, X_2, \dots, X_{k_1})$ ve zamana bağlı değişkenler kümesine $(X_1(t), X_2(t), \dots, X_{k_2(t)})$ şeklinde ifade edilecek olursa; bütün değişkenleri içine alan küme aşağıdaki şekilde ifade edilsin.

$$X(t) = \left(X_1, X_2, \dots, X_{k_1}, X_1(t), X_2(t), \dots, X_{k_2(t)}\right) \tag{3.60}$$

bu durumda tehlike fonksiyonu aşağıdaki formu alır[41].

$$\lambda(t|X(t)) = \lambda_0(t) e^{\sum_{i=1}^{k_1} \beta_i X_i + \sum_{j=1}^{k_2} \alpha_j X_j(t)} \tag{3.61}$$

Denklemde sırasıyla B ve α değişken katsayıları, OTV sağlayan değişken sayısı $k_1$ ve OTV sağlamayan değişken sayısı $k_2$ şeklindedir. Tehlike oranı genişletilmiş CRMO için aşağıdaki şekli alır.

$$e^{\sum_{i=1}^{k_1} \beta_i (X_i - X_i^*) + \sum_{j=1}^{k_2} \alpha_j (X_{j(t)} - X_j^*(t))} \tag{3.62}$$

Bu bilgiler ışığında tehlike fonksiyonu genişletilmiş CRMO altında

$$\lambda(t|X(t)) = \lambda_0(t) e^{\sum_{i=1}^{k} \beta_i X_i + \sum_{j=1}^{k} \alpha_j X_j g_j(t)} \tag{3.63}$$

olup denklemde $g_j(t)$ zamana bağlı bir fonsiyon olup, aşağıdaki ihtimallerden birisi seçilebilir.

$$g_j(t) = 0 \tag{3.64}$$

$$g_j(t) = t \tag{3.65}$$

$$g_j(t) = \log t \tag{3.66}$$

$$g_j(t) = \text{birim basamak fonksiyonu} \tag{3.67}$$

Zamana bağlı değişkenlerin modele girmesi modeli daha karmaşık yapmaktadır. Zamana bağlı değişkenlerde artık sabit bir tehlike oranından bahsedilemez. Tehlike oranı hesaplamalarında zaman değişkenide hesaplamalara dahil edilmelidir.

## 3.7 Parçalı ve Kübik Spline CRMO

Tehlike fonskiyonu için OTV sağlanmadığı taktirde, Parçalı CRMO(PCRMO) ve Kübik Spline CRMO (KSCRMO) kullanılarak tahminlenebilir. Zamana bağlı değişkeni zaman ekseninde parçalara ayrılarak her bir parçada farklı bir polinomla modellenebilir. Parçalanan noktalara düğüm denir.

Denklem 3.6'dan tehlike fonksiyonu aşağıdaki şekilde yazılabilir.

$$\lambda(t|X) = \lambda_0(t)\exp(\beta X) \tag{3.68}$$

Denklem 3.68 'da temel tehlike fonskiyonu olan $\lambda_0(t)$ PCRMO veya KSCRMO ile tahminlenebilir. KSCRMO için aşağıdaki şekilde tahminlenir[31].

$$\lambda_0(t) = \exp\left(\phi_0 + \phi_1 \log t + \sum_{j=1}^{N} \phi_{j+1} v_j(\log t)\right) \tag{3.69}$$

3.69 denkleminde n adet eşsiz düğüm noktaları $t_1 < t_2 < \cdots < t_n$ ve $t_1 > t_{min}$ ve $t_m < t_{max}$ olsun, $v_j$ kübik spline terimleri olmak üzere aşağıdaki şekilde tanımlanır [42, 43].

$$v_j(\log t) = \left(\log\left(t - t_j\right)\right)_+^3 - \lambda_j(\log\left(t - t_{min}\right))_+^3 - \left(1 - \lambda_j\right)(\log\left(t - t_{max}\right))_+^3 \tag{3.70}$$

3.70 denkleminde $\log\left(t - t_j\right)_+$:

$$\log\left(t - t_j\right)_+ = \max\left(0, \log\left(t - t_j\right)\right) \tag{3.71}$$

3.70 denkleminde $\lambda_1$:

$$\lambda_j = \frac{t_{max} - t_j}{t_{max} - t_{min}} \tag{3.72}$$

şeklindedir.İki düğümlü bir tahminleme algoritması için 3.70 formulü aşağıdaki formu alır $(t_1 < t_2)$

$$\lambda_0(t) = \exp\left(\phi_0 + \phi_1 \log t + \phi_2 v_1(\log t) + \phi_3 v_2(\log t)\right) \tag{3.73}$$

3.73 denkleminde sırasıyla $v_1(\log t)$:

$$v_1(\log t) = (\log(t - t_1))_+^3 - \lambda_1(\log(t - t_{min}))_+^3 - (1 - \lambda_1)(\log(t - t_{max}))_+^3 \quad (3.74)$$

3.73 denkleminde $v_2(\log t)$:

$$v_2(\log t) = (\log(t - t_1))_+^3 - \lambda_2(\log(t - t_{min}))_+^3 - (1 - \lambda_2)(\log(t - t_{max}))_+^3 \quad (3.75)$$

şeklindedir.

β regresyon katsayıları, [36] KOF denkleminde önerilen bir şekilde güncellenip katsayılar tahminlenir.

Bu modellemede zor bir süreç, düğüm noktalarını adet ve lokasyon olarak doğru seçmektir. Pratikte üç,dört veya beş düğüm seçilebilir. 3 düğüm seçildiğinde 10,50,90 yüzdelikleri, 4 düğüm seçildiğinde 5, 35, 65 ve 95 yüzdelik dilimler, 5 düğümde 5, 27, 5 , 50, 72,5 ve 95 yüzdelik dilimlerin seçilmesi tavsiye edilir[44, 45].

# 4
# PARAMETRİK REGRESYON MODELİ

## 4.1 Hızlandırılmış Başarısızlık Süresi Modeli

### 4.1.1 Model

CRMO son derece başarılı, popüler ve esnek bir model olmasına rağmen, modelin ön şartı OTV'nin sağlanmasıdır[46]. HBSM işte bu noktada devreye girer ve OTV sağlanmadığı durumlarda veri setine uygulanabilir. HBSM'de yaşam süresi bilinen bir olasılık dağılımına sahip olduğu varsayılır. Weibull, Log-Lojistik ve Log-Normal bu alanda kullanılan ve popüler olan bazı olasılık dağılımlarıdır. Doğru dağılım seçildiği taktirde daha düzgün bir model yapısı karşımıza çıkmakta olup, daha doğru ve efektif parametre tahminleri elde edilir[47]. Yaşam fonksiyonu, verinin analiz edilmesi için çok iyi bir betimsel istatistik olduğundan, denkleminin bilinmesi, açıklayıcı değişkenlerin tanınması ve yorumlanması adına büyük avantajdır[48].

### 4.1.2 Metot

HBSM aşağıdaki şekilde tanımlanır.

$$\mathrm{S(t)} = S_0(\eta(\mathrm{x})t), t \geq 0 \tag{4.1}$$

Denklemde $X_1, X_2, \dots, X_P$ açıklayıcı bağımsız değişkenler ve $\beta_0, \beta_1, \dots, \beta_P$ regresyon katsayıları olmak üzere, $\eta$ hızlandırma faktörü denir ($\eta \in \mathbb{R}$) ve formulü şu şekildedir.

$$\eta(x) = e^{\beta X} = e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_\mathrm{p}} \tag{4.2}$$

Hızlandırma faktörünü örnekle iki grup karşılaştırılıp açıklanacak olursa, bir grupta ki bireylerin mevcut x değerleri ile t yaşam süresinde sağkalım olasılığı ile diğer grubun $e^{\beta X}\mathrm{t}$ yaşam süresinde sağkalım olasılığı aynıdır. Yani ortalama yaşam süresi $e^{\beta X}$ kadar hızlandırılmış veya yavaşlatılmıştır.

Hızlandırma faktörü eğer:

- $e^{\beta X} > 1$, değişken yaşam süresini uzatır

- $e^{\beta X} < 1$ değişken yaşam süresini kısaltır

Denklem 4.1 yaşam süresine herbir açıklayıcı değişkenin etkisini ölçer ve istatistiksel çıkarımlar yapmamızı sağlar.

HBSM örnekle açıklanacak olursa;

$$S_1(t) = \text{Köpeklerin yaşam eğrisi}$$
$$S_2(t) = \text{İnsanların yaşam eğrisi} \tag{4.3}$$

Köpeklerin insanlardan yedi kat daha hızlı yaşlandığı bilinmektedir. Bu durumda sağkalım eğrileri arasındaki ilişki aşağıdaki şekildedir;

$$S_1(t) = S_2(7t), \text{ t yıl bazında } t \geq 0 \tag{4.4}$$

Sözle ifade edilirse, bir köpeğin ortalama 10 yıl ve daha fazla sağkalım ihtimali ile bir insanın ortalamada 70 yıl ve daha fazla sağkalım ihtimali aynıdır.

HBSM tehlike fonksiyonu aşağıdaki şekilde bulunur. Denklem 2.25'i kullanıp S(t) yerine denklem 4.1 yazılırsa;

$$\lambda(t) = \frac{S_0'(\eta(x)t)}{S_0(\eta(x)t)} \tag{4.5}$$

$$= -\eta(x)\lambda_0(\eta(x)t) \tag{4.6}$$

HBSM tehlike fonskiyonu elde edilir.

Her bir bireyin yaşam süresi $T_i$, pozitif olduğu için logaritması alınabilir ve logaritması alındığında HBSM'nin logaritmik doğrusal formu aşağıdaki şekilde elde edilir.

$$\log T_i = \mu + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \sigma\varepsilon_i \tag{4.7}$$

4.7 denkleminde p adet açıklayıcı değişken için $[\beta_1 \ldots \beta_p]$ katsayı vektörü, $[X_1, X_2, \ldots, X_p]$ açıklayıcı değişken vektörü, $\mu$ ise X eksenini kestiği nokta $(\beta_0)$, $\sigma$ ölçek parametresi, $\varepsilon_i$ ise hata terimleridir.

Yaşam süresi T'nin dağılımına göre karşlık gelen Log T dağılımları tablo 4.1'da verilmiştir.

**Tablo 4.1** Log T dağılımı

| T Dağılımı | Log T Dağılımı |
|---|---|
| Weibull | Uç Değer (2 Parametre) |
| Log-Lojistik | Lojistik |
| Log-Normal | Normal |

4.7 denklemi durdurulmuş veri olmadığı durumlarda geleneksel regresyon problemidir. Geleneksel regresyon analizi durdurulmuş verilerle analiz yapabilme yeteneğine sahip değildir. HBSM bu yeteneği ile regresyon analizinden ayrılır. Literatüre durdurulmuş veri içeren veri setlerine regresyon analizi yapılan çalışmalar ve denemeler mevcuttur[49,50,51].

### 4.1.3 Weibull HBSM

Weibull dağılımı iki parametreye sahip bir dağılımdır. Parametreler (λ,p) olarak temsil edilecek olursa, λ >0 ve p>0 olmak üzere, p şekil parametresi, λ ölçek parametresidir.

Weibull Dağılımının sırasıyla f(t), F(t), S(t) ve $\lambda(t)$ formülleri aşağıdaki şekildedir.

**f(t):** Farz edelim ki $T \backsim$ Weibull(λ,p) olsun. Bu durumda olasılık yoğunluk fonksiyonu olan f(t) şu sekildedir;

$$f(t) = \lambda p t^{p-1} exp(-\lambda t^p) \text{ , p>0, t>0} \tag{4.8}$$

F(t) aşağıdaki şekilde bulunur.

$$F(t) = \int_{-\infty}^{t} f(x)dx \tag{4.9}$$

$$= \int_{0}^{t} \lambda p x^{p-1} \exp(-\lambda x^p)\, dx \tag{4.10}$$

Burada değişken değiştirme yöntemi kullanılarak integral alınırsa,

$$u = \lambda x^p => x = 0 \; için \; u = 0 \tag{4.11}$$

$$=> x = t \; için \; u = \lambda t^p \tag{4.12}$$

$$du = \lambda p x^{p-1} \tag{4.13}$$

,4.11'da yerine yazıldığında,

$$F(t) = \int_0^{\lambda t^p} e^{-u} du \tag{4.14}$$

$$= -e^{-u}]_0^{\lambda t^p} \tag{4.15}$$

$$= 1 - e^{-\lambda t^p} \tag{4.16}$$

şeklinde elde edilir.

**S(t):** 4.16 denklemi kullanıldığında

$$S(t) = 1 - F(t) \tag{4.17}$$

$$= e^{-\lambda t^p} \tag{4.18}$$

şeklinde ifade edilir.

**$\lambda(t)$:** S(t) ve f(t) kullanılırsa λ(t) şu şekilde bulunur.

$$\lambda(t) = \frac{f(t)}{S(t)} \tag{4.19}$$

$$= \frac{\lambda p t^{p-1} \exp(-\lambda t^p)}{e^{-\lambda t^p}} \tag{4.20}$$

$$= \lambda p t^{p-1} \tag{4.21}$$

Eğer şekil parametresi p:

- p>1 ise tehlike artar
- p=1 ise tehlike sabittir
- p<1 is tehlike azalır

#### 4.1.3.1 Weibull HBMS Varsayımı

Weibull modelinin HBSM varsayımı sağlandığında S(t) aşağıdaki şekilde hesaplanır. Denklem 4.7 ve tablo 4.1'da ε dağılım Uç Değer (2 parametre) olduğundan sağ kalım fonksiyonu:

$$S_{\varepsilon_i}(\varepsilon) = \exp(-\exp(\varepsilon)) \tag{4.22}$$

şeklindedir. 4.7 denkleminden $\varepsilon$ yalnız bırakılıp 4.22 denkleminde yerine yazılırsa

$$S_{\varepsilon_i}(\varepsilon) = \exp\left[-\exp\left(\frac{\log t - \mu - \beta_1 X_1 - \beta_2 X_2 - \cdots - \beta_P X_P}{\sigma}\right)\right] \tag{4.23}$$

$$= \exp\left[-\exp\left(\frac{-\mu - \beta_1 X_1 - \beta_2 X_2 - \cdots - \beta_P X_P}{\sigma}\right) t^{\frac{1}{\sigma}}\right] \tag{4.24}$$

denklemi, Weibull dağılımı olup parametreleri sırasıyla

$$\lambda = \exp\left(\frac{-\mu - \beta_1 X_1 - \beta_2 X_2 - \cdots - \beta_P X_P}{\sigma}\right), p = \frac{1}{\sigma} \tag{4.25}$$

şeklinde bulunur. Son olarak HBSM temel varsayımları olan Tablo 4.1'da dağılımlardan T dağılımı Weibull dağılımı olduğunda, log T dağılımının uç değer dağılımı olduğu belirtilmişti. Weibull dağılımı için bunun ispatı aşağıdaki şekilde yapılabilir.

**Varsayım:** T'nin dağılımı Weibull olduğunda, Y = log T'nin dağılımı Uçdeğer bir dağılımdır

**İspat:** $T_i$ Varsayalım ki Weibull olarak dağılsın ($T_i \sim$Weibull(λ,p))

$$F_Y(y) = P(Y \leq y) \tag{4.26}$$

$$= P(\log T_i \leq y) \tag{4.27}$$

$$= P(T_i \leq e^y) \tag{4.28}$$

$$= F_T(e^y) \tag{4.29}$$

T, Weibull dağılımına sahip olduğu için F(t) denklemi bellidir ve $e^y$ denklemde yerine yazılıp;

$$= 1 - e^{-\lambda(e^y)^P} \tag{4.30}$$

buradanda $-\lambda(e^y)^P$ ifadesini $-\left(e^y \lambda^{\frac{1}{p}}\right)^P$ olacak şekile, $\lambda$ parantez içine alınıp, denklemde yerine yazılırsa,

$$= 1 - e^{-\left(e^y \lambda^{\frac{1}{p}}\right)^P} \tag{4.31}$$

$-\left(e^{y}\lambda^{\frac{1}{p}}\right)^{P}$ ifadesinde $\lambda$ logaritmik bir şekilde yazılırsa aşağıdaki şekilde ifade edilir.

$$= 1 - e^{-e^{\frac{y+\log\lambda^{\frac{1}{p}}}{\frac{1}{p}}}} \tag{4.32}$$

$$= 1 - e^{-e^{\frac{y-(-\log\lambda^{\frac{1}{p}})}{\frac{1}{p}}}} \tag{4.33}$$

Bu denklem 2 parametreli ve sırasıyla parametreleri

$$\sigma = \frac{1}{p}, \mu = -\log\lambda^{\frac{1}{p}} \tag{4.34}$$

olan Uç Değer dağılımını verir.

### 4.1.4 Log-Lojistik HBSM

Analiz edilen veri setinde, tehlike fonksiyonu grafikleri monotonsa Weibull dağılımı ile modellenebilir. Weibull dağılımını sadece bu durumlar için sınırlandırır. Tehlike fonksiyonu, monoton olmayan bir eğriye sahip olabilir. Bu durumda Log-Lojistik dağılımı uygulanabilir.

Log Lojistik dağılımı, Log-Normal dağılımına çok benzer. Durdurulmuş veri için daha kolay takip edilebilir, işlenebilir olması açısından Log-Normal dağılımından daha üstündür[48].

İki parametreye sahiptir ve bu parametreler; λ ölçek parametresi olup açıklayıcı değişkenler hakkında bilgi verir ve diğer parametre olan p>0 şekil parametresidir. Log-Lojistik dağılımının sırasıyla f(t),F(t),S(t) ve $\lambda(t)$ formülleri aşağıdaki şekildedir.

**f(t):**

$$f(t) = \frac{\lambda p t^{p-1}}{(1+\lambda t^{p})^{2}} \tag{4.35}$$

f(t) kullanlaraktan F(t) aşağıda ki şekilde bulunabilir.

$$F(t) = \int_{-\infty}^{t} f(x)dx \tag{4.36}$$

$$= \int_{0}^{t} \frac{\lambda p x^{p-1}}{(1+\lambda x^{p})^{2}} dx \tag{4.37}$$

Burada değişken değiştirme yöntemi kullanılarak integral alınıp

$$u = 1 + \lambda x^{p} => x = 0 \; için \; u = 1 \tag{4.38}$$

$$=> x = t \; için \; u = 1 + \lambda t^{p} \tag{4.39}$$

$$du = \lambda p x^{p-1} \tag{4.40}$$

Denklem 4.37'de yerlerine yazıldığında:

$$F(t) = \int_{0}^{1+\lambda t^{p}} \frac{1}{u^{2}} du \tag{4.41}$$

$$= -u^{-2+1}]_{1}^{1+\lambda t^{p}} \tag{4.42}$$

$$= -\left(\frac{1}{1+\lambda t^{p}} - 1\right) \tag{4.43}$$

$$= 1 - \frac{1}{1+\lambda t^{p}} \tag{4.44}$$

$$= \frac{\lambda t^{p}}{1+\lambda t^{p}} \tag{4.45}$$

F(t) elde edilir.

**S(t):** şu şekilde ifade edilir:

$$S(t) = 1 - F(t) \tag{4.46}$$

$$= 1 - \frac{\lambda t^{p}}{1+\lambda t^{p}} \tag{4.47}$$

$$= \frac{1}{1+\lambda t^{p}} \tag{4.48}$$

$\boldsymbol{\lambda(t)}$: şu şekilde ifade edilir:

$$\lambda(t) = \frac{f(t)}{S(t)} \tag{4.49}$$

$$= \frac{\frac{\lambda p t^{p-1}}{(1+\lambda t^{p})^{2}}}{\frac{1}{1+\lambda t^{p}}} \tag{4.50}$$

Buradan denklem yeniden düzenlenirse:

$$= \frac{\lambda p t^{p-1}}{(1+\lambda t^{p})^{2}} (1 + \lambda t^{p}) \tag{4.51}$$

$$= \frac{\lambda p t^{p-1}}{1+\lambda t^{p}} \quad p > 0, \lambda > 0 \tag{4.52}$$

λ(t) 'nin son hali elde edilir. λ(t) grafiği için aşağıdaki çıkarımlar yapılabilir.

- p<1 için tehlike fonksiyonu azalır
- p≥1 tehlike fonksiyonu önce artar daha sonra azalır

#### 4.1.4.1 Log-Lojistik HBSM Varsayımı

Log-Lojistik modelinin HBSM varsayımı sağlandığında S(t) aşağıdaki şekilde hesaplanır. Denklem 4.8 ve tablo 4.1'da ε Lojistik dağılımına (2 parametre) sahip olduğundan sağkalım fonksiyonu aşağıdaki şekildedir:

$$S_{\varepsilon_i}(\varepsilon) = \frac{1}{(1 + \exp(\varepsilon))} \tag{4.53}$$

4.8 denkleminde $\varepsilon$ yalnız bırakıp 4.51 denkleminde yerine yazılırsa

$$S_{\varepsilon_i}(\varepsilon) = \frac{1}{1 + \exp\left(\frac{\log t - \mu - \beta_1 X_1 - \beta_2 X_2 - \cdots - \beta_P X_P}{\sigma}\right)} \tag{4.54}$$

$$= \left[1 + t^{\frac{1}{\sigma}} \exp\left(\frac{-\mu - \beta_1 X_1 - \beta_2 X_2 - \cdots - \beta_P X_P}{\sigma}\right)\right]^{-1} \tag{4.55}$$

denklemi, Log-Lojistik dağılımının sağkalım fonksiyonu olup parametreleri sırasıyla

$$\lambda = \exp\left(\frac{-\mu - \beta_1 X_1 - \beta_2 X_2 - \cdots - \beta_P X_P}{\sigma}\right), p = \frac{1}{\sigma} \tag{4.56}$$

şeklindedir.

### 4.1.5 Log-Normal HBSM

Yaşam sürelerinin dağılımı Log-Normal olduğu varsayılırsa (T~lognormal(μ,σ)), S(t) HBSM varsayımı altında şu şekilde hesaplanır.

$$S(t) = P(T > t) \tag{4.57}$$

$$= P(Log\ T > \log t) \tag{4.58}$$

$$= 1 - \phi\left(\frac{\log(t)-\mu}{\sigma}\right) \tag{4.59}$$

Denklemde $\phi$ standart normal dağılımın birikimli yoğunluk fonksiyonudur. Log-Normal modelinin HBSM varsayımı sağlandığında S(t) aşağıdaki şekilde bulunur. Denklem 4.8 ve tablo 4.1'da ε normal dağılıma sahip olduğu için sağkalım fonksiyonu aşağıdaki şekilde ifade adilir.

$$S_{\varepsilon_i}(\varepsilon) = 1 - \phi(\varepsilon) \tag{4.60}$$

$$= 1 - \varphi\left(\frac{\log t - \mu - \beta_1 X_1 - \beta_2 X_2 - \cdots - \beta_P X_P}{\sigma}\right) \tag{4.61}$$

Bu denklem Log-Normal dağılımının yaşam fonksiyonu denklemi olup parametreleri sırasıyla

$$\mu = \mu + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_P X_P \tag{4.62}$$

$$\sigma = \sigma \tag{4.63}$$

şeklindedir.

## 4.2 HBSM Olabilirlik Fonksiyonu

Olabilirlik fonksiyonu kullanaraktan en çok olabilirlik yöntemi ile regresyon parametreleri tahmin edilir. Gözlenen sıralı başarısızlık süreleri $t_1, t_2, \ldots, t_n$ olmak üzere aşağıdaki şekildedir.

$$L(\beta, \mu, \sigma) = \sum_{i=1}^{n} \{f_i(t_i)\}^{\delta_i} \{S_i(t_i)\}^{1-\delta_i} \tag{4.64}$$

İ bireyi için zaman $t_i$'de, $f_i(t_i)$ olasılık yoğunluk fonksiyonu, $S_i(t_i)$ yaşam fonksiyonu, $\delta_i$ başarısızlık gösterge fonksiyonudur.

4.64 denkleminde S(t) aşağıdaki şekilde bulunur.

$$S_{\mathrm{i}}(t) = P(T_i > t) \tag{4.65}$$

Denklemde her iki tarafın logu alınıp

$$= P(\log T_i > \log t) \tag{4.66}$$

$\mathrm{Log}\,T_i$ yerine denklem 4.8'i kullanarak 4.65 denkleminde yerine yazıp ifade edildiğinde

$$= P\left(\mu + X_1\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \sigma\varepsilon_i > \log t\right) \tag{4.67}$$

$\varepsilon_i$ denklemde tek bırakıldığında

$$= S_{\varepsilon_i}\left(\varepsilon_i > \frac{\log t - \mu - \beta X_i}{\sigma}\right) \tag{4.68}$$

hata terimininin yaşam fonksiyonu elde edilir.

$$= S_{\varepsilon_i}\left(\frac{\log t - \mu - \beta X_i}{\sigma}\right) \tag{4.69}$$

Denklemde $S_{\varepsilon_i}(t)$, tablodaki Log T dağılımlarına göre aşağıdaki formları alabilir.

Normal Dağılım ;

$$S_{\varepsilon_i}(t) = 1 - \phi(t) \tag{4.70}$$

Denklemde $\phi(t)$ standart normal dağılımın F(t) denklemidir.

Uçdeğer Dağılımı ;

$$S_{\varepsilon_i}(t) = e^{-e^t} \tag{4.71}$$

Lojistik Dağılımı;

$$S_{\varepsilon_i}(t) = (1 + e^t)^{-1} \tag{4.72}$$

f(t)'yi bulabilmek için 4.69 denkleminin t'ye göre türevi alınırsa aşağıda ki sonuç elde edilir.

$$f_{\mathrm{i}}(t) = (\sigma t)^{-1} f_{\varepsilon_i}\left(\frac{\log t - \mu - \beta X_i}{\sigma}\right) \tag{4.73}$$

Denklem 4.7'i kullanaraktan $z_i$ aşağıdaki şekilde tanımlanıp

$$z_i = \frac{\log t_i - \mu - \beta X_i}{\sigma} \tag{4.74}$$

ve olabilirlik fonskiyonun logaritmasını alınırsa, karşımıza Log-olabilirlik fonksiyonu çıkar ve aşağıdaki şekilde ifade edilir.

$$l(\beta, \mu, \sigma) = \log L(\beta, \mu, \sigma) \tag{4.75}$$

$$= \sum_{i=1}^{n} \{-\delta_i \log(\sigma t_i) + \delta_i \left(\log f_{\varepsilon_i}(z_i)\right) + (1 - \delta_i) \log S_{\varepsilon_i}(z_i)\} \tag{4.76}$$

$\left(x_{i1}, \ldots, x_{ij}, \ldots, x_{ip}\right)$ i deneğinin açıklayıcı değişken vektörü olmak üzere, tahminlenecek p+2 tane parametreler vektörü $\left(\sigma, \mu, \beta_1, \beta_2, \ldots, \beta_p\right)$ 'dir. Log olabilirlik fonksiyonu maksimize edilerek parametre tahmini yapılır. Aşağıda örnek olarak seçilen iki parametre $\hat{\sigma}$, $\widehat{\beta}$ ,sırasıyla $\frac{\partial l}{\partial \sigma} = 0$ ,$\frac{\partial l}{\partial \beta} = 0$ çözülerek bulunur.

$$\frac{\partial l}{\partial \beta_j} = -\frac{1}{\sigma} \sum_{i=1}^{n} [\delta_i \frac{\partial \log f_{e_i}(z_i)}{\partial z_i} + (1 - \delta_i) \frac{\partial \log S_{\varepsilon_i}(z_i)}{\partial z_i}] x_{ij} \tag{4.77}$$

$$\frac{\partial l}{\partial \sigma} = -\frac{r}{\sigma} - \frac{1}{\sigma} \sum_{i=1}^{n} \left[\delta_i z_i \frac{\partial \log f_{e_i}(z_i)}{\partial z_i} + (1 - \delta_i) z_i \frac{\partial \log S_{\varepsilon_i}(z_i)}{\partial z_i}\right] \tag{4.78}$$

## 4.3 Model Seçimi

HBSM ve CRMO modelleri içerisinde en iyi modeli seçmek için AIC metriği kullanılmış olup tanımı aşağıdaki şekildedir.

### 4.3.1 Akaike Information Criterion (AIC)

Birden fazla yarı parametrik ve parametrik modelleri karşılaştırmak için kullanılır. Akaike tarafından önerilen bu metot [52], tahmin edilmiş, hesaplanmış istatistiksel modelin uyum denencesi ölçümüdür. AIC aşağıda ki şekilde bulunur.

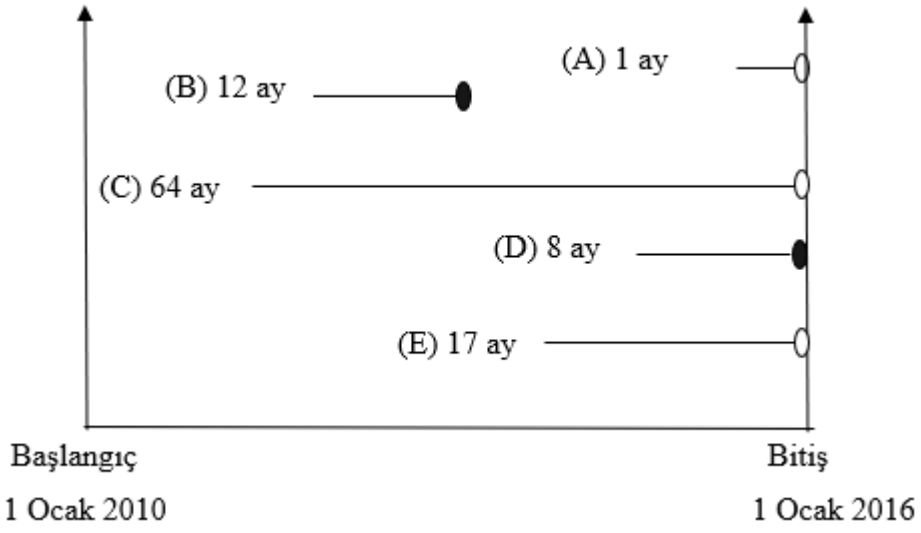
$$\text{AIC}=\text{-2(Log Likelihood)} +2(\text{P}+\text{K}) \tag{4.79}$$

Olabilirlik denkleminden çıkan sabit bir değerdir. P dağılımın sahip olduğu parametre sayısı, K'da sabit terim dışında toplam katsayı adetidir. AIC değeri düştükçe model performansı artar.

# 5
# TELEKOM SEKTÖRÜ MÜŞTERİ KAYIP VERİLERİNE UYGULANIŞI

## 5.1 Müşteri Kayıp Verileri

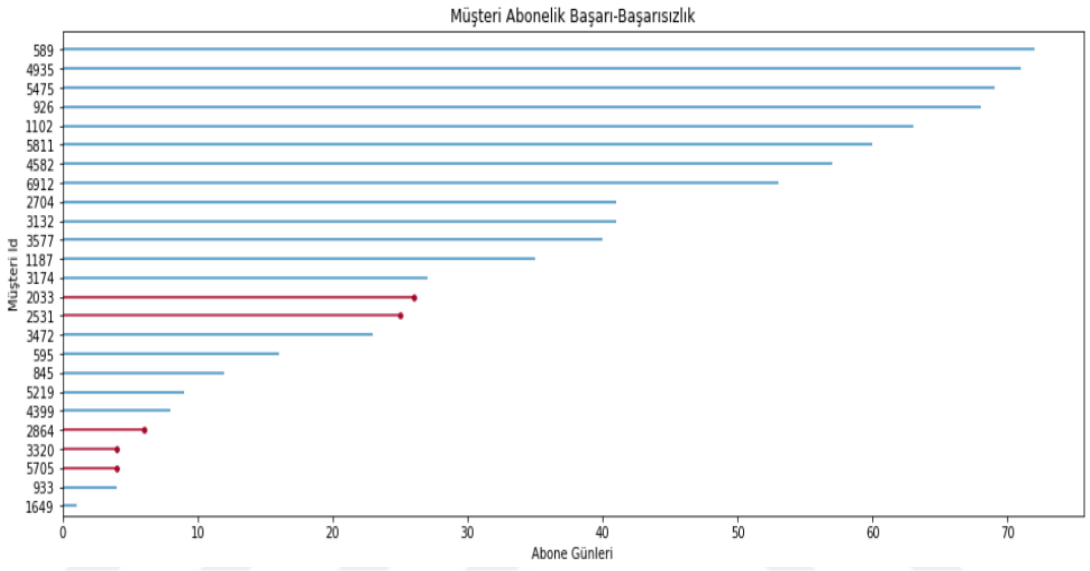
Çalışmamızda, Kaggle [13] sitesinden mobil operatör müşteri kayıp veri seti kullanılmıştır. 18 açıklayıcı değişken içeren ve 7032 adet müşterinin 72 ay boyunca izlenmesiyle oluşturulmuştur. Çalışmanın başlangıç ve bitiş tarihi veri setinde belirtilmemiştir. Muhtemelen sabit bir bitiş tarihi seçilmiş olup; örnek olarak 2015 Ekim, Kasım ve Aralık tarihlerinde aktif olan müşterilerden rastgele 7032 müşteri örneklemi alınıp, 6 yıl geriye gidilmiştir. Müşterilerin aldığı ürünler, servisler, paketler ve müşteri bilgileri ile birlikte veri seti oluşturulmuştur. 72 ay boyunca 1869 müşteri kaybı yaşanmış, 5163 müşteri sağdan durdurulmuş olup, çalışma bitmesine rağmen kayıp yaşamamıştır. Burada önemli nokta çalışma takvimi ile çalışma süresi arasındaki farkı dikkate almaktır. Çalışma süresi olan 6 yıl içinde müşterilik ömürleri not edilmiştir.




**Şekil 5.1** Çalışma takvimi

Şekil 5.1 'den örnek verilecek olursa, A müşterisi, müşteri olduktan 1 ay sonra çalışma bitmiş ve hala kayıp yaşamamış olup sağdan durdurulmuş veridir. B müşterisi, 12 ay sonra kayıp yaşamış, C müşterisi, müşteri olduktan 64 ay sonra

çalışma bitmiş ve kayıp yaşamamıştır ve sağdan durdurulmuş veridir. D abonesi 8 ay sonra kayıp yaşamış, E müşterisi müşteriliğe başladıktan 17 ay sonra çalışma bitmiş ve hala kayıp yaşamadığından sağdan durdurulmuş veridir. Şekilde 5.2'de rastgele seçilen 25 müşterinin çalışma süresi boyunca başarı-başarısızlık eğrileri çizilmiştir. Kırmızı müşteriler için müşterilik iptali gerçekleşmiş, mavi renkteki aboneler için ise abonelik iptali gerçekleşmemiş olup durdurulmuş bireyler olarak tanımlanmaktadırlar.




**Şekil 5.2** Çalışma süresi

Tablo 5.1'de müşterilerin aldığı ürünler, hizmetler ve tarifeler mevcuttur. Tarifede dikkat çeken nokta, faturasız hatların çok oluşudur. Faturasız hatların müşterilik ömrüne olan etkisi incelenmek istenmiş olabilir. Faturasız hatlar, eknomisi gelişmekte olan bazı Latin Amerika ve Afrika ülkelerinde revaçta olup faturalı hatlara oranı 80:20'dir[53]. Eknonomik durumu iyi olmayan müşterilerin, kendilerini zorunlu ve taahhütlü hatlara maddi olarak bağlamak istememeleri ve istedikleri operatörler arasında geçiş yapabilme özgürlüğünü istemeleridir. Bu ülkelerde mobil operatörlerin gelirlerinin büyük bir kısmı bu müşterilerden oluştuğu için operatörlerin bir amacıda bu abonelerin etkili bir şekilde elde tutulması için stratejiler ve modeller geliştirmek olabilir. SKA bu çalışma için büyük bir avantajdır. Müşterilerin, müşterilik ömürlerini uzatmak ve müşteri kaybını minimize etmek en büyük hedeftir. Yaş/cinsiyet gibi demografik değişkenlerin, herhangi bir partnerinin aynı mobil operatörde müşteri olması,

güvenlik ve film gibi ürün paketlerinin, ortalamada müşterilik ömrünü ne kadar uzattığı veya kısaltığı konuları bu çalışma ile cevaplanmış olup, cevaplanan bu sorularla pazarlama stratejilerine ışık tutulmak istenmiştir.

SKA modelleme çalışmaları Python'da Lifelines paketi kullanılarak analiz edilmiştir [31].

**Tablo 5.1** Değişken tanımları ve düzeyleri

| Değişken | Tür | Değişken Düzeyleri | Olay | Durdurulmuş Olay Sayısı |
|---|---|---|---|---|
| Cinsiyet | Kategorik | Kadın (%49,5)<br>Erkek (%50.5) | 939<br>930 | 2.544<br>2.619 |
| Yaşlı mı | Kategorik | Evet (%16,2)<br>Hayır (%83,8) | 476<br>1.393 | 666<br>4.497 |
| Online Fatura | Kategorik | Evet (%59,2)<br>Hayır (%40,8) | 1,400<br>469 | 2.768<br>2.395 |
| Partneri Var mı | Kategorik | Evet (%48,2)<br>Hayır (%51,8) | 669<br>1.200 | 2.724<br>2.439 |
| Sabit Telefon | Kategorik | Evet (%90,3)<br>Hayır (%9,7) | 1.699<br>170 | 4.653<br>510 |
| Yakını Var mı | Kategorik | Evet (%29,8)<br>Hayır (%70,2) | 326<br>1.543 | 1.773<br>3.390 |
| Cihaz Koruma | Kategorik | Evet (%34,4)<br>Hayır (%44)<br>İnternet Hizmeti Yok (%21.6) | 545<br>1.211<br>113 | 1.873<br>1.883<br>1.407 |

**Tablo 5.1** Değişken tanımları ve düzeyleri (devamı)

| | | | | |
|---|---|---|---|---|
| Birden Fazla Hattı Var mı | Kategorik | Evet (%42,2)<br>Hayır (%48,1)<br>Telefon Servisi Yok (%9,7) | 850<br>849<br>170 | 2.117<br>2.536<br>510 |
| Internet Hizmeti | Kategorik | DSL (%34,4)<br>Fiber Optik (%44)<br>Hayır (%21,6) | 459<br>1.297<br>113 | 1.957<br>1.799<br>1.407 |
| Tarife | Kategorik | Faturasız Hat (%55)<br>Yıllık Kontrat (%21)<br>İki Yıllık Kontrat (%24) | 1.655<br>166<br>48 | 2.220<br>1.305<br>1.638 |
| Fatura Ödeme Şekli | Kategorik | Elektronik Çek (%33,1)<br>İban Transfer Otomatik (%22)<br>Kredi Kartı Otomatik (%22)<br>Posta Çeki (%22,9) | 1071<br>258<br>232<br>308 | 1.294<br>1.284<br>1.289<br>1.296 |
| Online Güvenlik | Kategorik | Evet (%29)<br>Hayır (%50)<br>Internet Hizmeti Yok (%21) | 295<br>1461<br>113 | 1.720<br>2.036<br>1.407 |
| Online Yedekleme | Kategorik | Evet (%35)<br>Hayır (%44)<br>Internet Hizmeti Yok (%21) | 523<br>1233<br>113 | 1.902<br>1.854<br>1.407 |
| Kesintisiz Film | Kategorik | Evet (%39)<br>Hayır (%40)<br>Internet Hizmeti Yok (%21) | 818<br>938<br>113 | 1.913<br>1.843<br>1.407 |

**Tablo 5.1** Değişken tanımları ve düzeyleri (devamı)

| | | | | |
|---|---|---|---|---|
| Kesintisiz TV | Kategorik | Evet (%39)<br>Hayır (%40)<br>Internet Hizmeti Yok (%21) | 814<br>942<br>113 | 1.889<br>1.867<br>1.407 |
| Teknik Destek | Kategorik | Evet (%29)<br>Hayır (%49)<br>Internet Hizmeti Yok (%22) | 310<br>1446<br>113 | 1.730<br>2.026<br>1.407 |
| Toplam Fatura | Sürekli | | 1869 | |
| Aylık Fatura | Sürekli | | 1869 | |

%73 sağdan durdurulmuş veri içeren kayıp veri setine daha önceden teorik bilgileri verilen aşağıda ki modeller sırasıyla başvurulmuştur.

- KM
- CRMO
- HBSM

## 5.2 Kaplan Meier

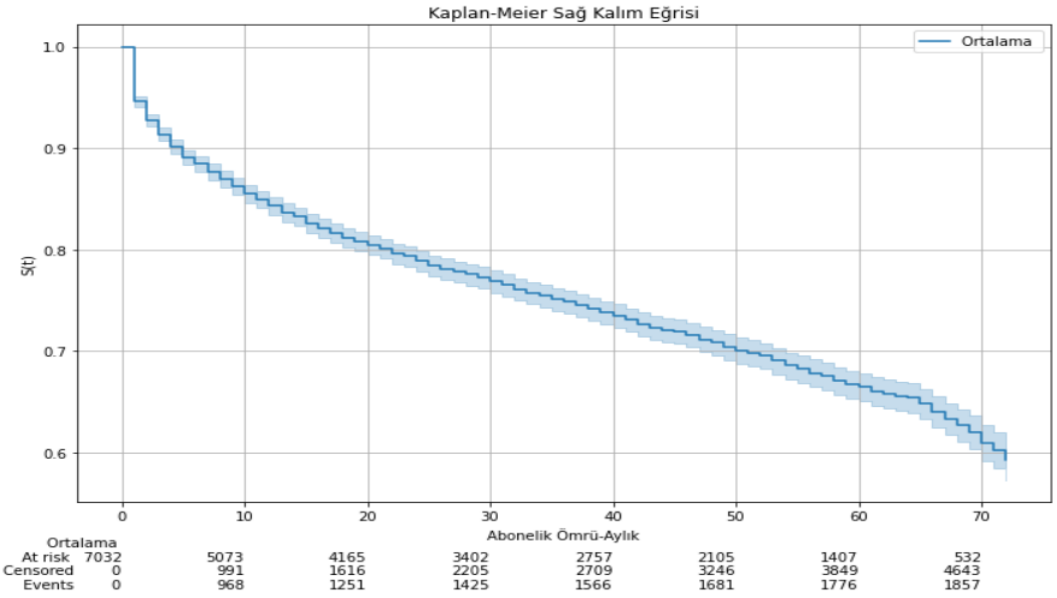
Şekil 5.3'de, çalışmadaki bütün müşterilerin müşterilik süresi sağkalım eğrisi çizilmiş ve müşterilerin 10,20,30,40,50,60,70 ay ve daha fazla müşterilik ömürleri sağkalım olasılıkları tablo 5.2'de belirtilmiştir.

**Tablo 5.2** Sağkalım olasılık tablosu

| Ay | Sağkalım Olasılığı |
|---|---|
| 10 | 0,85 |

**Tablo 5.2** Sağkalım olasılık tablosu (devamı)

| | |
|---|---|
| 20 | 0,8 |
| 30 | 0,77 |
| 40 | 0,73 |
| 50 | 0,7 |
| 60 | 0,66 |
| 70 | 0,61 |



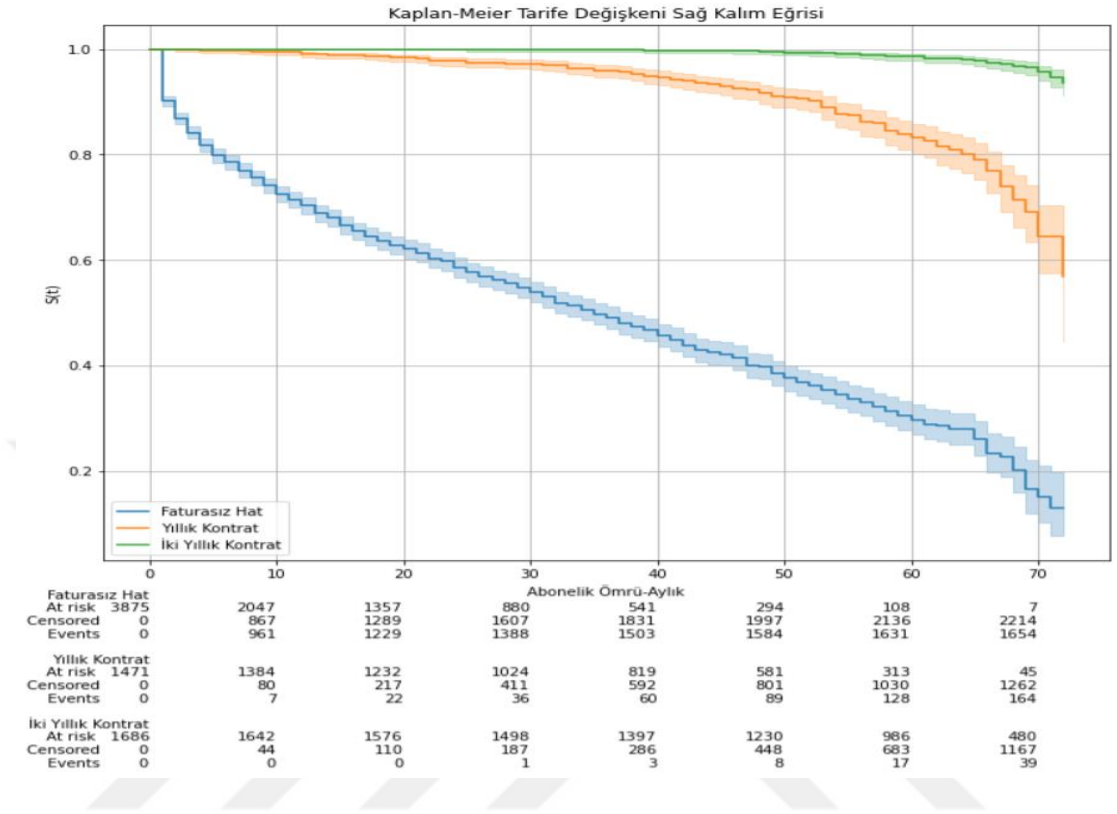

**Şekil 5.3** KM sağkalım eğrisi

Müşterilerin 10 ay ve daha fazla yaşama ihtimali

$$S(10) = P(10 \geq t) \tag{5.1}$$

$$= S(9) * \left(1 - \frac{968}{5073}\right) = 0{,}85 \tag{5.2}$$

70 ay ve daha fazla yaşama ihtimali tablo 5.2'ye göre 0,61'dir. Telekom firması uzun süreli sadık abone kitlesine sahip olmak istiyorsa bu olasılık değerlerini nasıl yukarı çekebilirim sorusuna cevap arayabilir.




**Şekil 5.4** KM tarife sağkalım eğrisi

Bazı açıklayıcı değişkenler için KM sağkalım grafikleri çizilip, değişken içerisindeki grupların sağkalım olasılığı üzerine etkileri incelenmiştir. İlk açıklayıcı değişken olan tarife, üç grup olduğundan çok değişkenli LRT uygulanmış, sonuçları tablo 5.3'de verilmiştir. Tarifelerin sağkalım olasılığına farklı etkileri vardır ($p<0,005$). Şekil 5.4'de görüldüğü gibi sağkalım olasılığı en yüksek tarife iki yıllık kontrattır. Bu tarifenin müşterilik ömrünü uzatma açısından en başarılı tarife olduğu gözükmektedir.

Faturasız tarifeli abonelerin abonelik ömürleri medyanı 35 aydır. Eğrilerin paralel olmamasıda zamana bağlı bir değişken olabileceğini göstermektedir.

İkinci açıklayıcı değişkenimiz olan teknik destek için LRT uygulanmıştır. Tablo 5.4'de teknik destek ürününün sağkalım olasılığına farklı etkileri vardır. ($p<0,005$).

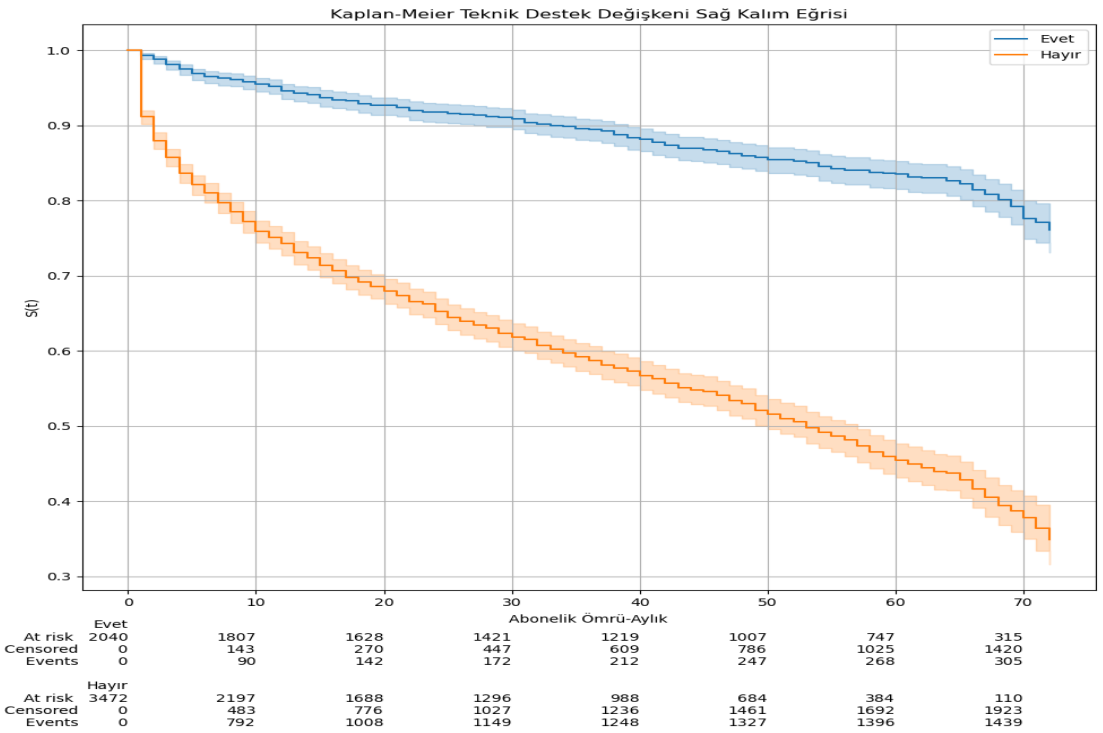
Şekil 5.5'da, teknik desteğe sahip müşterilerin sağkalım olasılığı, teknik desteğe sahip olmayan müşterilere göre daha yüksektir. Eğrilerin paralel olmasıda zamandan bağımsız bir değişken olabileceğini göstermektedir.

**Tablo 5.3** LRT tarife

| Testin Adı | Boş Dağılım | Serbestlik Derecesi | Test İstatistiği | p |
|---|---|---|---|---|
| Çok Değişkenli Rank | Ki-Kare | 2 | 2352,87 | <0,005 |

**Tablo 5.4** LRT teknik destek

| Testin Adı | Boş Dağılım | Serbestlik Derecesi | Test İstatistiği | p |
|---|---|---|---|---|
| Çok Değişkenli Rank | Ki-Kare | 1 | 639,35 | <0,005 |




**Şekil 5.5** KM teknik destek sağkalım eğrisi

## 5.3 Ön Hazırlık

Yarı parametrik ve parametrik modellere başvurmadan önce yapılan ön hazırlık çalışmaları aşağıdaki şekildedir.

### 5.3.1 Kukla Değişken Ataması

Modelleri uygulamadan önce açıklayıcı değişkenlere kukla değişken ataması yapılmıştır. Kukla değişken atamasından sonra referans değişkene göre tehlike fonksiyonu ve yaşam süreleri etkileri ölçülmek istendiğinden referans değişken modele dahil edilmemiştir. Tarife değişkeni üzerinden örnek verilecek olursa, üç tarife olduğundan, sırasıyla tarife faturasız hat, tarife yıllık kontrat ve tarife iki yıllık kontrat olacak şekilde üç değişken oluşturulmuş ve referans değişken olarak seçilen faturasız hat modellemeden çıkarılmıştır.

### 5.3.2 Çoklu Bağımlılık ve VIF (Variance Inflation Factor)

Veri setinde açıklayıcı değişkenler arasında çoklu bağımlılık tablo 5.6'ya göre mevcuttur. Çoklu bağımlılık regresyon katsayılarında büyük ölçekte değişimlere sebebiyet vermektedir. VIF değeri kullanılarak veride korelasyona sahip değişkenler modellemeden çıkarılmışlardır. VIF değeri aşağıdaki şekilde hesaplanır:

$$VIF = \frac{1}{1 - R^2} \tag{5.3}$$

**Tablo 5.5** VIF bağlılık seviyesi

| VIF | Bağımlılık Seviyesi |
|---|---|
| VIF>10 | Yüksek |
| 10≥VIF>5 | Orta |
| 5≥VIF | Zayıf veya sıfır |

Burada $R^2$ determinasyon katsayısıdır. $R^2$ ne kadar 1'e yakınsarsa, değişkenler arasında korelasyon artar, VIF değeride yükselir. VIF değeri, 1'den başlar ve üst sınıra sahip değildir. Tablo 5.5'de VIF değeri bağımlılık seviyesini özetlemiştir.

Tablo 5.6'da listelenen, VIF değeri 10'dan büyük olan açıklayıcı değişkenler modelden çıkarılmıştır.

**Tablo 5.6** Değişken VIF değerleri

| **Değişken** | **VIF Değeri** |
|---|---|
| Aylık Fatura | 29,1 |
| Kesintisiz TV_internet hizmeti yok | sonsuz |
| Kesintisiz TV_internet hizmeti yok | sonsuz |
| Cihaz Koruma_internet hizmeti yok | sonsuz |
| Kesintisiz Film_internet hizmeti yok | sonsuz |
| Online Güvenlik_internet hizmeti yok | sonsuz |
| İnternet Hizmeti_Hayır | sonsuz |
| Birden Fazla Hattı Var mı_telefon servisi yok | sonsuz |
| Sabit Telefon_Hayır | sonsuz |
| Teknik Destek_internet hizmeti yok | sonsuz |
| Online Yedekleme_internet hizmeti yok | sonsuz |

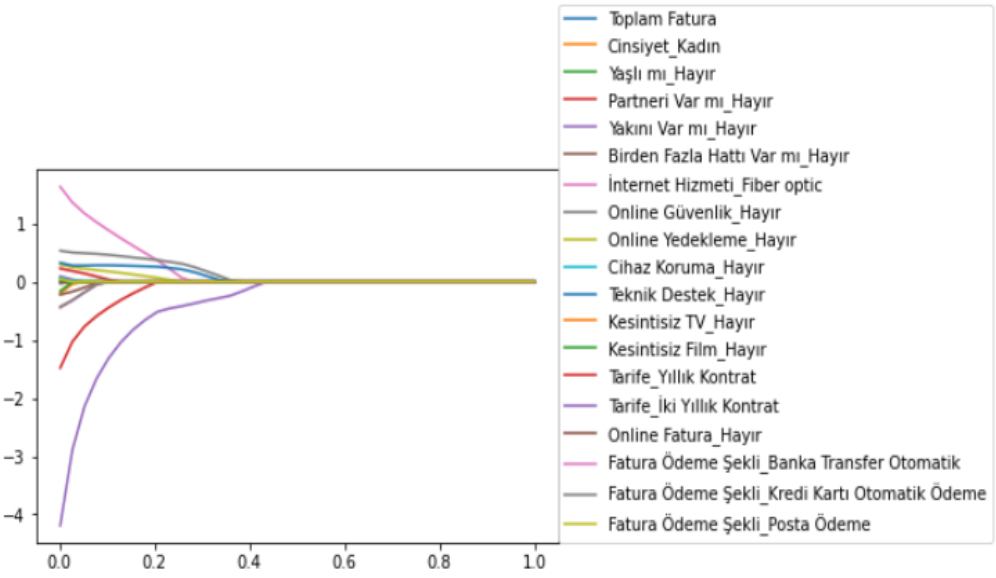
## 5.4 Yarı Parametrik Cox Regresyon Modeli

### 5.4.1 Panelizer Katsayısı

CRMO'de karşımıza çıkan diğer bir durum regresyon katsayılarının stabilizasyonudur. Stabilize edilmiş regresyon katsayıları sonuçların güvenilir olması için gereklidir. Penalizer terimi eklenerek bu stabilizasyon sağlanabilir [31]. Terim aşağıdaki şekilde elde edilir.

$$\frac{1}{2} penalizer\left((1 - l1\ Oranı) \cdot \|\beta\|_2^2 + l1\ Oranı \cdot \|\beta\|_1\right) \qquad (5.4)$$

L1 Oranı için herhangi bir seçim yapılmayarak sadece penalizer değeri 0,4 seçilmiştir. Şekil 5.6'de CoxPHFitter kullanılarak tahmin edilen parametrelerde oynaklık 0,4 değerine yakın noktada son bulmuştur ve 0,4 değerinde stabilizasyonun sağlandığı görülmüştür. Bu nedenle penalizer değeri 0,4 seçilmiştir.




**Şekil 5.6** Penalizer seçim grafiği

Python lifelines paketinde CoxPHFitter kullanılarak CRMO'a başvurulmuştur. CRMO sonuçları tablo 5.7'de paylaşılmıştır.

P-değeri <0,05'den küçük açıklayıcı değişkenlerin tehlike üzerine etkisi mevcuttur. Sürekli değişkenler için katsayı (coef) değeri pozitif ise referans gruba göre mevcut değişkende her bir birim artış daha riskli yani tehlike daha yüksek,

katsayı değeri negatif ise her bir birim artışta risk daha az yani tehlike daha düşüktür.

Kategorik değişkenler için katsayı (coef β değeri) pozitif ise bu kategorinin referans gruba göre daha riskli olduğu, katsayı değeri negatif ise bu kategorinin referans gruba göre daha az riskli olduğunu gösterir.

Tablo 5.7'ye göre kategorik değişkenlerden Partneri Var mı Hayır değişkeninin tehlikeye etkisi mevcuttur. Mobil operatörde partnerininin müşteri olmaması kayıp riskini ortalamada %25 artırmaktadır. Tarife değişkeninin tehlikeye etkisi mevcut olup, faturasız hatta göre, iki yıllık kontrat kayıp riskini ortalamada %41 düşürmüş, yıllık kontratta ortalamada %31 düşürmüştür. Sürekli değişkenlerden toplam fatura değişkeninin tehlikeye etkisi mevcuttur. Sürekli bir değişken olduğundan ayrıklaştıma yapılarak kontrol edilmesi daha doğru olacağında bir sonraki bölümde başvurulan tabakalandırılmış CRMO'da değişik ayrıklaştıma yöntemleri denenerek tehlikeye olan etkisi ölçülmüştür.

Burada şerh düşülmesi gereken nokta OTV şartı sağlamayan değişkenlerin etkisi CRMO'da ölçülemediğinden yukarıda yapılan yorumlar bu değişkenler için geçersizdir. OTV şartı kontrol edilip bu şartı sağlamayan değişkenlere uygun yöntemler seçilip, yapılan yorumlar daha doğru olacaktır. OTV şartı bir sonraki konu başlığında kontrol edilmiş ve OTV şartına göre yeni modeller belirlenmiştir.

Uygulanan CRMO çalışmamızda yarı parametrik modeller arasında eşik başarısı için referans model olup, tablo 5.10'de $CRMO_{REFERANS}$ modeli altında performans sonuçları paylaşılmıştır.

**Tablo 5.7** CRMO sonuçları

| Değişken | $\widehat{\boldsymbol{\beta}}$ | Tehlike Oranı | p-değeri |
|---|---|---|---|
| Toplam Fatura | 0,000 | 1,000 | <0,0005 |
| Cinsiyet Erkek | Referans Grup | | |

**Tablo 5.7** CRMO sonuçları (devamı)

| | | | |
|---|---|---|---|
| Cinsiyet Kadın | 0,02 | 1,021 | 0,484 |
| Yaşlı mı Evet | Referans Grup | | |
| Yaşlı mı Hayır | -0,09 | 0,913 | 0,017 |
| Partneri Var mı Evet | Referans Grup | | |
| Partneri Var mı Hayır | 0,229 | 1,257 | <0,0005 |
| Yakını Var mı Evet | Referans Grup | | |
| Yakını Var mı Hayır | 0,149 | 1,160 | <0,0005 |
| Birden Fazla Hattı Var mı Evet | Referans Grup | | |
| Birden Fazla Hattı Var mı Hayır | 0,105 | 1,111 | <0,0005 |
| İnternet Hizmeti DSL | Referans Grup | | |
| İnternet Hizmeti Fiber Optic | 0,288 | 1,334 | <0,0005 |
| Online Güvenlik Evet | Referans Grup | | |
| Online Güvenlik Hayır | 0,343 | 1,409 | <0,0005 |
| Online Yedekleme Evet | Referans Grup | | |
| Online Yedekleme Hayır | 0,292 | 1,339 | <0,0005 |
| Cihaz Koruma Evet | Referans Grup | | |

**Tablo 5.7** CRMO sonuçları (devamı)

| | | | |
|---|---|---|---|
| Cihaz Koruma Hayır | 0,232 | 1,261 | <0,0005 |
| Teknik Destek Evet | Referans Grup | | |
| Teknik Destek Hayır | 0,307 | 1,359 | <0,0005 |
| Kesintisiz TV Evet | Referans Grup | | |
| Kesintisiz TV Hayır | 0,076 | 1,079 | 0,013 |
| Kesintisiz Film Evet | Referans Grup | | |
| Kesintisiz Film Hayır | 0,089 | 1,093 | 0,004 |
| Online Fatura Evet | Referans Grup | | |
| Online Fatura Hayır | -0,148 | 0,863 | <0,0005 |
| Tarife Faturasız Hat | Referans Grup | | |
| Tarife Yıllık Kontrat | -0,367 | 0,693 | <0,0005 |
| Tarife İki Yıllık Kontrat | -0,52 | 0,594 | <0,0005 |
| Fatura Ödeme Şekli Elektronik Ödeme | Referans Grup | | |
| Fatura Ödeme Şekli Banka Transfer Otomatik | -0,248 | 0,78 | <0,0005 |
| Fatura Ödeme Şekli Kredi Kartı Otomatik | -0,261 | 0,77 | <0,0005 |

**Tablo 5.7** CRMO sonuçları (devamı)

| | | | |
|---|---|---|---|
| Fatura Ödeme Şekli Posta Ödeme | -0,005 | 0,995 | 0,888 |

### 5.4.2 Oransal Tehlike Varsayımı Kontrolü

OTV, CRMO için bir ön şarttır. OTV sağlamayan değişkenler için tehlike oranı etki analizi yapmak CRMO modeli için geçerli değildir. OTV sağlayan değişkenlerin tehlikeye olan etkisi ölçülebilir ve yorumlanabilir. OTV sağlayan değişkenleri belirlemek için pythonda log-rank testine başvurulmuştur. Tablo 5.8'de Ki-Kare uyum iyiliği testi sonuçlarına göre p-değerleri <0,05 olan değişkenler için sağlanmamaktadır. Bu değişkenler sırasıyla, Birden Fazla Hattı Var mı, Fatura Ödeme Şekli, Kesintisiz Film, Kesintisiz TV, Tarife, Toplam Fatura, Partneri var mı, Internet Hizmeti değişkenleridir.

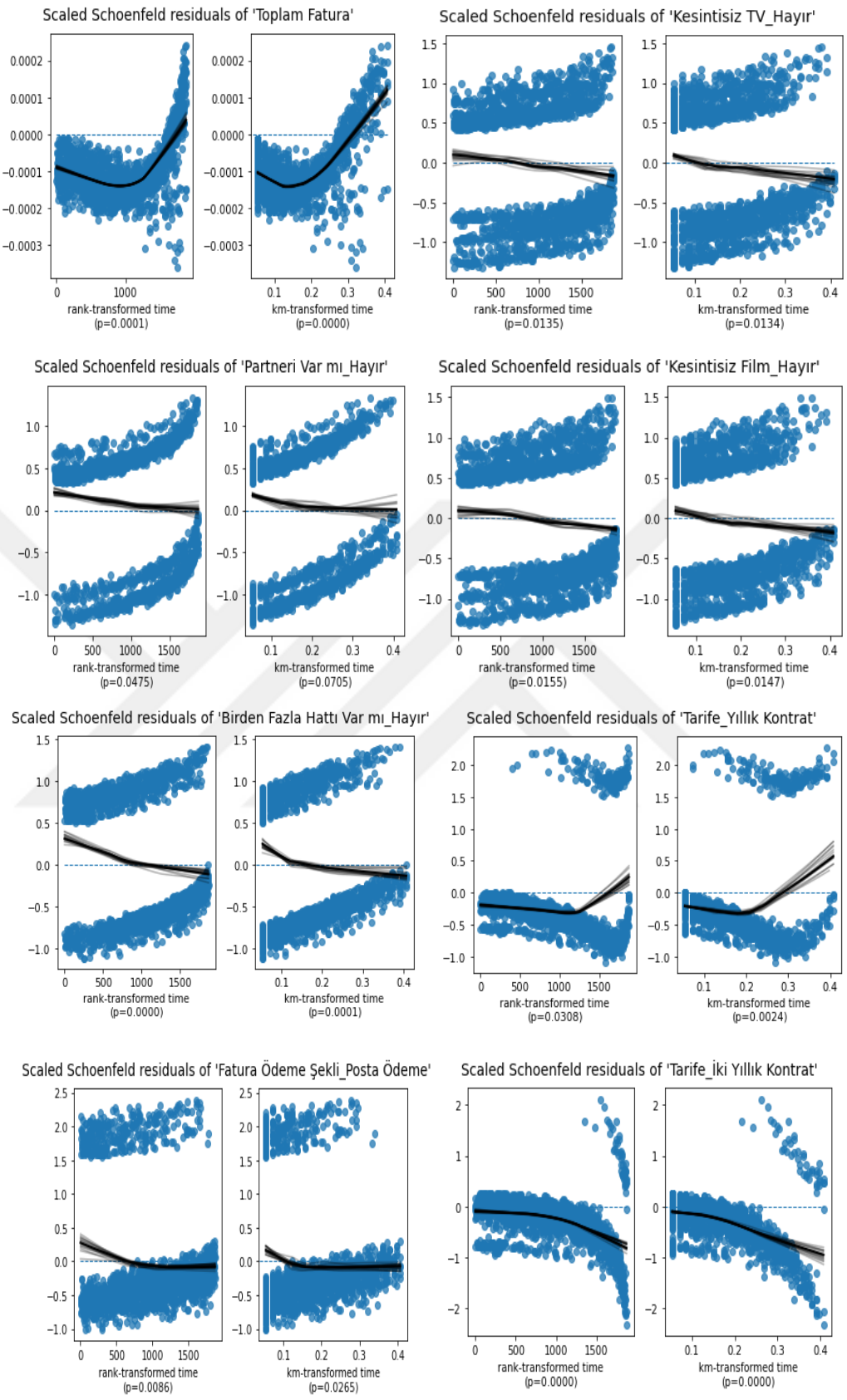
**Tablo 5.8** Log-rank testi sonuçları

| Değişken | Test istatistik | p |
|---|---|---|
| Birden Fazla Hattı Var mı Hayır | 16,47 | 0,005 |
| Cihaz Koruma Hayır | 0,16 | 0,69 |
| Cinsiyet Kadın | 0,21 | 0,65 |
| Fatura Ödeme Şekli Banka Transfer Otomatik | 0,04 | 0,84 |
| Fatura Ödeme Şekli Kredi Kartı Otomatik Ödeme | 0,49 | 0,48 |
| Fatura Ödeme Şekli Posta Ödeme | 6,91 | 0,01 |
| Kesintisiz Film Hayır | 5,86 | 0,02 |

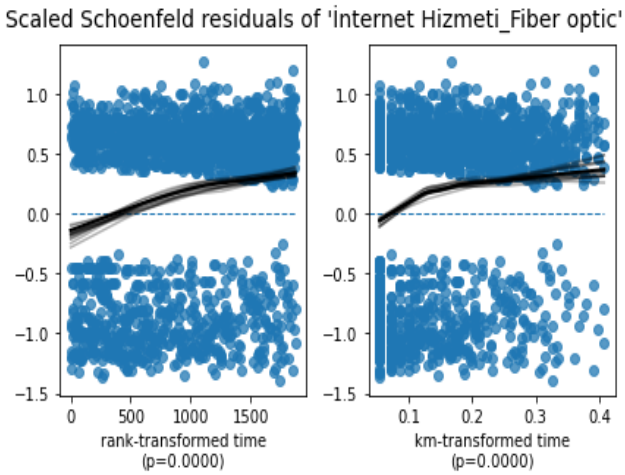
**Tablo 5.8** Log-rank testi sonuçları (devamı)

| Değişken | Test istatistik | p |
|---|---|---|
| Kesintisiz TV Hayır | 6,11 | 0,01 |
| Online Fatura Hayır | 2,58 | 0,11 |
| Online Güvenlik Hayır | 1,28 | 0,26 |
| Online Yedekleme Hayır | 0,18 | 0,68 |
| Partneri Var mı Hayır | 3,93 | 0,05 |
| Tarife Yıllık Kontrat | 4,66 | 0,03 |
| Tarife İki Yıllık Kontrat | 27,71 | 0,005 |
| Teknik Destek Hayır | 0,64 | 0,42 |
| Toplam Fatura | 15,77 | 0,005 |
| Yakını Var mı Hayır | 0,01 | 0,91 |
| Yaşlı mı Hayır | 2,44 | 0,12 |
| İnternet Hizmeti Fiber Optic | 19,76 | 0,005 |

OTV sağlamayan değişkenlerin, Schoenfeld Artıklarının başarısızlık zamanına karşı şekil 5.7 'de grafikleride incelenecek olursa, yaşam süresinin log değerine göre dağınık bir görünüm sergilemediği görülebilir.

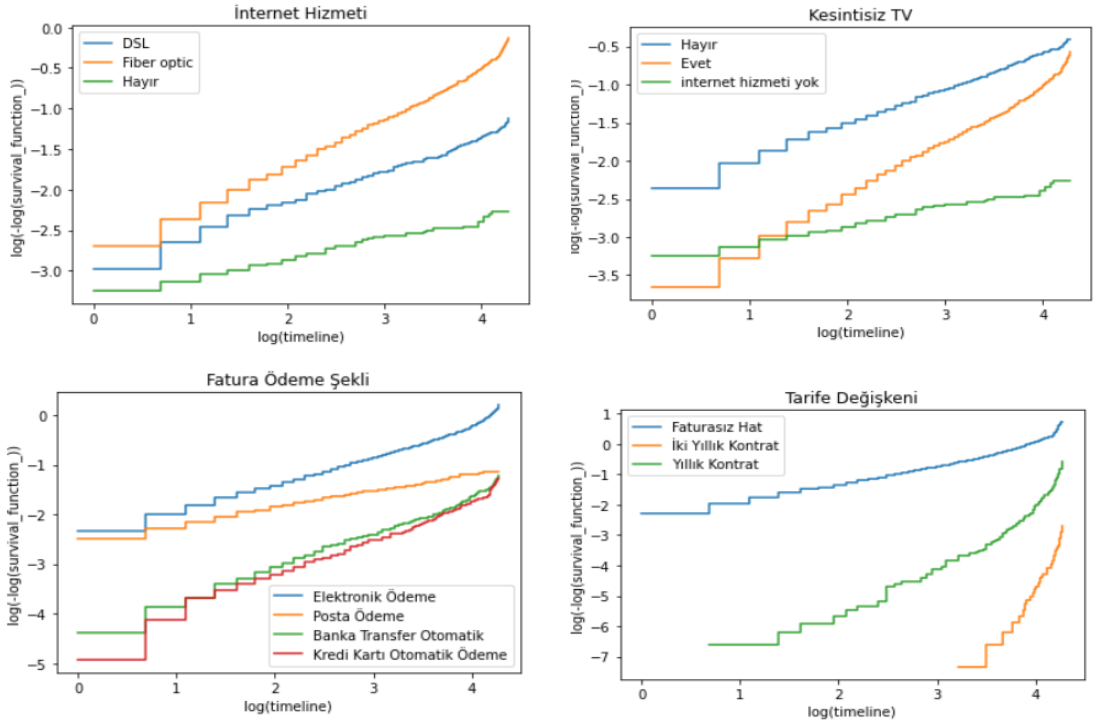



**Şekil 5.7** OTV sağlamayan değişkenlerin Schoenfeld artıkları grafikleri

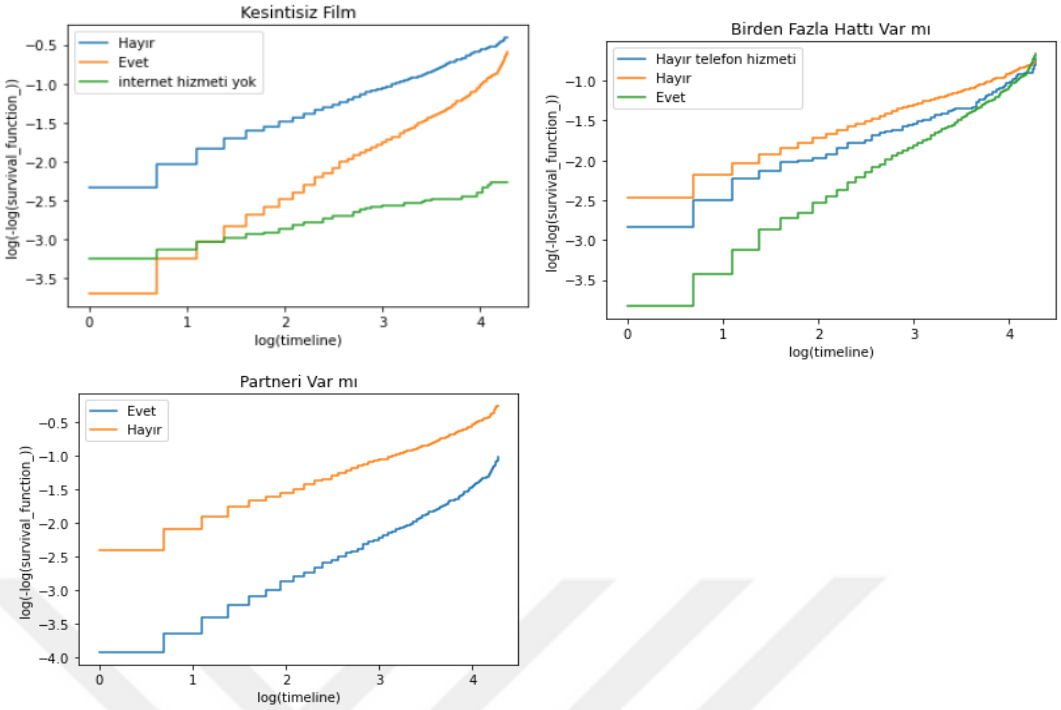



**Şekil 5.7** OTV sağlamayan değişkenlerin Schoenfeld artıkları grafikleri (devamı)

Şekil 5.8'da OTV sağlamayan değişkenler için diğer bir yöntem olan (ln-lnS(t)) grafiğinin log t ye göre çizilmiş grafikleri incelenecek olursa, eğriller arasında paralellik olmadığı ve zamanla tehlike oranının değiştiği görülmektedir.




**Şekil 5.8** OTV sağlamayan değişkenlerin log t ye göre (ln-lnS(t)) grafikleri




**Şekil 5.8** OTV sağlamayan değişkenlerin log t ye göre (ln-lnS(t)) grafikleri (devamı)

CRMO'da OTV kontrolü ile OTV'yi sağlamayan değişkenler olduğundan, literatürde OTV ön şartı sağlanmadığı durumlarda önerilen modellere başvurulmuştur. Sırasıyla modeller tanımlarıyla aşağıdaki şekilde verilmiştir.

### 5.4.3 Tabakalandırılmış CRMO

Bu başlıkta 7 model denenmiştir. Veri setimizde mevcut olan OTV sağlamayan kategorik değişkenler için tabakalandırma yapılmış olup, tek sürekli değişken olan toplam fatura, değişik ayrıklaştırma yöntemleri ile ayrıklaştırma yapılarak, kategorik eşdeğerine dönüştürülüp daha sonra tabakalandırma yapılarak modele dahil edilmiştir. Toplam fatura için yapılan ayrıklaştırma yöntemleri aşağıda verilmiş olup, model performans sonuçları tablo 5.10'de paylaşılmıştır.

$CRMO_{TABAKALANDIRILMIŞ,4YA}$: Toplam fatura yüzdelik olarak [0 %25 %50 %75 %100] dilimlerde ayrıklaştırma yapılarak 4 eşit parçaya bölünmüş ve kategorik eşdeğerine dönüştürülüp tabakalandırılmıştır.

$CRMO_{TABAKALANDIRILMIŞ,5YA}$:Toplam fatura yüzdelik olarak [0 %20 %40 %60 %80 %100] dilimlerde ayrıklaştırma yapılarak 5 eşit parçaya bölünmüş ve kategorik eşdeğerine dönüştürülüp tabakalandırılmıştır.

$CRMO_{TABAKALANDIRILMIŞ,10YA}$:Toplam fatura yüzdelik olarak [0 %10 %20 %30 %40 %50 %60 %70 %80 %90 %100] dilimlerde ayrıklaştırma yapılarak 10 eşit parçaya bölünmüş ve kategorik eşdeğerine dönüştürülüp tabakalandırılmıştır.

$CRMO_{TABAKALANDIRILMIŞ,2A}$: Toplam fatura değeri {[0 ,4500),[4500,9000]} aralıklarında ayrıklaştırma yapılarak 2 eşit parçaya bölünmüş ve kategorik eşdeğerine dönüştürülüp tabakalandırılmıştır.

$CRMO_{TABAKALANDIRILMIŞ,3A}$: Toplam fatura değeri {[0 ,3000),[3000,6000),[6000,9000)} aralıklarında ayrıklaştırma yapılarak 3 eşit parçaya bölünmüş ve kategorik eşdeğerine dönüştürülüp tabakalandırılmıştır.

$CRMO_{TABAKALANDIRILMIŞ,9A}$: Toplam fatura değeri {[0 ,1000),[1000,2000),[2000,3000) ,...,[8000,9000]} aralıklarında ayrıklaştırma yapılarak 9 eşit parçaya bölünmüş ve kategorik eşdeğerine dönüştürülüp tabakalandırılmıştır.

$CRMO_{TABAKALANDIRILMIŞ,18A}$: Toplam fatura değeri {[0 ,500),[500,1000),[1000,1500) ,...,[8500,9000]} aralıklarında ayrıklaştırma yapılarak 18 eşit parçaya bölünmüş ve kategorik eşdeğerine dönüştürülüp tabakalandırılmıştır.

### 5.4.4 Genişletilmiş CRMO

Modelimizde OTV sağlamayan değişkenler, zamana bağlı değişkenler olup bu değişkenlerin zamanla olan etkileşimleri modele dahil edilmiştir. Literatürde önerilen g(t) için t ve log t seçilmiş olup, Genişletilmiş CRMO başlığında 2 model denenmiştir. Denenen modellerin tanımları aşağıda verilmiş olup, sonuçlar tablo 5.10'de paylaşılmıştır.

$CRMO_{GENİŞLETİLMİŞ,t}$:OTV varsayımı sağlamayan değişkenler t ile çarpılıp modele dahil edilmiştir.

$CRMO_{GENİŞLETİLMİŞ,logt}$ :OTV varsayımı sağlamayan değişkenler log(t) ile çarpılıp modele dahil edilmiştir.

## 5.5 PRCRMO ve KSCRMO

Çalışmadaki süre ay olarak 1,2,3,...,69,70,71,72 olduğundan, sırayla önerilen düğüm yüzdelerine karşılık gelen t süreleri hesaplanmış ve tablo 5.9'da gösterilmiştir. Tablo 5.9'daki düğümler sadece PRCRMO için kullanılmıştır. Python Lifeline paketi KSCRMO'da düğüm sayısına göre t'yi eşit parçalara ayırmaktadır. Bu başlık altında toplamda altı model uygulanmıştır. Modellerin tanımları aşağıda olup, model performans sonuçları tablo 5.10'da paylaşılmıştır.

$PRCRMO_{3D\text{düğüm}}$: Parçalı Regresyon modeli olup, düğüm adeti 3 olacak şekilde düğüm lokasyonları tablo 5.9'da verilmiştir

$PRCRMO_{4D\text{düğüm}}$: Parçalı Regresyon modeli olup, düğüm adeti 4 olacak şekilde düğüm lokasyonları tablo 5.9'da verilmiştir

$PRCRMO_{5D\text{düğüm}}$: Parçalı Regresyon modeli olup, düğüm adeti 5 olacak şekilde düğüm lokasyonları tablo 5.9'da verilmiştir

$KSCRMO_{3D\text{düğüm}}$: Kübik Spline modeli olup, düğüm adeti 3 olacak şekilde, t ekseni üç eşit parçaya bölünmüştür.

$KSCRMO_{4D\text{düğüm}}$: Kübik Spline modeli olup, düğüm adeti 4 olacak şekilde, t ekseni dört eşit parçaya bölünmüştür.

$KSCRMO_{5D\text{düğüm}}$: Kübik Spline modeli olup, düğüm adeti 5 olacak şekilde, t ekseni beş eşit parçaya bölünmüştür.

**Tablo 5.9** Düğüm sayısı ve düğüm yüzdeleri

| Düğüm Sayısı | Düğüm Yüzdesi | Yüzdeye Karşılık Gelen t |
|---|---|---|
| 3 | 10 | 7,3 |
| | 50 | 36,5 |
| | 90 | 65,7 |

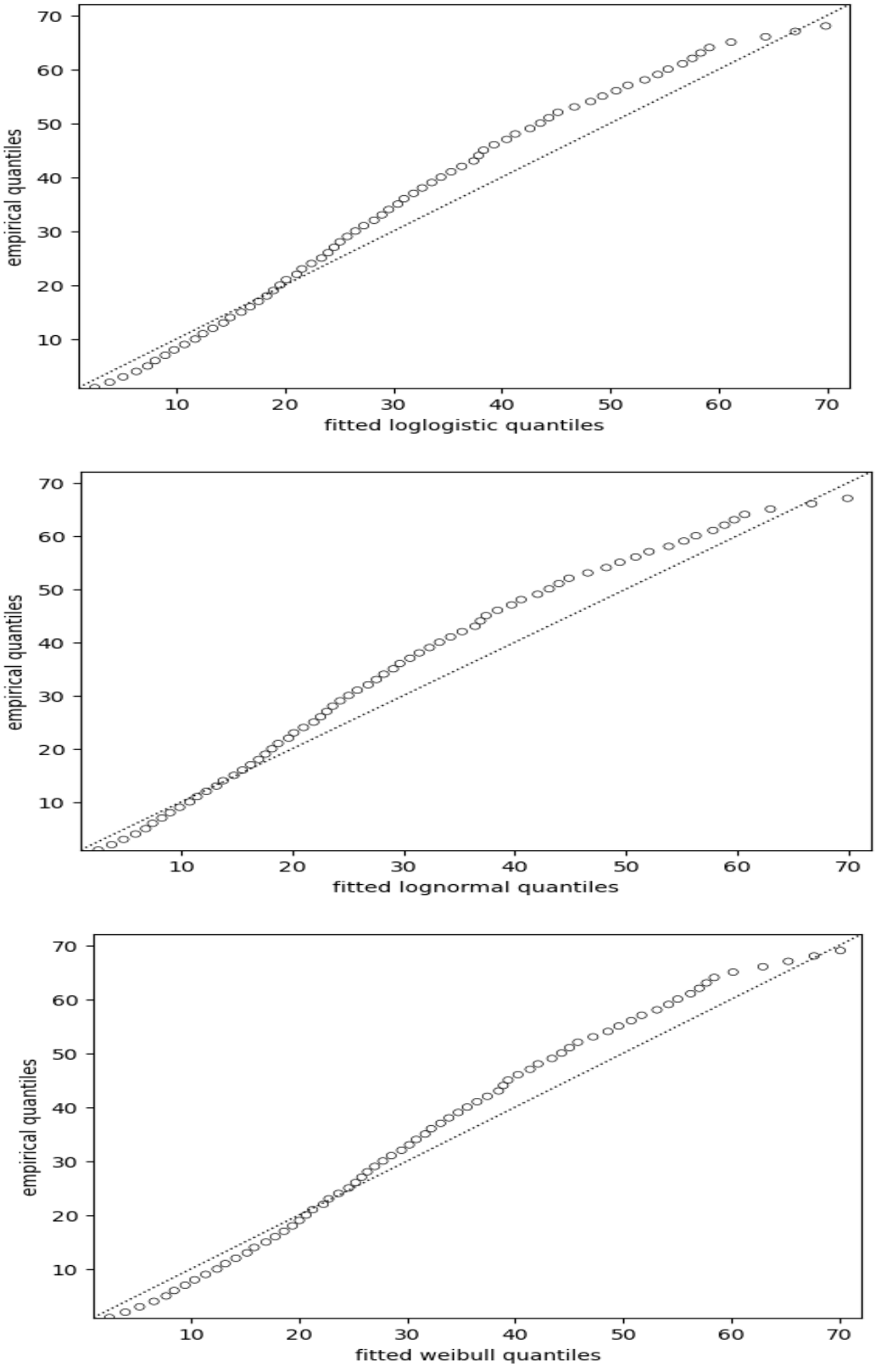
**Tablo 5.9** Düğüm sayısı ve düğüm yüzdeleri (devamı)

| | | |
|---|---|---|
| 4 | 5<br>35<br>65<br>96 | 3,65<br>25,55<br>47,45<br>69,35 |
| 5 | 5<br>27,5<br>50<br>72,5<br>95 | 3,65<br>20,074<br>36,5<br>53,925<br>69,35 |

## 5.6 HBSM

HBSM modelleri olan Weibull, Log-Lojistik, Log-Normal dağılımları uygulanmıştır. Python Lifelines paketinde WeibullAFTFitter, LogLogisticAFTFitter, LogNormalAFTFitter kodları kullanılmıştır.

HBSM performansları tablo 5.11'de verilmiş olup, görsel olarak şekil 5.9 'de QQ grafiği kullanılmıştır. QQ grafiği, bir dağılım grafiği olup bir noktanın konumunu tanımlayan her koordinat çifti, bir veri değerinden ve bu veri değeri için uygun dağılımın nicel fonksiyonundan türetilen ilgili tahminden oluşur. Tahmin değeri ile gerçek değer ne kadar örtüşüyorsa yani doğru üzerinde noktalar sıralanıyorsa bu model iyi bir modeldir. Şekil 5.9'da Log-Lojistik dağılımında noktalar diğer dağılımlara göre doğru üzerinde daha fazla yoğunlaştığı görülmektedir.




**Şekil 5.9** HBSM qq grafikleri

**Tablo 5.10** Yarı parametrik modellerin performansları

| Model | -2LogL | Serbestlik Derecesi | AIC |
|---|---|---|---|
| $CRMO_{REFERANS}$ | 29.161,93 | 19 | 29.199,93 |
| $CRMO_{TABAKALANDIRILMIŞ,4YA}$ | 10.131,92 | 10 | 10.151,92 |
| $CRMO_{TABAKALANDIRILMIŞ,5YA}$ | 10.786,56 | 10 | 10.806,56 |
| $CRMO_{TABAKALANDIRILMIŞ,10YA}$ | 8.062,7 | 10 | 8.072,7 |
| $CRMO_{TABAKALANDIRILMIŞ,2A}$ | 13.995,39 | 10 | 13.997,39 |
| $CRMO_{TABAKALANDIRILMIŞ,3A}$ | 13.174,61 | 10 | 13.194,12 |
| $CRMO_{TABAKALANDIRILMIŞ,9A}$ | 10.485,88 | 10 | 10.505,88 |
| $CRMO_{TABAKALANDIRILMIŞ,18A}$ | 8.626,49 | 10 | 8.646,49 |
| $CRMO_{GENİŞLETİLMİŞ,t}$ | 28.405,7 | 28 | 28.378,7 |
| $CRMO_{GENİŞLETİLMİŞ,\log t}$ | 28.482,92 | 28 | 28.408,59 |
| $PRCRMO_{3\ Düğüm}$ | 18.836,55 | 21 | 18.882,5 |
| $PRCRMO_{4\ Düğüm}$ | 18.842,67 | 21 | 18.890,67 |
| $PRCRMO_{5\ Düğüm}$ | 18.846,13 | 21 | 18.896,13 |
| $KSCRMO_{3\ Düğüm}$ | 19.629,99 | 21 | 19.640,34 |
| $KSCRMO_{4\ Düğüm}$ | 19.623,9 | 21 | 19.671,9 |

**Tablo 5.10** Yarı parametrik modellerin performansları (devamı)

| | | | |
|---|---|---|---|
| $KSCRMO_{5\ Düğüm}$ | 19.617,45 | 21 | 19.667,45 |

**Tablo 5.11** Parametrik modellerin performansları

| **Model** | **-2LogL** | **Serbestlik Derecesi** | **AIC** |
|---|---|---|---|
| Log-Lojistik | 14.234,99 | 19 | 14.282,99 |
| Log-Normal | 14.277,30 | 19 | 14.325,34 |
| Weibull | 14.418,57 | 19 | 14.466,57 |

**Tablo 5.12** Log-Lojistik HBSM çıktısı

| **Değişken** | $\widehat{\boldsymbol{\beta}}$ | **Hızlandırma Faktörü** | **p-değeri** |
|---|---|---|---|
| Cinsiyet Erkek | Referans Grup | | |
| Cinsiyet Kadın | -0,027 | 0,973 | 0,226 |
| Yaşlı mı Evet | Referans Grup | | |
| Yaşlı mı Hayır | 0,003 | 1,003 | 0,924 |
| Partneri Var mı Evet | Referans Grup | | |
| Partneri Var mı Hayır | -0,086 | 0,917 | <0,001 |
| Yakını Var mı Evet | Referans Grup | | |

**Tablo 5.12** Log-Lojistik HBSM çıktısı (devamı)

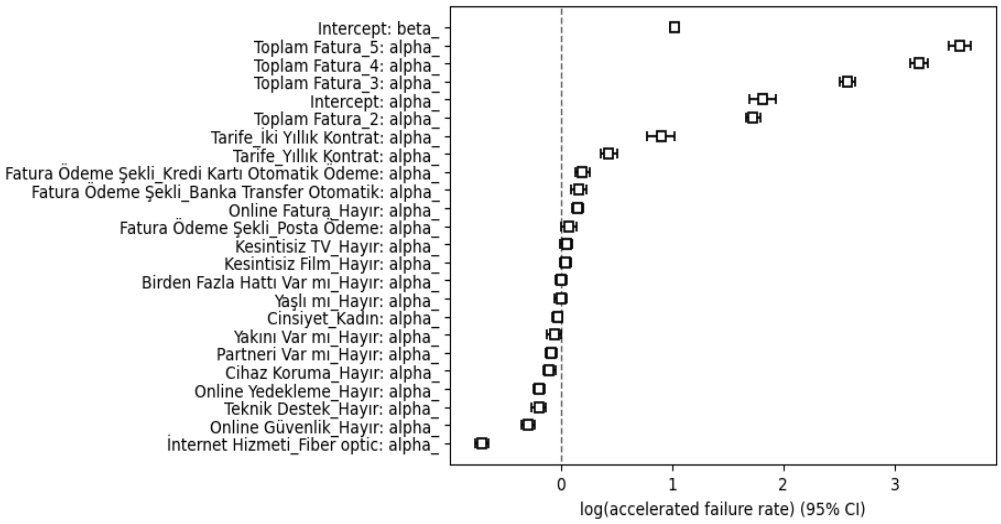
| | | | |
|---|---|---|---|
| Yakını Var mı Hayır | -0,059 | 0,942 | 0,064 |
| Birden Fazla Hattı Var mı Evet | Referans Grup | | |
| Birden Fazla Hattı Var mı Hayır | 0,005 | 1,005 | 0,83 |
| İnternet Hizmeti DSL | Referans Grup | | |
| İnternet Hizmeti Fiber Optic | -0,709 | 0,492 | <0,0005 |
| Online Güvenlik Evet | Referans Grup | | |
| Online Güvenlik Hayır | -0,291 | 0,747 | <0,0005 |
| Online Yedekleme Evet | Referans Grup | | |
| Online Yedekleme Hayır | -0,191 | 0,826 | <0,0005 |
| Cihaz Koruma Evet | Referans Grup | | |
| Cihaz Koruma Hayır | -0,1 | 0,905 | <0,0005 |
| Teknik Destek Evet | Referans Grup | | |
| Teknik Destek Hayır | -0,196 | 0,822 | <0,0005 |
| Kesintisiz TV Evet | Referans Grup | | |
| Kesintisiz TV Hayır | 0,05 | 1,051 | 0,065 |
| Kesintisiz Film Evet | Referans Grup | | |

**Tablo 5.12** Log-Lojistik HBSM çıktısı (devamı)

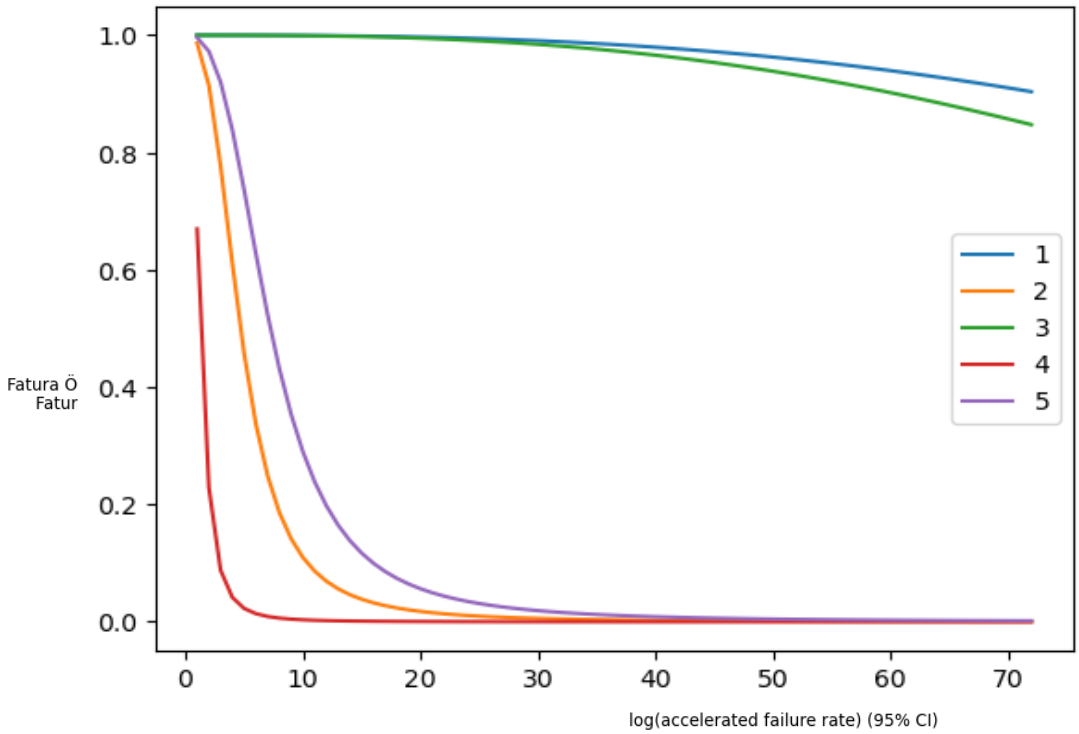
| | | | |
|---|---|---|---|
| Kesintisiz Film Hayır | 0,046 | 1,047 | 0,092 |
| Online Fatura Evet | Referans Grup | | |
| Online Fatura Hayır | 0,152 | 1,164 | <0,0005 |
| Tarife Faturasız | Referans Grup | | |
| Tarife Yıllık Kontrat | 0,433 | 1,543 | <0,0005 |
| Tarife İki Yıllık Kontrat | 0,901 | 2,461 | <0,0005 |
| Fatura Ödeme Şekli Elektronik Ödeme | Referans Grup | | |
| Fatura Ödeme Şekli Banka Transfer Otomatik | 0,162 | 1,176 | <0,0005 |
| Fatura Ödeme Şekli Kredi Kartı Otomatik Ödeme | 0,196 | 1,217 | <0,0005 |
| Fatura Ödeme Şekli Posta Ödeme | 0,074 | 1,077 | 0,034 |
| Toplam Fatura [0,265) | Referans Grup | | |
| Toplam Fatura [265,941) | 1,726 | 5,617 | <0,0005 |
| Toplam Fatura [941,2.045) | 2,574 | 13,12 | <0,0005 |
| Toplam Fatura [2.045,4.473) | 3,217 | 24,95 | <0,0005 |

**Tablo 5.12** Log-Lojistik HBSM çıktısı (devamı)

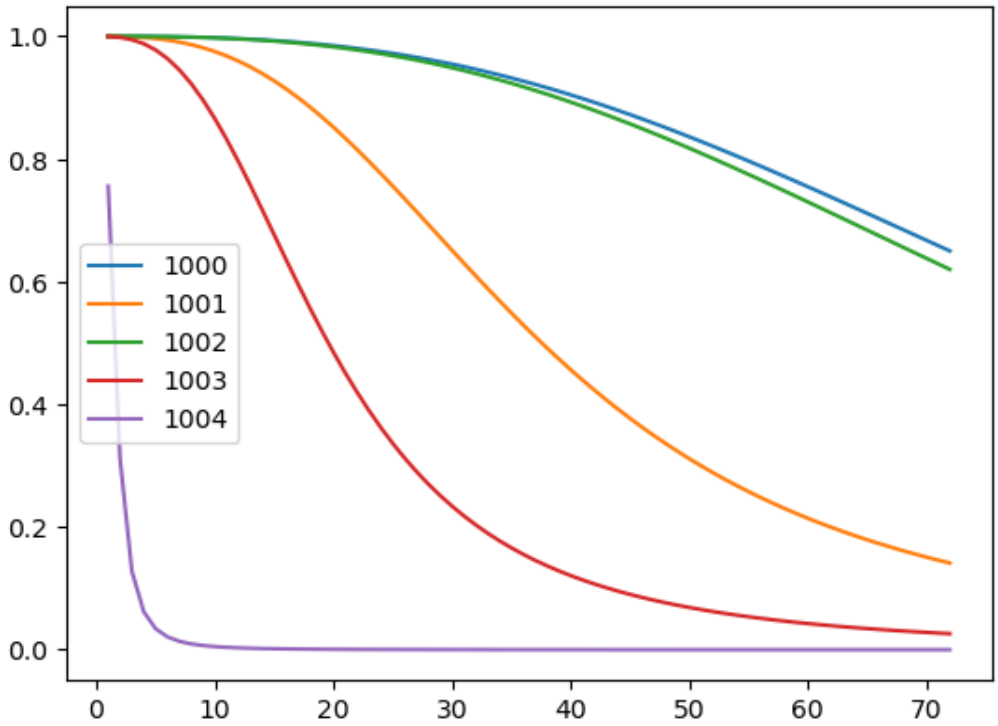
| | | | |
|---|---|---|---|
| Toplam Fatura [4.473,9.000] | 3,586 | 36,088 | <0,0005 |




**Şekil 5.10** Weibull için açıklayıcı değişkenlerin %95 güven aralığı




**Şekil 5.11** 1-5 müşteriler için Log-Lojistik sağkalım eğrisi tahmini




**Şekil 5.12** 1000-1004 müşteriler için Log-Lojistik sağkalım eğrisi tahmini

# 6
# SONUÇ VE ÖNERİLER

Müşteri kayıp veri setine toplam 19 adet model uygulanmış olup Akaike Information Criterion (AIC) ve -2log likelihood (-2LogL) değerleri model seçim metrikleri olarak kullanılmıştır.

İlk başvurulan modeller tablo 5.10'da listelenen yarı parametrik modellerdir. $CRMO_{REFERANS}$ baz alınırsa, $CRMO_{TABAKALANDIRILMIŞ,10YA}$ yarı parametrik modeller arasında en düşük AIC değeri ile en iyi performansa sahiptir. AIC değeri 29.161,93'den 8.062,7'ye düşmüştür. Model, en iyi performansa sahip olmasına rağmen OTV sağlamayan değişkenlerin etkisi ölçülemediğinden modelde bilgi kaybı yaşanmıştır. Bu nedenle parametre sayısı $CRMO_{REFERANS}$ 'da 19 iken $CRMO_{TABAKALANDIRILMIŞ,10YB}$ 'da 10'a düşmüştür. Toplam fatura değişkenini onluk yüzdelik dilimlere bölmek, eşit parçalara bölünüp ayrıklaştırma yapılan modellere göre çok daha iyi sonuç vermiştir. Ayrıklaştırma yapmak için yüzdesel değerlere bölmek eşit parçalara bölmekten daha avantajlı olduğu gözlemlenmiştir. Ayrıklaştırma sayısının artırılması model performansını iyileştirsede, ayrıklaştırma sayısının çok fazla artması aşırı öğrenmeye sebebiyet verebileceğinden daha fazla artırılmamıştır.

$CRMO_{GENİŞLETİLMİŞ,t}$, Genişletilmiş Modeller arasında en düşük AIC değeri ile en iyi performansa sahip modeldir. Modelde seçilen g(t) fonksiyonunun denklem 6.1'e göre seçilmesi model performansını $CRMO_{REFERANS}$ 'a göre ve Genişletilmiş Modeller arasında en iyi sonuçu vermesini sağlamıştır. AIC değeri 29.161,93'den 28.378,7'ye düşmüştür. $CRMO_{REFERANS}$ 'a göre daha başarılı olup, $CRMO_{TABAKALANDIRILMIŞ,10YB}$'a göre başarısı daha düşük çıkmıştır. Parametre sayısı, genişletilmiş CRMO'da 28'ye yükselmiş olup artan parametre sayısı modeli daha kompleks hale getirmiştir.

$$g(t) = t \tag{6.1}$$

PRCRMO ve KSCRMO arasında en iyi performansa sahip model, en düşük AIC değeri ile $PRCRMO_{3\ Düğüm}$'dir. $CRMO_{REFERANS}$'a görede daha başarılıdır. AIC değeri 29.161,93'den 18.836,55'e düşmüştür. PRCRMO arasında en iyi model $PRCRMO_{3\ Düğüm}$, KSCRMO arasında en iyi model $KSCRMO_{4\ Düğüm}$ 'dür

Telekom sektörü kayıp verileri için CRMO performans sonuçları, akciğer kanseri sağkalım verilerine uygulananan CRMO performansları ile paraleldir[41]. Tabakalandırılmış CRMO, Genişleştilmiş CRMO'ya göre daha iyi sonuçlar vermiştir. OTV sağlanmadığından dolayı literatürde önerilen modellerin hepsi referans modeline göre çok başarılı sonuçlar vermiştir.

OTV sağlanmadığından ve daha düzgün efektif parametre tahminlerinden dolayı, ikinci başvurulan modeller tablo 5.11'de listelenen Parametrik Modeller olmuştur. Parametrik Modeller arasında en düşük AIC değerine sahip Log-Lojistik, en iyi performans sergileyen modeldir.

Yarı parametrik modellerde hesaplanan kısmi olabilirlik fonksiyonundan ötürü parametrik modellerle kıyaslama yapılamaz. HBSM avantajlarından dolayı ve OTV varsayımı sağlanamadığı için bir paramerik model seçilecekse bu model en düşük AIC değerine sahip Log-Lojistik olmalıdır.

Log-Lojistik HBSM model çıktısı tablo 5.12'de verilmiştir. P değeri 0,05'den küçük ise açıklayıcı değişkenin müşterilik yaşam süresine (referans değişkene) etkisi vardır. CRMO'da OTV testinde, dokuz adet değişkenin OTV sağlanamadığı için tehlikeye olan etkisi ölçülememiş ve yorumlanamamıştı. Log-Lojistik HBSM'de OTV olmadığından herbir açıklayıcı değişkenin müşterilik ömrüne olan etkisi rahatlıkla ölçülebilir ve yorumlanabilir. Eğer $\hat{\beta}$ değeri pozitif ise müşterilik ömrü ortalamada hızlandırma faktörü kadar artar, eğer $\hat{\beta}$ değeri negatif ise müşterilik ömrü ortalamada hızlandırma faktörü değeri kadar azalır(referans gruba göre).

Tablo 5.12 'de etkileri analiz edilecek olursa, iki yıllık kontrat, faturasız hatlara göre müşterilik ömrünü ortalamada 2,46 kat artırırken, yıllık kontrat ortalamada 1,5 kat artırmaktadır. İnternet hizmeti fiber optik, DSL hizmetine göre müşterilik ömrünü ortalamada 0,49 kat düşürmektedir. Müşterinin partnerinin aynı operatörde müşteri olmayışı müşterilik ömrünü 0,9 kat düşürmektedir. Online güvenlik ve online yedekleme ürünlerininin olmayışı müşterilik ömürlerini

kısaltan değişkenler olarak dikkat çekmektedirler. Toplam fatura değişkeni [0 %20 %40 %60 %80 %100] yüzdelik dilimlerinde beş aralığa bölünmüş ve her bir aralığın referans gruba göre müşterilik ömrüne etkisi ölçülmüştür. Toplam Fatura [4.473,9.000] aralığı referans [0,265) aralığına göre müşterilik ömrünü ortalamada 36 kat artırdığı gözlemlenmiştir. Bunun sebebi yüksek fatura tutarına sahip müşterilerin kontratlı hatlarda oluşudur. Müşteriyi değişik kampanyalarla, Toplam Fatura aralığı [4.473,9.000) aralığına çekilirse müşteriler çok daha uzun süre elde tutulmuş olur ve operatöre uzun süreli nakit akışı sağlanır.

Şekil 5.10'de Log-Lojistik modeline göre açıklayıcı değişkenlerin %95 güven aralığı grafiği incelenecek olursa, güven aralığı 0 içeren değişkenlerin müşterilik ömrüne bir etkisi olmadığı, güven aralığı pozitif bölgede olan değişkenlerin müşterilik ömrünü artırdığı, negatif bölgedeki değişkenlerin müşterilik ömrünü azaltığı söylenebilir.

Çalışmamızın bir amacıda müşteri sağ kalım eğrisini tahminlemektir. Model parametre tahminlerinden sonra bu rahatlıkla başarılabilir. En iyi model olan Log-Lojistik HBSM'ye göre sağ kalım eğrileri pythonda predict_survival_function kullanılarak tahminlenmiştir. Şekil 5.11'de müşteri numarası 1-5 olan müşteriler için çizilen tahminlenmiş sağ kalım eğrisine göre 4 nolu müşterinin 10 ay içerisinde kayıp yaşayacağı, 2 nolu müşterinin 22 ay sonra kayıp yaşayacağı, 5 nolu müşterinin 32 ay sonra kayıp yaşayacağı, 4 ve 5 nolu müşterilerin 70 ay içerisinde herhangi bir kayıp riski görülmemektedir. Şekil 5.12'de müşteri numarası 1000-1005 olan müşteriler için çizilen tahminlenmiş sağ kalım eğrisine göre 1004 nolu müşterinin 8 ay içerisinde kayıp yaşayacağı, diğer müşterilerde 70 ay içerisinde kayıp riski görülmediği söylenebilir. Bu bilgiler ışığında 1 ve 1004 nolu müşterilerin kısa zamanda operatörü terk edecek olmaları mobil operatör için bu müşterilere reaksiyon verme fırsatı adına değerli bir bilgidir.

Çalışmamızda kullanılan müşteri kayıp veri setinde, çalışma bitmesine rağmen kayıp yaşamayan ve sağdan durdurulmuş olan 5163 müşteriyi, geleneksel sınıflandırma modelleri, kayıp yaşamamış gibi modele dahil etmesi, modelde bilgi kaybı yaşanmasına ve eksik veri analizi yapılmasına neden olmaktadır. Çalışmamızda uyguladığımız SKA, geleneksel modellerin aksine modelleriyle

sağdan durdurulmuş müşteriler herhangi bir bilgi kaybı yaşanmadan modele dahil edilmişlerdir.

Çalışmamızda kullanılan Python lifeline paketi yarı parametrik modellerde parametre tahminlemesi için metot seçimlerinde esneklik sağlamamaktadır. Varsayılan Breslow olup, Efron, eş zamanlı olayların çokluğunda paket tarafından seçilip kullanılmaktadır. [21] çalışmasında R programı kullanıldığı için SKA paketinde bu esneklik sağlanıp üç metotta seçilebilmektedir. Çalışmamızda yarı parametrik modeler için Python yerine R kullanılıp Kesin, Breslow ve Efron metodlarının AIC performansları incelenebilir.

Çalışmamızda kullanılan yeni yöntemler, simülasyonla oluşturulan veri setlerine başvurulup performansları ölçülebilir ve sonuçlar çalışmamızda çıkan sonuçlarla kıyaslanabilir. Aynı şekilde bankacılık ve sigortacılık gibi farklı sektörlerden veri setlerine başvurulup etkinliği ve performansları ölçülebilir. SKA'da yeni yaklaşımlardan olan ve sınıflandırma modelleri ile uygulanan Destek Vektör Makineleri SKA ve Rastgele Orman Algoritması SKA ve Derin Öğrenme SKA metodlarına başvurularak, elde edilen sonuçlar çalışmamızda çıkan sonuçlarla karşılaştırılabilir[55,56,57]. Telekom sektöründe bu modeller sayesinde model performansları daha iyi seviyelere çıkartılabilir.

# KAYNAKÇA


[1] Berson, A., Smith, S. and Thearling, K. (2000) “Building data mining applications for CRM”. New York: McGraw-Hill.

[2] Owczarczuk, M. (2010) “Churn models for prepaid customers in the cellular telecommunication industry using large data marts,” Expert Systems with Applications, 37(6), pp. 4710–4712. Available at: https://doi.org/10.1016/j.eswa.2009.11.083.

[3] Pendharkar, P.C. (2009) “Genetic algorithm based neural network approaches for predicting churn in cellular wireless network services,” Expert Systems with Applications, 36(3), pp. 6714–6720.

[4] Idris, A., Rizwan, M. and Khan, A. (2012) “Churn prediction in telecom using Random Forest and PSO based data balancing in combination with various feature selection strategies,” Computers & Electrical Engineering, 38(6), pp. 1808–1819.

[5] Huang, Y. and Kechadi, T. (2013) “An effective hybrid learning system for telecommunication churn prediction,” Expert Systems with Applications, 40(14), pp. 5635–5647.

[6] XIA, G.-en and JIN, W.-dong (2008) “Model of customer churn prediction on supporvector machine,” Systems Engineering - Theory & Practice, 28(1), pp. 71–77. Available at: https://doi.org/10.1016/s1874-8651(09)60003-x.

[7] Kim, K., Jun, C.-H. and Lee, J. (2014) “Improved churn prediction in telecommunication industry by analyzing a large network,” Expert Systems with Applications, 41(15), pp. 6575–6584. Available at: https://doi.org/10.1016/j.eswa.2014.05.014.

[8] Lim, T. (2020) “Applying survival analysis for customer retention: A U.S. Regional Mobile Service operator,” 2020 Zooming Innovation in Consumer Technologies Conference(ZINC)[Preprint].Availableat:https://doi.org/10.1109/zinc50678.2020 .9161811.

[9] Lu, J. “Modeling Customer Lifetime Value Using Survival Analysis: An Applicationin the Telecommunications Industry SAS” User Groups International Proceedings Seattle Washington, 120 (28), 2003.

[10] King, Barry & Rice, Jennifer. (2019). “Analysis of Churn in Mobile Telecommunications: Predicting the Timing of Customer Churn”. AIMS International Journal of Management. 13. 127. 10.26573/2019.13.2.3.

[11] Suhartonoi,D., Saefuddin,A., Sumertajaya,M.,“Survival Analysis Of Customer In Postpaid Telecommunication Industry” Indonesian Journal of Statistics available online at: ISSN : 0853-8115 journal.ipb.ac.id/index.php/statistika Vol. 18 No.1, p: 1-10 (2013).

[12]Havrylovych ,M., Kuznietsova ,N.“Survival analysis methods for churn prevention in telecommunications industry” Institute for Applied System Analysis of the National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine.

[13]Kaggle,”Telco Customer Churn”,Available at: https://www.kaggle.com/ /datasets/blastchar/telco-customer-churn?resource=download

[14]R. Mattersion Telecom churn management, APDG Publishing, Fuquay-Varina, NC (2001).

[15]Fornell , C .( 1992 ) ‘ National Customer Satisfaction Barometer: The Swedish experience ’ , Journal of Marketing, Vol. 56 , January, pp. 6 – 21 .

[16]Safko, L., (2011). “The Fusion Marketing Bible: Fuse Traditional Media, Social Media, & Digital Media to Maximize Marketing ” McGraw-Hill Education - Europe 35, 344-356

[17]Colgate, M., Stewart, K., & Kinsella, R. (1996). Customer defection: A study of the student market in ireland. International Journal of Bank Marketing, 14(3), 23–29.

[18]Ganesh, J., Arnold, M., & Reynolds, K. (2000). Understanding the customer base of service providers: An examination of the differences between switchers and stayers. Journal of Marketing, 64(3), 65–87.

[19]Zeithaml, V., Berry, L., & Parasuraman, A. (1996). The behavioural consequences of service quality. Journal of Marketing, 60(2), 31–46.

[20]Li, M.-L., Green, R.D.: A mediating influence on customer loyalty: the role of perceived value. J. Manag. Market. Res. 7, 1–12 (2011)

[21]Masarifoglu, M. and Hakan Buyuklu, A. (2019) “Applying survival analysis to Telecom Churn Data” American Journal of Theoretical and Applied Statistics, 8(6), p. 261. Available at: https://doi.org/10.11648/j.ajtas.20190806.18.

[22]Kaplan, L. E., (1958). ” Non-parametric estimation from incomplete observations”, Journal of American Statistical Association 53, 457-81.

[23]Cox, D. R., (1972). “Regression Model and Life-Tables” Journal of the Royal Statistical Society 34,187-220.

[24]Ranganathan, P., Pramesh, C.S, (2012). ” Censoring in survival analysis: Potential for bias”, Perspectives in Clinical Research, 3 (1), 40.

[25]Shihcorresponding, W. J., (2011). ”Problems in dealing with missing data and informative censoring in clinical trials”, Curr Control Trials Cardiovasc Med, 35 (1), 21-31.

[26]Collett D.,(2003) “Modelling Survival Data in Medical Research”, Chapman&Hall London, 1-271.

[27]Shenyang, Guo, '4 The Cox Proportional Hazards Model', Survival Analysis, PKT GD TOSOCIAL WORK RESEARCH METHODS (New York, 2009; online edn, Oxford Academic, 1 Jan. 2010).

[28]Gonzalez, D., Pina, M. and Torres, P., (2008). ” Estimation of Parameters in Cox's Proportional Hazard Model: Comparisons between Evolutionary Algorithms and the Newton-Raphson Approach Mexican International Conference on Artificial Intelligence”, Springer Link, 513-523

[29]Breslow, N., (1974). “Covariance Analysis of Censored Data.”, Biometrics, 30, 89:99.

[30]SAS Institute Inc., (2010). ”SAS/GRAPH 9.2 Reference”, Second Edition. Cary,NC: SAS Institute Inc., SAS Institute Inc.

[31]Davidson-Pilon, (2019). “lifelines: survival analysis in Python.” Journal of Open Source Software, 4(40), 1317, https://doi.org/10.21105/joss.01317.

[32]R Development Core Team, (2011). “R: A Language and Environment forStatistical Computing. R Foundation for Statistical Computing”, Vienna, Austria. ISBN 3-900051-07-0, URLchttp://www.R-project.org/ ,URL http://www.R-project.org/.

[33]Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. Journal of the American Statistical Association 72, 557-565.

[34]Hertz-Picciotto, I. & Rockhill, B., (1997). “Validity and Efficiency of Approximation Methods for Tied Survival Times in Cox Regression”, Biometrics, 53, 1151-1156.

[35]Schoenfeld, D., (1982). “Partial Residuals fort he Proportional Hazards Regression Model”, Biometrika, Vol 69, No 1, 239-241.

[36]Therneau, Terry M., and Patricia M. Grambsch. 2000. “Modeling Survival Data: Extending the Cox Model”. New York: Springer.

[37]Abrahamowicz, M., MacKenzie, T., Esdaile, J. M., “Time-dependent hazard ratio: modelling and hypothesis testing with application in lupus nephritis, Journal of the American Statistical Association”, 91(1996), 1432-1439.

[38]Schemper, M., Wakounig, S., Heinze, G., “The estimation of average hazard ratios by weighted Cox regression”, Statistics in Medicine, 28: 2473-2489, (2009).

[39]K.R. Hess, Assessing time-by-covariate interactions in proportional hazards regression models using cubic spline functions, Stat. Med. 13 (1994) 104551062.

[40]Wei, L.J., (1992). “The Accelerated Failure Time Model: A Useful Alternative to the Cox Regression Model in Survival Analysis”, Statistics in Medicine,Vol. 11, 1871-1879.

[41]Ata, N., & Sözer, M. T. (2007). “Cox regression models with nonproportional hazards applied to lung cancer survival data”. Hacettepe Journal of Mathematics and Statistics, 36(2), 157-167.

[42]Royston, P. and Parmar, M.K. (2002) “Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects,”

Statistics in Medicine, 21(15), pp. 2175–2197. Available at: https://doi.org/10.1002/sim.1203.

[43]Wegman EJ, Wright IW. Splines in statistics. Journal of the American Statistical Association 1983; 78(382): 351-65.

[44]Harrell FE Jr. Regression Modeling Strategies. 2nd ed. New York, NY: Springer-Verlag; 2015.

[45]Austin PC, Fang J, Lee DS. Using fractional polynomials and restricted cubic splines to model non-proportional hazards or time-varying covariate effects in the Cox regression model. Stat Med. 2022 Feb 10;41(3):612-624.

[46]Klein, P., J,., Goel, P., (2011). “Survival Analysis: State of the Art”, Kluwer Academic Publishers, Series E:Applied Sciences, 211.

[47]Cox, D.R., Oakes, D., 1984. “Analysis of Survival Data”. Chapman and Hall, London, 193.

[48]Royston, P., Lambert, P., (2011). ”Flexible Parametric Survival Analysis Using Stat: Beyond the Cox Model”, Stata Press, 13.

[49]Miller, R. ‘Least squares regression with censored data’, Biometrika, 63, 449-464 (1976).

[50]Prentice, R. L. ‘Linear rank tests with right censored data’, Biometrika, 65, 167-179 (1978).

[51]Buckley, I. V. and James, I. ‘Linear regression with censored data’, Biometrika, 66, 429-436 (1979)

[52]Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Pet rov & F. Caski (Eds.), Proceedings of the Second International Symposium on Information Theory (pp. 267-281). Budapest: Akademiai Kiado

[53]Vargas, R. (2018) “Survival analysis of mobile prepaid customers using the Weibull distribution.” Available at: https://doi.org/10.31235/osf.io/baq76.

[54]Harrell, F.E. (1982) “Evaluating the yield of medical tests,” JAMA: The Journal of the American Medical Association, 247(18), p. 2543. Available at: https://doi.org/10.1001/jama.1982.03320430047030.

[55]Eleuteri, A. (2008) “Support vector survival regression,” 4th IET International Conference on Advances in Medical, Signal and Information Processing (MEDSIP 2008) [Preprint]. Available at: https://doi.org/10.1049/cp:20080436.

[56]Husband, S.M. and Roberts, J. (2017) “Survival random forest to predict time to fill,” 2017 IEEE International Conference on Data Mining Workshops (ICDMW) [Preprint]. Available at: https://doi.org/10.1109/icdmw.2017.32.

[57]Lee, C. et al. (2018) “DeepHit: A deep learning approach to survival analysis with competing risks,” Proceedings of the AAAI Conference on Artificial Intelligence, 32(1). Available at: https://doi.org/10.1609/aaai.v32i1.11842.

# TEZDEN ÜRETİLMİŞ YAYINLAR

## Makaleler

1. Masarifoglu, M. and Hakan Buyuklu, A. (2019) “Applying survival analysis to Telecom Churn Data” American Journal of Theoretical and Applied Statistics, 8(6), p. 261. Available at: https://doi.org/10.11648/j.ajtas.20190806.18.