T.C.
SİNOP ÜNİVERSİTESİ
FEN BİLİMLER ENSTİTÜSÜ
MATEMATİK ANABİLİM DALI

ÖZDEN İŞÇİMEN

YÜKSEK LİSANS TEZİ

COX REGRESYON MODELİNDE KAYIP VERİ PROBLEMİ İÇİN ÇÖZÜM ÖNERİLERİ

DANIŞMAN
DOÇ. DR. NESRİN ALKAN

T.C.
SİNOP ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

MATEMATİK ANABİLİM DALI

YÜKSEK LİSANS TEZİ

COX REGRESYON MODELİNDE KAYIP VERİ PROBLEMİ İÇİN ÇÖZÜM ÖNERİLERİ

YAZAR
ÖZDEN İŞÇİMEN

DANIŞMAN
DOÇ. DR. NESRİN ALKAN

SİNOP – 2019

TEZ KABUL

Özden İŞÇİMEN tarafından hazırlanan "Cox Regresyon Modelinde Kayıp Veri Problemi İçin Çözüm Önerileri " başlıklı bu çalışma, 24.06.2019 tarihinde yapılan savunma sınavı sonucunda başarılı bulunarak, jürimiz tarafından **YÜKSEK LİSANS tezi** olarak kabul edilmiştir.

Başkan	Prof. Dr. Kamil DEMİRCİ Sinop Üniversitesi/Fen Edebiyat Fakültesi	İmza
Üye	Prof. Dr. Yüksel TERZİ Ondokuz Mayıs Üniversitesi/Fen Edebiyat Fakültesi	İmza
Üye	Doç. Dr. Nesrin ALKAN Sinop Üniversitesi/ Fen Edebiyat Fakültesi	İmza

ETİK BEYANI

Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmasında; tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi, tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu, tez çalışmasında yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi, kullanılan verilerde herhangi bir değişiklik yapmadığımı, bu tezde sunduğum çalışmanın özgün olduğunu bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

Yazarın adı soyadı: Özden İŞÇİMEN

İÇİNDEKİLER

Sayfa

SEMBOLER VE KISALTMALAR LİSTESİ

SEMBOLLER

$\hat{\mu}$: Örneklem ortalama vektörü

Σ^*: Simüle edilmiş kovaryans matrisi

$\hat{\lambda}$: Kayıp oranı

$S(t)$: Sağkalım fonksiyonu

$f(t)$: Olasılık yoğunluk fonksiyonu

$H(t)$: Birikimli hazard fonksiyonu

$h(t)$: Hazard fonksiyonu

$S_0(t)$: Temel sağkalım fonksiyonu

$\widehat{\beta}$: En çok olabilirlik tahmini

W: Wald testi istatistiği

S: Skor testi istatistiği

T: Sağkalım süresi

L:Sansürleme zamanı

$F(t)$: Kümülatif dağılım fonksiyonu

$L(\hat{\theta})$: Olabilirlik fonksiyonu

$I(\beta)$: Logaritmik kısmi olabilirlik

$\hat{\theta}$: En çok olabilirlik tahmin edicisi

KISALTMALAR

CCA: Eksiksiz veri analizi (Complete Case Analysis)

EM: Beklenti maksimizasyon (Expectation Maksimizasyon)

LR: Olabilirlik oran

MAR: Rasgele kayıp (Missing at random)

MCAR: Tamamen rasgele kayıp (Missing completely at random)

MCMC: Markov Zinciri Monte Carlo (Markov Chain Monte Carlo)

MEAN: Ortalama değer atama (Mean Imputation)

MI: Çoklu değer atama (Multiple Imputation)

MNAR: Rasgele olmayan kayıp (Missing not at random)

REG: Regresyon değer atama (Regression Imputation)

ŞEKİLLER LİSTESİ

ÇİZELGE LİSTESİ

ÖZET

COX REGRESYON MODELİNDE KAYIP VERİ PROBLEMİ İÇİN ÇÖZÜM ÖNERİLERİ

Tıp, biyoloji, mühendislik ve ekonomi gibi birçok alanda kullanımı olan sağkalım analizi, istenilen olayın gerçekleşmesine kadar geçen zamanın modellenmesini sağlar. Sağkalım analizinin en önemli özelliği veri setinde sansürlü verilerin olmasıdır. Sansürlü veri, araştırma sonlandırıldığında istenilen olayın gerçekleşmediği durumda ortaya çıkmaktadır. Sağkalım analizinin en çok kullanılan yöntemi, sağkalım süresini etkileyen faktörleri belirleyen Cox regresyon analizidir. Cox regresyon analizinde de diğer istatistiksel yöntemlerde olduğu gibi tüm gözlem değerlerinin biliniyor olması gerekir. Veri setinde ilgilenilen değişkenin bazı değerlerinin gözlenememesi kayıp veri olarak adlandırılır ve araştırmalar için önemli bir problemdir. Bu çalışmada amaç, kayıp veri problemini gideren yöntemlerin tanıtılması ve farklı kayıp veri mekanizmalarında performanslarının değerlendirilerek farklı örnek genişlikli ve kayıp oranlı veri setlerinde karşılaştırmaktır. Bu amaçla öncelikle simülasyon çalışması yaparak her bir kayıp veri mekanizması için N=50, 100, 300 birimlik örnek genişlikli veri setlerinde %5, %10, %20 ve %30 kayıp oranlı veri setleri türetilmiştir. Farklı senaryolar için elde edilen bu veri setlerinin herbirinde kayıp veri analiz yöntemleri uygulanarak, Cox regresyon analizi yapılmıştır. Ayrıca genellikle klinik çalışmalarda yaygın olarak kullanılan Cox regresyon analizi, ekonomi alanında uygulanmış ve gerçek veri seti olarak işsizlik verisi kullanılarak kayıp veri analiz yöntemleri, Cox regresyon analizi üzerinden karşılaştırılmıştır.

Anahtar Kelimeler: Sağkalım Analizi, Cox Regresyon Analizi, Kayıp Veri Analizi

ABSTRACT

THE SOLUTION PROPOSALS FOR MISSING DATA PROBLEMS IN COX REGRESSION MODEL

Survival analysis, which is used in many areassuch as medicine, biology, engineering and economics, allows modeling of the time until the desired event takes place. The most important feature of survival analysis is the presence of censoreddata in the data set. Censored data arises where the desired event does not ocur when the research is terminated. The most commonly used method of survival analysis is Cox regression analysis, which determines factors affecting survival time. In Cox regression analysis, all observation values should be known as in other statistical methods. The fact that some values of the variable of interest in the data set cannot be observed is called as missing data and it is an important problem for researchers. The aim of this study is to introduce the methods that solve the missing data problem and to evaluate the performance of different missing data mechanisms and to compare them in data sets with different sample size and missing rate. For this purpose, data sets with 5%, 10%, 20% and 30% missing rates and N = 50, 100, 300 units of sample size are derived for each missing data mechanism. Thus, Cox regression analysis was performed by using missing data analysis methods in each of these data sets for different scenarios. Cox regression analysis, which is commonly used in clinical studies, has also been applied in the field of economics. Missing data analysis methods were compared over Cox regression analysis using unemployment data as real data set.

Keywords: Survival Analysis, Cox Regression Analysis, Missing Data Analysis

TEŞEKKÜR

Tez çalışmamın her aşamasında değerli anlayış ve sabrıyla bana yol gösteren her defasında bana her anlamda desteğini esirgemeyen danışmanım Sayın Doç. Dr. Nesrin ALKAN'a, sonsuz saygı ve teşekkürlerimi sunarım. Yüksek lisans sürecim boyunca yardım ve destekleri için Sayın Doç. Dr. Barış ALKAN'a teşekkür ederim. Ayrıca tez çalışmam sırasında göstermiş oldukları maddi manevi anlayış, sabır ve desteklerinden dolayı annem, babam, ablam ve sevgili eş adayıma teşekkür ederim.

1. GİRİŞ

Sağkalım analizi, tıp, biyoloji, mühendislik, halk sağlığı ve ekonomik çalışma alanlarının pek çoğunda kullanım alanına sahiptir. Araştırmaların çoğunda farklı grupların araştırılması söz konusudur. Gruplardaki bireyler birbirleri ile benzer birçok özelliğe sahiptir. Bu özellikler, yaş, cinsiyet gibi demografik değişkenler kandaki glikoz düzeyleri, kan basıncı gibi, fizyolojik değişkenler diyet, sigara içme durumu gibi davranışsal değişkenler şeklinde sıralanabilir.

Sağkalım analizi ilk olarak 17. yüzyılda kullanılmaya başlamıştır. Edmund Halley, 1687-1691 yılları arasında ilk sağkalım tablosunu tasarlamıştır. Günümüzde demografi ve aktüerya çalışmalarında Halley'in tasarladığı sağkalım tablosu kullanılmakta olan sağkalım tabloları ile çok benzerlik göstermektedir. İkinci Dünya Savaşı sırasında 20. Yüzyılda sağkalım süreleri üzerine araştırmalar hızlanmış ve özellikle askeri teçhizatların güvenilirliği ile ilgili çalışmalar yapılmıştır. Savaş sonrasında da özellikle elektronik endüstrisi alanında sağkalım analizlerinin önemi artmıştır. Sağkalım sürelerini etkilediği düşünülen bağımsız değişken ya da ortak değişkenler belirlenmeye çalışılmıştır. Bu tür verilerin modellenmesi için sansürlü verilerinde kullanıldığı Cox regresyon analizi kullanılmıştır.

1972 yılında sağkalım analizinde önemli adımlar atılmış, Cox tarafından geliştirilen regresyon modeli kullanılmaya başlanmıştır. Daha sonra Kalbfleisch ve Prentice (1980) tarafından yapılan katkılarla bugünkü önemini kazanmıştır. Son yıllarda Cox regresyon modeli bağımsız açıklayıcı değişkenlerin yanı sıra zamana bağımlı açıklayıcı değişkenleri de içeren Cox regresyon modeli ile ilgili çalışmalar yapılmıştır.

Kayıp veri sorunu istatistiksel analizlerde, önemli bir tartışma alanıdır. Tartışmaların bazıları, standart istatistiksel yöntemlerin kendisinden kaynaklanmaktadır. Veri düzeninin bir matris şeklinde olduğu düşünülürse satırlar gözlemleri, sütunlar ise değişkenleri temsil etmektedir. Bir gözlemde bir değişkene yönelik değerin bulunmaması durumunda, o hücre boş kalır ve kayıp veri oluşur. Sonuçta kayıp değer içeren veri setinde yer alan bazı değişkenlere ilişkin bilgilerin ya da gözlemlenen değerlerin kayıp olduğu anlamına gelmektedir (Little ve Rubin, 1987).

Kayıp verinin oluşma aşaması bireylerin sorulan soruya cevap vermekten kaçınması (Finch ve Margref, 2008), verilen sürenin yetersiz olması (Custer, Sharairi ve Swift,

2012), dikkat dağınıklığı ve uykusuzluk, gibi kişisel durumlar (Garson, 2015); araştırmanın konusundan kaynaklı yalnızca bazı maddelerin cevaplanması (Graham, Hofer ve MacKinnon,1996; Sinharay, Stern ve Russell, 2001) veya bireyin doğru cevabı bilmemesi (Finch ve Margref, 2008) gibi durumlarda araştırmacılar kayıp verilerle karşılaşabilirler. Veri setinde kayıp değerin varlığı, istatistiksel analiz sonuçlarını etkileyerek ve hatalı sonuçlara neden olabilir. Kayıp verilerin neden olacağı problemler kayıp değerin miktarına, kayıp değerin dağılımının oluşturduğu yapıya ve bu iki durumun etkileşimine bağlı olarak değişebilir.

Bilim insanları uzun yıllardır çalışmasına rağmen kayıp veri probleminin giderilmesi konusunda en büyük gelişmeyi Demster, Laird ve Rubin tarafından 1977 yılında EM algoritmasını geliştirmeleriyle ile elde etmişlerdir. Ayrıca 1987 yılında Rubin tarafından bulunan Çoklu Değer Atama yöntemi ile kayıp veri problemini gideren yöntemler elde edilmiştir. Kayıp verilerin veri setindeki dağılımlarına ilişkin Little ve Rubin (1987) tarafından bir sınıflama yapılmıştır. Üç veri seti tanımlanmıştır. Oluşan yapılar, kayıp veri olma olasılığı ile ölçülen değişkenler arasındaki ilişkiyi tanımlamaktadır. Yapılar; eğer bir değişkende kayıp veri olma olasılığı, değişkenin kendisine bağlı olmayıp veri setindeki diğer değişkenlere bağlı ise bu tür verilerdeki kayıp veri oluşumu rasgele kayıp (missing at random- MAR) olarak adlandırılır. Eğer bir değişkendeki kayıp veri olma olasılığı ne değişkenin kendisine nede veri setinde yer alan diğer değişkenlerle ilişkili olmadığı durumlarda kayıp veriler tamamen rasgele kayıp (missing at completely random, MCAR) olarak adlandırılır. MAR kayıp veri yapıları rastlantısal olarak adlandırılmasına rağmen bu yapılarda sistematik bir kayıp veri mekanizması vardır (Baraldi ve Enders, 2010). Son olarak, bir değişkende kayıp veri olma olasılığı değişkenin kendisine bağlı ise bu kayıp veriler rasgele olmayan kayıp (missing not at random- MNAR) olarak adlandırılır.

Bu çalışmada amaç, kayıp veri problemini gideren yöntemlerin tanıtılması ve farklı kayıp veri mekanizmalarında performanslarının değerlendirilerek farklı örnek genişlikli ve kayıp oranlı veri setlerinde karşılaştırmaktır. Bu amaçla öncelikle simülasyon çalışması yaparak her bir kayıp veri mekanizması için N=50, 100, 300 birimlik örnek genişlikli veri setlerinde %5, %10, %20 ve %30 kayıp oranlı veri setleri türetilmiştir. Farklı senaryolar için elde edilen bu veri setlerinin her birinde kayıp veri analiz yöntemleri uygulanarak Cox regresyon analizi yapılmıştır. Ayrıca genellikle klinik çalışmalarda yaygın olarak kullanılan Cox regresyon analizi, ekonomi alanında

uygulanmış ve gerçek veri seti olarak işsizlik verisi kullanılarak kayıp veri analiz yöntemleri, Cox regresyon analizi üzerinden karşılaştırılmıştır.

Bu tez çalışması beş bölüm ve kaynaklardan oluşmaktadır. Çalışmanın ikinci bölümünde sağkalım analizinin temel tanımları, fonksiyonlar ve ilişkileri verilmiştir. Ayrıca bu bölümde Cox regresyon modeli, modeldeki parametre tahmin yöntemleri ve önemlilik testleri verilmiştir.

Üçüncü bölümde yani materyal ve yöntemler kısmında kayıp verinin tanımı, kayıplığın veri ile ilişkisini veren kayıp veri mekanizması ve kayıp veri problemini gideren yöntemler açıklanmıştır.

Çalışmanın dördüncü bölümü olan bulgular ve tartışma kısmında kayıp veri mekanizması olan MAR, MCAR ve MNAR'nın her biri için farklı örnek genişlikli ve farklı kayıp oranlı veri setleri türetilerek kayıp veri problemini gideren yöntemler karşılaştırılmıştır. Böylece yöntemlerin hangi kayıp veri mekanizmasında, hangi örnek genişliğinde ve kayıp oranında daha iyi performansa sahip olduğu belirlenmiştir. Yöntemlerin simülasyon çalışması üzerinden karşılaştırılmasının yanı sıra gerçek veri seti uygulaması da yapılmıştır. Ayrıca genellikle klinik çalışmalarda yaygın olarak kullanılan Cox regresyon analizi, gerçek veri seti olarak işsizlik verisi kullanılarak Cox regresyonun ekonomi alanında ekonomi alanındaki kullanımı gösterilmiştir.

Çalışmanın son bölümünde ise elde edilen tüm sonuçlar tartışılarak gelecekteki çalışmalar için önerilerde bulunulmuştur.

2.GENEL BİLGİLER

2.1. Sağkalım Analizi

Sağkalım analizi, istenilen olayların gerçekleşmesine kadar geçen sağkalım zamanının modellenmesini sağlayan bir analizdir. Tıp alanında bir hastalığın gelişme süreci, tedaviye cevabı ve ölüm sıklıkla incelenmiş olaylardır. Sağkalım analizi ekonomi, sosyal yaşam, zooloji, botanik gibi birçok alanda da kullanılabilmektedir. Örneğin tıp alanında kanserli hastaların yaşam süresini etkileyen faktörleri belirlemede ya da farklı tedavi yöntemleri uygulanmış hasta grupları arasında karşılaştırma yapmada, ekonomi alanında bir şirketin daha ne kadar süre aynı şartlar altında varlığını sürdürebileceği konularında sağkalım analizi kullanılabilir. Ayrıca araştırmalarda ilgilenilen olay çalışma alanına göre farklılık gösterilebilir. Tıp alanında ölüm, bebeğin anne sütünü bırakması, gebeliğin sonlandırılması, mühendislik alanında makinanın bozulması, bir ağrının başlaması, trafikte bir kazanın olması gibi de olabilir (İnceoğlu, 2013).

Tıp alanında yapılan bir sağkalım analizinde, hastaların sağkalım süreleri, demografik özellikleri, tedaviye cevabı veya ölümü, tedavi bilgisi, muayene verileri gibi daha pek çok özellik sağkalım verisini oluşturur. Genellikle sağkalım analizinde, sağkalma olasılığı, dağılımları ve belirlenen risk faktörleri hesaplanır ve gruplar arasında karşılaştırma yapılır. Ayrıca sağkalım analizinde sağkalım sürelerini etkileyen açıklayıcı faktörler belirlenebilmektedir.

Sağkalım verilerinin analizinde klasik istatistiksel analiz yöntemleri yetersiz kalmaktadır. Bunun en önemli nedeni, çoğu kez araştırmanın değerlendirilmesinin tüm hastalar ölmeden veya ilgilenilen olay her bir birey için ortaya çıkmadan yapılması gerektiğidir. Aksi durumda hangi tedavi yönteminin daha iyi olduğunun ve hastalığı etkileyen faktörlerin saptanmasının uzun süre alabileceğidir. Diğer bir neden ise, hastalara uygulanan tedavilerin aynı zamanda başlamamasıdır. Çünkü hastalar sisteme farklı zamanda girmiş olabilir. Bu tip çalışmalarda sağkalım analizi yöntemleri daha uygun sonuçlar vermektedir (Cox, 1972).

2.1.1. Sansürleme

Sağkalım analizinin en önemli özelliği veri setinde sansürlü verilerin olmasıdır. Sansürlü veriler ise araştırma sonlandırıldığında istenilen olayın gerçekleşmediği durumda ortaya çıkar. Örneğin tıp alanında yapılan çalışmalarda bazı hastalar çalışma

sona erdiğinde hala yaşıyor olabilir. Trafik kazası ya da başka bir sebepten ölmüş veya başka bir yere taşındığı için çalışmadan ayrılmış olabilir. Bu tür hastalar için sağkalım süresi kesin olarak bilinemeyeceğinden bu tür gözlemlere sansürlü denir.

i) 1. Tip Sansürleme

Bu tip sansürlemede sağkalım zamanı en az çalışma süresinin uzunluğu kadardır. Örneğin farelerin beyin ölümü ile ilgili bir çalışmada sınırlı maliyet ve kısıtlı zamandan dolayı bütün farelerin beyin ölümünün gerçekleşmesi beklenemez. Beyin ölümü gerçekleşmeyen fareler sansürlü veriler olarak ele alınır.

Herbiri birbirinden bağımsız F dağılım fonksiyonuna sahip $T_1, T_2, \ldots T_N$ rasgele değişkenleri ele alınsın. T_i i.nci sağkalım zamanını ve C_i önceden belirlenen sabit sağkalım sansür zamanı ise T_i nin C_i ye göre büyük yada küçük olma durumuna göre Y_i rasgele değişkeni tamamlanmış olsun.

$$Y_i = \begin{cases} T_i , & T_i \leq C_i \\ C_i , & T_i > C_i \end{cases} \qquad (2.1)$$

Burada $P\{T > C\} > 0$ ve $Y_i = C_i$'dir ve Y'nin dağılım fonksiyonunun pozitiftir. Ayrıca tüm C_i ler birbirine eşit ve sabittir (Terzi, 2003).

ii) 2. Tip Sansürleme

Bu tip sansürlemede araştırmacı belirlediği ölüm sayısına ulaştığında çalışmasını sonlandırır ve yaşayan gözlemleri sansürlü olarak adlandırır. Çalışmada sansürlü bireylerin sağkalım süresi, en son ölen bireyin sağkalım süresi olarak değerlendirilir.

$T_1, T_2, \ldots T_n$ rasgele değişkenlerinin sıra istatistikleri $T_{(1)} < T_{(2)} < \cdots < T_{(n)}$olsun. Bu sansürlemede $r < n$ olmak üzere r-inci başarısızlıktan sonra çalışma sona erdirilir Böylelikle çalışmada $T_{(1)} < T_{(2)} < \cdots < T_{(r)}$ gözlenmiş olur. Fakat çalışma tüm gözlemler başarısız olana kadar devam etmez ve çalışma r-inci bireyin başarısız olmasıyla durdurulur (Leung ve ark.,1997).

iii) 3. Tip Sansürleme

$C_1, C_2, \ldots, C_n$ sansürleme zamanları ve $Y_1, Y_2, \ldots, Y_n$ değişkenleri birbirinden bağımsız olsun.

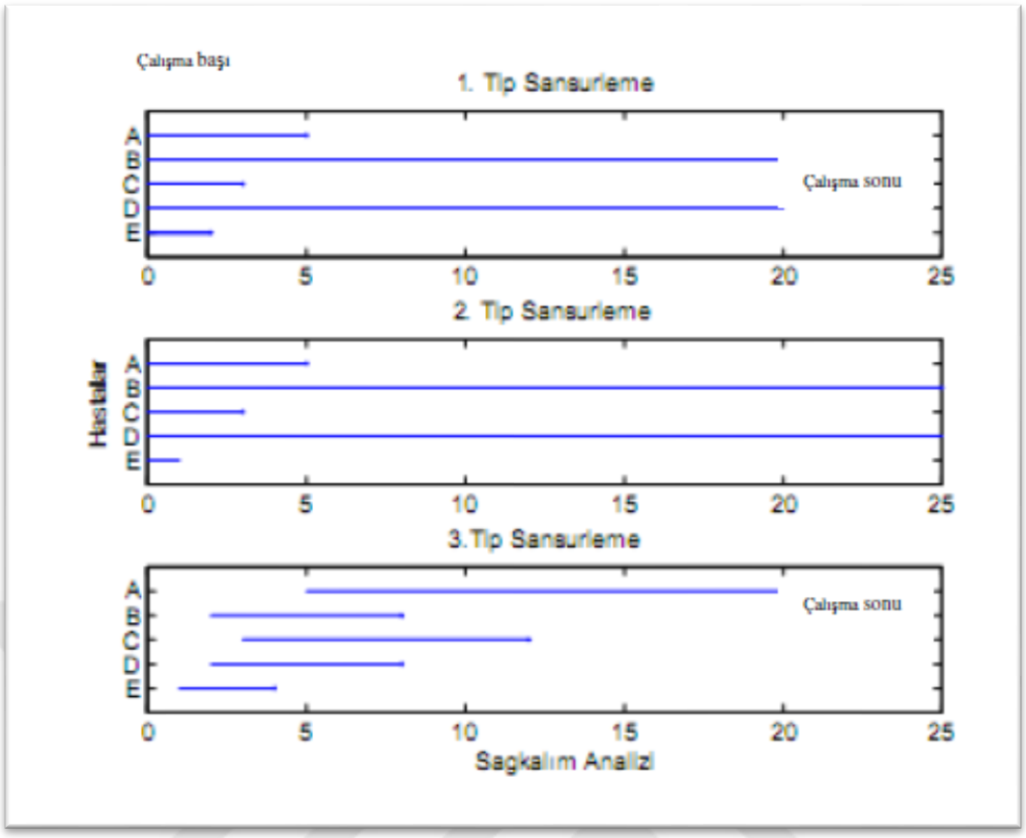
$$Y_i = \min(T_i, C_i)$$

$$\delta_i = I(T_i \leq C_i) = \begin{cases} 1 & , \ T_i \leq C_i \\ 0 & , \ T_i > C_i \end{cases} \qquad (2.2)$$

Burada $\delta_1, \ldots, \delta_n$ değerleri sansürleme hakkında bilgi vermektedir. Eğer $\delta = 0$ $(T_i > C_i)$ ise veriler sansürlü, $\delta = 1$ $(T_i \leq C_i)$ olduğunda ise sansürleme yoktur. Yani olumsuzluk (ölüm) söz konusudur (Terzi, 2003).

Pek çok araştırmada çalışma zamanı sınırlı olarak belirlenmektedir. Gözlenen bireyler araştırma süresine farklı zamanlarda dâhil olabilmektedir. Gözlenen bireyler çalışma sona erdiğinde hala yaşıyorsa sansürlüdür. Sağkalım zamanı ise çalışmaya girdikleri zamanla çalışmanın sona erdiği zaman arasındaki farktır.

I. ve II. tip sansürlü gözlemler tek basına sansürlü gözlemler olarak adlandırılırken üçüncü tip sansürlü veriler kademeli veya sıralı sansürlü veri adını alırlar.



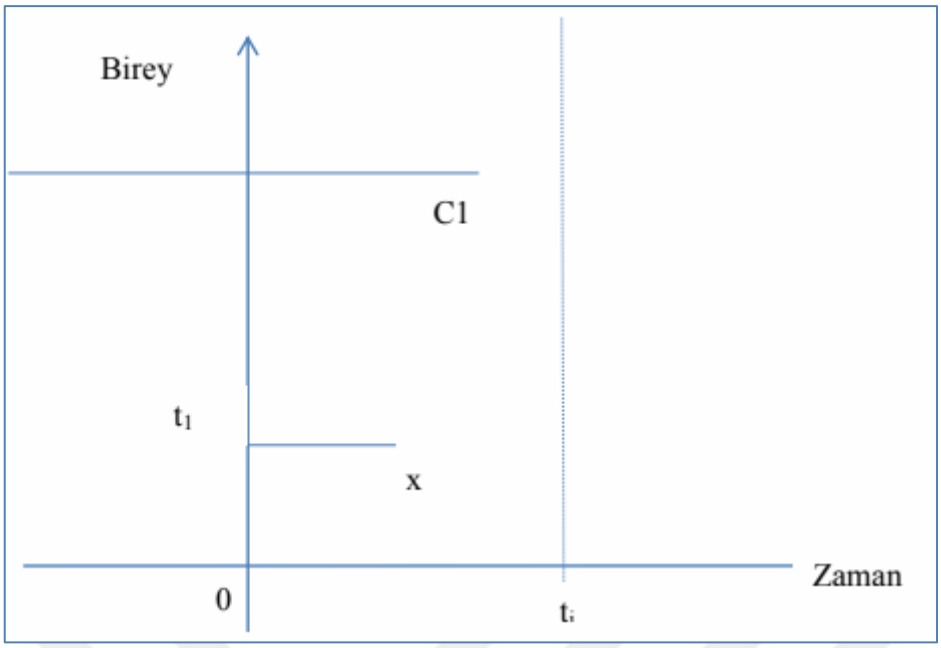
Şekil 2.1. Sansür Çeşitleri

Soldan Sansürleme

Soldan sansürlü verilerde olay belli bir t zamanından önce tamamlanmıştır. C_i sansürleme zamanı T_i bireyin sağkalım süresi olmak üzere; T_i<C_i olması durumunda bu bireyin yaş süresinin soldan sansürlenmiş olduğu söylenir (Lawless, 2011).

$$\partial_i = \begin{cases} 0, & T_i \leq C_i \\ 1, & T_i > C_i \end{cases} \tag{2.3}$$

Eğer $\partial_i = 0\ (T_i < C_i)$ ise birey sansürlenmiş, $\partial_i = 1\ (T_i > C_i)$ ise ilgilenilen olay gerçekleşmiştir. Örneğin AIDS hastalığında HIV virüsü taşıyan bir hasta başlangıç testinde pozitif olduğu andan itibaren gözlem altına alınmasına rağmen, başlangıçta HIV virüsünün bulaşma zamanı bilinemediğinden sağkalım zamanı soldan sansürlü denilmektedir.

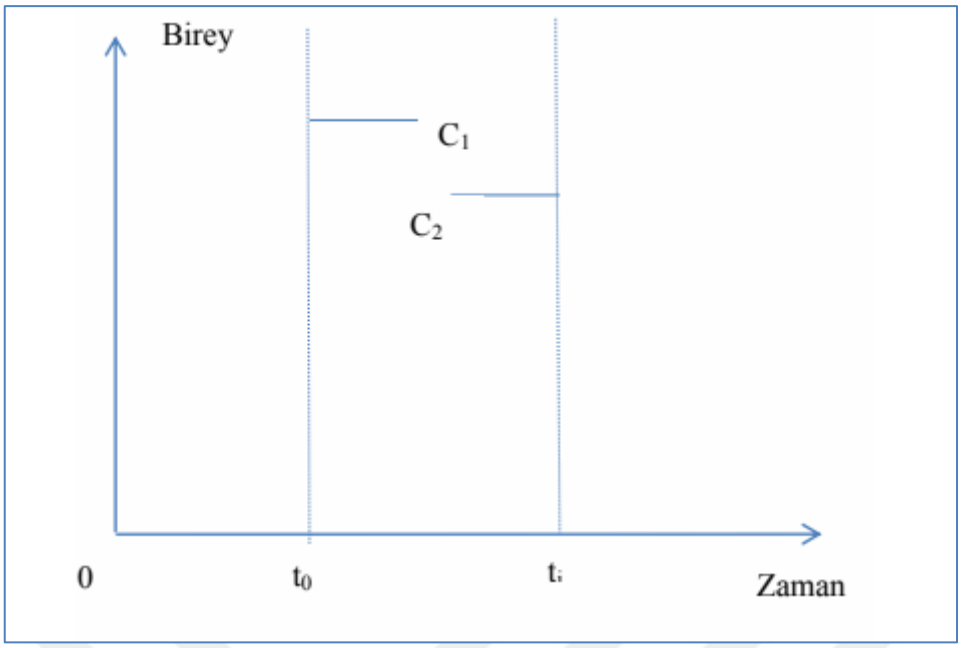


Şekil 2.2. Soldan sansürlü veri

Aralıklı Sansürleme

Aralıklı sansürlemesi, sansürleme çeşitlerinden genelleştirilmiş olanıdır. Takip gerektiren olaylarda aralıklı sansürleme genellikle kullanılmaktadır. Araştırılan konunun incelenen olayın meydana gelme süresi, çalışmaya konu olan olayın meydana gelme süresi, bir aralıkta ifade edilmesidir. Aralıklı sansürlemede ise olay t1 ve t2 gibi iki zaman dilimi arasında gerçekleşmektedir.

Sağkalım süresi $(R_i, L_i]$ aralığında yer alır. Sağdan sansürlemenin genelleştirilmiş biçimi olarak ifade edilmesi durumunda sol sınır noktası 0, sağ sınır noktası ise L_i olarak alınır. Eğer soldan sansürlemenin genelleştirilmiş biçimi olarak ifade edilirse sol sınır noktası L_i sağ sınır noktası ise (∞) olarak ele alınır. Kısacası aralık sansürlemesi, hem sağdan sansürlemenin hemde soldan sansürlemenin genelleştirilmiş şeklidir (Nelson, 1982).



Şekil 2.3. Aralıklı sansürlü veri seti

2.1.2. Sağkalım Zaman Fonksiyonları

Sağkalım süresi, bir araştırmada ilgilenilen olayın gerçekleşmesi anına kadar geçen zaman olarak ifade edilir. Sağkalım zamanının dağılımı birbiri ile ilişkili üç matematiksel fonksiyon ile tanımlanır. Bunlar sağkalım fonksiyonu, olasılık yoğunluk fonksiyonu, hazard fonksiyonudur.

2.1.2.1. Sağkalım Fonksiyonu

Bir bireyin sağkalım süresi T ile gösterilsin. Sağkalım fonksiyonu S(t) olsun ve T zamanından daha uzun sürede hayatta kalma olasılığını göstersin.

$$S(t) = P(T > t) \, , \, 0 \leq t < \infty \tag{2.4}$$

S(t) azalan bir fonksiyondur.

$$S(t) = \begin{cases} 0, & t = 0 \\ 1, & t = \infty \end{cases} \tag{2.5}$$

Zamanının sıfır olması durumunda tüm bireyler sağ olduğu için sağkalım olasılığı 1'dir. Süre ilerledikçe bireyin sağkalma olasılığı düşecek ve sonsuza doğru gittikçe bireyin yaşaması mümkün olamayacağından sağkalma olasılığı sıfır olacaktır.

2.1.2.2. Olasılık Yoğunluk Fonksiyonu

T sağkalım süresinin olasılık yoğunluk fonksiyonu f(t) ile kümülatif dağılım fonksiyonu ise F(t) ile gösterilir.

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t<T\leq t+\Delta t)}{\Delta t} \tag{2.6}$$

$$F(t) = 1 - S(t) \tag{2.7}$$

2.1.2.3. Hazard Fonksiyonu

Hazard fonksiyonu, çok küçük bir zaman aralığında başarısızlık oranını vermektedir. Bir bireyin t zamanına kadar yaşaması koşulu altında, $\Delta t \to 0$ iken [t, t+Δt] aralığında ölmesi olasılığına hazard fonksiyonu denir (Karasoy, 2008). Matematiksel olarak eşitlik (2.8) de verildiği gibidir.

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t\leq T<(t+\Delta t)/T\geq t)}{\Delta t} \tag{2.8}$$

Hazard fonksiyonu, sağkalım fonksiyonu ve olasılık yoğunluk fonksiyonu arasında eşitlik (2.8) daki gibi bir ilişki vardır. Hazard fonksiyonu ani başarısızlık riski, ölüm oranı gücü, şartlı ölüm oranı ve belirli yaştaki başarısızlık oranıdır.

2.1.2.4. Fonksiyonlar Arasındaki İlişki

Sağkalım fonksiyonu, olasılık yoğunluk fonksiyonu ve hazard fonksiyonlarından biri biliniyor ise diğerleri de bulunabilir. Bu fonksiyonlar aralarındaki ilişkiye göre yazılabilir (Lee, 1992)

$$h(t) = \frac{f(t)}{S(t)} \tag{2.9}$$

Olasılık yoğunluk fonksiyonu birikimli dağılım fonksiyonunun türevidir.

$$f(t) = \frac{d}{dt}[1 - S(t)] = -S'(t) \tag{2.10}$$

Buna göre

$$h(t) = -\frac{S'(t)}{S(t)} = -\frac{d}{dt}\log\big(S(t)\big) \tag{2.11}$$

S(0)=1 olması durumu kullanılarak 0'dan t'ye integral alınırsa,

$$-\int_0^t h(u)du = \log\big(S(t)\big) \tag{2.12}$$

elde edilir.

Böylece birikimli hazard fonksiyonu,

$$H(t) = -\ (\log\big(S(t)\big) \tag{2.13}$$

olarak bulunur.

$$S(t) = \exp[-H(t)] = \exp\left[-\int_0^t h(u)du\right] \tag{2.14}$$

$$f(t) = h(t)exp[-H(t)] \tag{2.15}$$

Örnek: T sağkalım zamanı λ parametreli üstel dağılıma uyuyorsa olasılık yoğunluk fonksiyonu aşağıdaki gibi tanımlanır.

$$f(t) = \begin{cases} \lambda e^{-\lambda t}, & t \geq 0, \lambda > 0 \\ 0, & dd \end{cases} \tag{2.16}$$

Birikimli dağılım fonksiyonu,

$$F(t) = 1 - S(\mathrm{t}) \tag{2.17}$$

$$F(t) = \int_0^x \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x} \tag{2.18}$$

Sağkalım fonksiyonu,

$$S(t) = e^{-\lambda t} \tag{2.19}$$

Hazard fonksiyonu,

$$h(t) = -\frac{S'(t)}{S(t)} = \lambda \tag{2.20}$$

olarak bulunur.

2.2. Cox Regresyon Modeli

Sansürlü verilerinde olduğu sağkalım analizinde bağımlı değişkenle bağımsız değişkenler arasındaki sebep-sonuç ilişkisinin açıklanmasına yardımcı olan regresyon yöntemine Cox regresyon denir. Yöntem, başta tıp olmak üzere yaygın biçimde kullanılmaktadır.

Bir araştırmada incelenen bağımlı değişken, bir hastalığa yakalanan bireylerin ölüm zamanlarına kadar geçen gözlem süreleri ise; bağımsız değişkenler, bağımlı değişken üzerinde etki yaratan değişkenler olur. Sağkalım analiz yöntemleri içerisinde Cox Regresyon Yöntemi sağkalım süresini etkileyen faktörlerin ortaya konması açısından güçlü bir istatistiksel yöntemdir. Tıp alanında yapılan pek çok hastalıkta olduğu gibi bir hastalığın varlığı ya da yokluğu tek başına sağkalım ya da ölüm açısından belirleyici değildir. Hastalığın dışında bireylerin yaşı, kilosu, hastalık tipi, çevresel faktörler ve daha pek çok özelliği sağkalım süresini uzatabilir veya kısaltabilir.

Birçok faktörü birlikte değerlendirmede en çok bilinen yöntem çoklu regresyon analizidir. Ancak sağkalım analizi çalışmalarında sansürlü veriler yer almaktadır. Çoklu regresyon modelleri, sonuç açısından olgular arasında zaman içinde oluşan bu önemli farkı dikkate alarak değerlendirme yapma imkânına sahip değildir (Terzi ve ark.,2005). Zaman içindeki değişimi dikkate alan ve sansürlü verilerle analiz yapmayı kolaylaştıran

bir yönteme ihtiyaç duyulmuştur. Bu yöntem Cox (1972) tarafından önerilen Cox regresyon modelidir. Cox Regresyon modeli (2.21) eşitliği ile verilir.

$$h(t;x) = h_0(t)exp(\beta' \boldsymbol{x}) \tag{2.21}$$

Burada h(t; x), hazard fonksiyonu **x**, px1 boyutlu bağımsız değişkenler vektörü $\beta' x$, 1xp boyutlu bilinmeyen regresyon katsayıları vektörü h_0(t), temel hazard fonksiyonudur. Modelde yer alan h_0 (t), **x**'lerden bağımsız, zamana bağlı parametrik olmayan tahminleri veren bir fonksiyondur.

Cox regresyonun en önemli varsayımı orantılı hazard varsayımıdır. Bu varsayıma göre herhangi iki bireyin hazard oranı modelde zaman ekseni boyunca sabittir. Bu varsayımdan dolayı Cox regresyon modeli orantılı hazard modeli olarak da bilinmektedir. Hazard fonksiyonunun temel hazard fonksiyonuna oranı nispi risk olarak adlandırılır.

$$RR = \frac{h(t)}{h_0(t)} = exp\{\sum_{i=1}^{p} \beta_i x_i\} \tag{2.22}$$

Bu eşitlikde $exp\{\sum_{i=1}^{p} \beta_i x_i\}$ bağımsız değişkenlerdeki artışa karşılık, nispi olarak risk oranındaki yüzde değişimi belirtir. Yani risk oranındaki yüzde değişim, bağımsız değişkendeki bir birimlik artışa karşılık gelir. Eşitlikte sağ tarafta t olmadığından, nispi risk tüm zaman değerleri için sabittir. Regresyon katsayıları sıfır olduğunda nispi risk bire eşit olur. Burada RR nin her iki tarafında logaritması alınırsa;

$$log\left[\frac{h(t)}{h_0(t)}\right] = \sum_{i=1}^{p} \beta_i X_i = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{2.23}$$

elde edilir. Birikimli hazard fonksiyonu,

$$H(t;x) = \int_0^t h_0(u)exp\left[\sum_{i=1}^{p} \beta_i x_i\right] du$$

$$= exp\left[\sum_{i=1}^{p} \beta_i X_i\right] \int_0^t h_0(u)du$$

$$=H_0(t)exp\left[\sum_{i=1}^{p}\beta_i x_i\right] \quad (2.24)$$

ile verilir. Birikimli sağkalım fonksiyonu,

$$S(t;x)=exp[-H_0(t)exp(\sum \beta_i x_i)] \quad (2.25)$$

$$=[S_0(t)]^{exp(\sum_{i=1}^{p}\beta_i x_i)} \quad (2.26)$$

ile verilir. $S_0(t)$ temel sağkalım fonksiyonudur. Sağkalım fonksiyonu ile temel sağkalım fonksiyonu arasında üstel bir ilişki vardır. Olasılık yoğunluk fonksiyonu eşitlik;

$$f(t;\ x) = h_{0(t)}exp\left(\sum_{i=1}^{p}\beta_i x_i\right)exp\left[-H_0(t)exp\left(\sum_{i=1}^{p}\beta_i x_i\right)\right] \quad (2.27)$$

dir.

2.2.1. Parametre Tahmini

En çok olabilirlik yöntemi, parametre tahmininde en sık kullanılmakta olan yöntemdir. En çok olabilirlik yöntemi olabilirlik fonksiyonunu en yüksek yapan parametre tahminlerini belirlemek için kullanılır. Cox Regresyon modelinin parametre tahmininde olabilirlik fonksiyonu yerine kısmi olabilirlik fonksiyonu kullanılır. Kısmi olabilirlik kavramı, olasılıkların sadece tamamlanmış olan birimler için ele alınmasından kaynaklanmaktadır (Kleinbaum ve Klein, 1996).

Gözlemlenen n tane sağ kalım zamanı arasından k tanesi küçükten büyüğe sıralanmış olarak $t_1 < t_2 < \cdots < t_k$, ölüm sonucu olan verileri göstersin. t_i zamanında değerleri tespit edilen x_i açıklayıcı değişken vektörü belirlenmiş olsun. Bir açıklayıcı değişkenin, sağkalım zamanı üzerine etkide bulunan tüm değişkenler dikkate alınarak belirlenecek genel risk içindeki oranı riskler oranı biçiminde belirlenir ve bu oran;

$$Riskler\ oranı = \frac{\exp(\beta' x_i)}{\sum_{j\epsilon R(t_i)}\exp(\beta' x_j)} \quad (2.28)$$

Kısmi olabilirlik fonksiyonu;

$$L(\beta) = \prod_{i=1}^{k}\frac{\exp(\beta' x_i)}{\sum_{j\epsilon R(t_i)}\exp(\beta' x_j)} \quad (2.29)$$

Logaritmik kısmi olabilirlik fonksiyonu;

$$\ell(\beta) = lnL(\beta) = \sum_{i=1}^{k} \beta' x_i - \sum_{i=1}^{k} ln\left[\sum_{j\epsilon R(t_i)} \text{ex}\ell(\beta) = \ln L(\beta)\, \text{p}(\beta' x_j)\right] \quad (2.30)$$

Yukarıdaki eşitlikte β katsayılarına göre türev alınır. Olabilirlik fonksiyonunun birinci türevi ile eşitlik aşağıda verilen skor fonksiyonu elde edilir;

$$U(\beta) = \frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^{k} x_i - \sum_{i=1}^{k} \frac{\sum_{j\epsilon R(t_i)} \exp(\beta' x_j) x_j}{\sum_{j\epsilon R(t_i)} \exp(\beta' x_j)} \quad (2.31)$$

Veride benzer süreli gözlemler olduğunda logaritmik kısmi olabilirlik fonksiyonunun maksimum yapılmasında farklı yaklaşımlar önerilmiştir. Bunlar Breslow (1974) ve Efron (1975)'un önerdiği kısmi olabilirlik fonksiyonlarıdır. Birçok uygulamada Breslow ve Efron yaklasımlarından elde edilen parametre tahminleri arasında oldukça küçük ve önemsiz farklılıklar bulunmuştur (Lee ve Wang, 2003). Bu nedenle çalışmamızda bunlar arasında en yaygın bir yaklaşım olan Breslow kullanıldı.

$$L_B(\beta) = \prod_{i=1}^{k} \frac{\exp(\beta' x_i)}{\left[\sum_{j\epsilon R(t_i)} \exp(\beta' x_j)\right]^{d_i}} \quad (2.32)$$

Burada d_i, $t_i$zamanındaki ölü gözlem sayısını R(t_i), $d_i$zamanından önce risk altındaki bireylerin sayısını göstermektedir. β 'ların en çok olabilirlik tahminleri $\hat{\beta}$ ile gösterilir. Elde edilen tahminlerin anlamlı olup olmadıklarının test edilmesi gerekmektedir.

2.2.2. Parametrelerin Önemlilik Testleri

Cox regresyon analizinde parametreler tahmin edildikten sonra ve parametrelerin önemlilik testi için H_0 : β =0 hipotezi test edilir. Bu hipotezi test etmek için

• Olabilirlik Oran Testi

• Wald Testi

• Skor Testi

kullanılır.

2.2.2.1. Olabilirlik Oran Testi

Değişkenlerin iki veya daha fazla olduğu durumlarda ve Cox modelinde aynı anda birden çok açıklayıcı değişkenin test edilmesinde kullanılır. Cox regresyon modelinde; regresyon katsayılarının önemliliği $H_0: \beta_i$ i=1,2, … ,kvarsayımı test edilerek bulunur. Bu modele sıfır model denilmektedir. Modelin en çok olabilirlik değeri L_0 ile gösterilmektedir. Modelde i tane birbirinden farklı ve sağ kalım süresi üzerine önemli etkide bulunan başka bir modelin en çok olabilirlik değeri $L_i$belirlenerek olabilirlik oran testi hesaplanır.

$$LR = -2\ln\left[\frac{L_0}{L_i}\right] \sim \chi_p^2$$

$$\text{LR=-2ln}\left[\frac{tam\ model\ için\ L}{indirgenmiş\ model\ için\ L}\right] \sim \chi^2_{(p-p_1)} \qquad (2.33)$$

Olabilirlik oran testi p serbestlik dereceli $(sd = p)$ ki –kare dağılımı göstermektedir. Önemliliğinin tespit edilmesinde $p$serbestlik dereceli ki-kare dağılımının kritik değerlerinden yararlanılır.

2.2.2.2. Wald Testi

Wald testi, en çok olabilirlik tahminlerinin örnekleme dağılımının normal varsayımına dayanmaktadır ve tahmin edilen parametrenin standart hatasına oranlanması ile Wald test istatistiği bulunur.

$$Z = \frac{\widehat{\beta}}{SH_{\widehat{\beta}}} \qquad (2.34)$$

$$W = Z^2 = \left[\frac{\widehat{\beta}}{SH_{\widehat{\beta}}}\right]^2 = \frac{\widehat{\beta}^2}{Var(\widehat{\beta})} \qquad (2.35)$$

ile elde edilir. Yani standart normal dağılımın tamamının karesine eşittir. Bu test istatistiği Z_H testi standart normal dağılım gösterir ($Z_H \sim N(0,1)$). Böylece Wald test

istatistiği 1 serbestlik dereceli ki-kare dağılımı gösterir ve 1 serbestlik dereceli ki-kare dağılımının kritik değerleri ile karşılaştırılarak önemliliği belirlenir.

2.2.2.3. Skor Testi

Skor test istatistiği, logaritmik olabilirlik fonksiyonlarından yararlanılarak hesaplanmaktadır. Modeli oluşturan değişkenlerin sürekli olduklarında veya birden fazla değişken bulunduğunda kullanılır. Test istatistiği;

$$S = \frac{\partial \ell(\beta)/\partial \beta}{\sqrt{I(\beta)}} \tag{2.36}$$

biçimindedir $\ell(\beta)$ logaritmik kısmi olabilirlik, $I(\beta)$, bilgi matrisidir. Skor istatistiği standart normal dağılım gösterir. Bu istatistiğin karesi, 1 serbestlik dereceli ki-kare dağılımı gösterir (Kleinbaum ve Klein, 1996).

Skor test istatistiği de Olabilirlik oran testi ve Wald istatistiği gibi p tahmin edilmekte olan parametre sayısı olmak üzere p serbestlik dereceli ki-kare dağılımı gösterir. Önemliliği p serbestlik dereceli ki-kare dağılımının kritik değeri ile karşılaştırılarak yapılır (Özdamar, 2010).

3. METERYAL VE YÖNTEMLER

3.1. Kayıp Veri

İstatistiksel analiz yapacak araştırmacıların karşılaşacağı sorunlardan biri kayıp veri problemidir. Kayıp veri, araştırma sürecinde ilgilenilen değişkenin bazı değerlerinin gözlenememesidir. Araştırmaya ilişkin belli bir değişkene ait özellik çalışma sürecinde elde edilemeyebilir. Örneğin bilgi toplamak amacıyla yapılan anketlerde cevaplayıcı soruya yanıt vermek istemeyebilir, soruyu yönelttiğimiz kişinin sorunun cevabı hakkında bir fikri olmayabilir ya da ankette kişinin durumu ile ilgisi olmayan sorular yöneltilmiş olabilir. Araştırma yapılan konulardan bilgi toplanmak istendiğinde idari kayıtlarda bir kayıplık ya da yanlış bilgi toplandığında da kayıplar ortaya çıkabilir. Araştırmalarda, tüm verileri tam olan kayıpsız veri setleri elde etmek her zaman kolay değildir. Araştırmalarda, veri setinde kayıp değer olması önemli bir problemdir. Çünkü istatiksel metotlar ve bilgisayar programları tüm değerleri tam olan eksiksiz veri setleri için tasarlanmıştır. Bu nedenle veri setlerindeki kayıp değer problemi çözülmelidir (Alkan ve ark., 2013).

3.1.1. Kayıp Veri Mekanizması

Kayıp veri analizi yapılırken kayıp veri sorununu giderecek uygun yöntemin seçilmesi önemlidir. Kayıp değerin veri ile nasıl bir ilişki içinde olduğunu gösteren sınıflandırma kayıp veri mekanizması olarak adlandırılır. Kayıp veri mekanizmasına göre belirlenen uygun kayıp veri analiz yöntemi yansız ve doğru tahminler üretmektedir. Bu nedenle veri setindeki kayıplığın değişkenlerle olan ilişkisi yani kayıp verinin oluşumunu içeren kayıp veri mekanizmasını anlamak kayıp veri problemlerini gidermek için önemli adımlardan biridir. Çalışmanın durumuna göre kayıp verilerin oluşması farklılık gösterebilir. Rubin (1976), Little ve Rubin (2002) kayıp veri oluşumunu rasgele kayıplar (Missing at Random, MAR), tamamen rasgele kayıp (Missing Completely Random, MCAR) ve rasgele olmayan kayıplar (Missing not at Random, MNAR) olmak üzere 3 şekilde sınıflandırmıştır.

3.1.1.1. Rasgele Kayıp (Missing at Random, MAR)

X değişkeni üzerindeki kayıp veri olasılığı *X*'in kendisiyle ilgili olmayıp, modeldeki ölçülen bir başka değişken (ya da değişkenler) ile ilgili olduğunda veri rastgele kayıptır (MAR). Bir başka değişle, isminde geçen Rastgele terimi biraz çeldiricidir, çünkü verinin yazı tura atmaya benzer bir şekilde tesadüfi olarak kaybolduğunu çağrıştırır. Aslında MAR bir veya birden çok ölçülmüş değişken ile kayıp veri olasılığı arasında sistemli bir ilişki olduğu anlamına gelir (Demir, E., & Parlak, B., 2012).

Örneğin, bir eğitim araştırıcısının okuma başarısı çalıştığını ve İspanyol öğrencilerin Kafkas öğrencilerden daha yüksek bir kayıp veri oranına sahip olduğunu varsayalım. İkinci bir örnek olarak, bir psikoloğun bir grup kanser hastasının yaşam kalitesi ile ilgili çalışmasında yaşlı ve eğitim düzeyi düşük olan hastaların yaşam kalitesi sorusuna yanıtlamamaya eğilimli olduğu sonucuna varmaktadır. Verilen bu örneklerle kayıp değerler başka bir değişkenin etkisiyle oluşmaktadır. Yaşam kalitesi yaş ve eğitimi dışta tuttuktan sonra, kayıp veri olasılığı yaşam kalitesiyle ilişkili değildir. Bu nedenle bu verideki kayıp veri mekanizması MAR dır. Eğitim örneğinde ise, okuma başarısı ile kayıp veri arasında, etnik kimliği yokladıktan sonra bile bir ilişki vardır. Bu durum MAR mekanizmasıyla uyumsuzdur. Bununla birlikte, kayıp veri başarı sonuçlarını bilmedikçe, araştırmacının bu ilişkinin varlığı ya da yokluğunu kanıtlaması imkânsızdır.

Matematiksel olarak X, daima gözlenen değişken, Y kayıp veri içeren değişken olmak üzere MAR varsayımı eşitlik 3.1 de verildiği gibidir.

$$P(Y_{kayıp}/Y, X) = P(Y_{kayıp}/X) \quad (3.1)$$

Bu gösterim, Y ve X'in verilmesi durumunda Y'deki kayıp veri koşullu olasılığı ile sadece X'in verilmesi durumunda Y'deki kayıp veri koşullu olasılığının eşit olduğu anlamına gelmektedir.

3.1.1.2. Tamamen Rasgele Kayıp (Missing Completely At Random, MCAR)

MCAR, test edilebilir olan tek kayıp veri mekanizmasıdır. Tamamen rasgele kayıp mekanizmasında Y değişkeni üzerindeki kayıp verilerin olasılığı, diğer ölçülen değişkenlerle ve Y değişkeninin değerleri ile ilişkili değildir. MCAR, MAR'dan daha

kısıtlayıcıdır çünkü kayıp olma durumu tamamen veriden ilişkisiz olduğunu varsayar (Enders, 2010). Yani veri seti veya setleri oluşturulurken istek dışı gerçekleşen veri kayıplarının tümüne denilmektedir.

Örneğin bir anket çalışmasında soruyu görmeyip cevaplamama veya istenmeyen durumlarda verilerden bazılarının kaybolması gibi durumlarda MCAR'a uygun kayıp oluşumudur. Yada eğitim örneğinde çocukların beklenmedik kişisel olaylar (örneğin bir hastalık, bir cenaze, aile tatili, başka bir okul bölgesine taşınma) veya zamanlama sorunları (araştırmacılar okulu ziyaret ettiğinde sınıf bir okul gezisinde olması) nedenleriyle başarı puanlarının kayıp olması durumu MCAR'dır. Bir başka örnekte bir psikolog bir grup kanser hastasında yaşam kalitesini araştırmakta ve hastalara yaşam kalitesi anketini uygulamaktadır. Anketteki yaşam kalitesi ile ilgili soruları cevaplamayı reddedenlerin yaş ortalaması anketi tamamlayan hastalara göre daha düşüktür. Bu ortalama farklar, verilerin MCAR olmadığına dair güçlü kanıtlar sunar ve demografik değişkenler ile kayıp verilerin olasılığı arasında olası bir ilişki olduğunu gösterir.

3.1.1.3. Rasgele Olmayan Kayıp (Missing not at Random, MNAR)

Y değişkeninde kayıp veri olasılığı, Y değerleriyle ilişkili olduğunda veriler rastgele değildir (MNAR). Sonuç olarak, iş performansı değerlendirme değişkenindeki kayıp olma olasılığı kişinin kendi iş performansına bağlıdır (Craig K. Enders, 2010).

Eğitim örneği için zayıf okuma becerisine sahip öğrencilerin okuma başarısı değişkeninde kayıp değere sahip olması durumu rastgele olmayan kayıptır. Çünkü okuma anlama zorlukları yaşadıkları için cevap verilmemiş olabilir. MAR mekanizması gibi, kayıp değişkenlerin değerlerini bilinmeden puanların MNAR olduğunu doğrulamanın bir yolu yoktur.

Kayıp veri mekanizmasını daha anlaşılır olması açısından Çizelge 3.1.'de iş performans örneği verilmiştir. Bu örnekte 20 kişi bir şirkete iş başvurusunda bulunmuştur. Şirket başvuruda bulunanlara IQ testi uygulamış ve daha sonra 6 aylık deneme süresine tabi tutmuştur. Şirket tarafından adayların deneme süresi sonunda iş performansları değerlendirilmiştir.

Çizelge 3.1. İş Performans Örneğinde Kayıp Veri Mekanizması

	İş Performans Değerlendirme			
IQ	TAM VERİ	MCAR	MAR	MNAR
78	9	-	-	9
84	13	13	-	13
84	10	-	-	10
85	8	8	-	-
87	7	7	-	-
91	7	7	7	-
92	9	9	9	9
94	9	9	9	9
94	11	11	11	11
96	7	-	7	-
99	7	7	10	-
105	10	10	10	10
105	11	11	11	11
106	15	15	15	15
108	10	10	10	10
112	10	-	10	10
113	12	12	12	12
115	14	14	14	14
118	16	16	16	16
134	12	-	12	12

MAR sütunundaki iş performans değerleri, düşük IQ puanları olan adaylarda kayıp olduğunu görülmektedir. Yani iş performans değişkenindeki kayıp, başka bir değişkenden kaynaklı oluşmaktadır.

Çizelge 3.1.'de rastgele puanlar silinmiş ve MCAR sütunu oluşturulmuştur. Silinen rastgele sayılar IQ ve iş performansı ile ilişkili olmadığından kayıp olma durumu verilerden ilişkisizdir. Örneğin, bir çalışan 6 aylık değerlendirme öncesi doğum izni alabilir veya eşi başka bir eyalette iş kabul ettiği için şirketten ayrılabilir ya da adayı değerlendiren şirket elemanı başka bir birime terfi etmiş olabilir. Yani iş performans değişkenindeki kayıp veriden tamamen bağımsız oluşmaktadır.

Kayıp veri mekanizmasının MNAR olması durumunda ise adayların 6 aylık deneme süresi sonundaki iş performans puanı 8 ve altında olanların işlerine son verilmesi sonucu oluşan kayıptır. Yani iş performans değişkenindeki kayıp kişinin iş performansına bağlı oluşmuştur.

3.1.2. Kayıp Veri Sürecinde Rasgeleliğin Sorgulanması

Kayıp veri mekanizmaları içerisinde test edilebilen tek yöntem tamamen rasgele kayıp (MCAR)'dır. Rasgele Kayıp (MAR) ve Rasgele Olmayan Kayıp (MNAR) mekanizmaları için herhangi bir test bulunmamaktadır. Kayıplığın nedenini belirlemek için MCAR testlerini kullanmak, veri setindeki kayıp veri mekanizmasının MCAR olup olmadığını değerlendirmekle ilgilenilmiyor olsa bile kayıp veri mekanizmasının sorgulanmasında iyi bir başlangıç noktası olacaktır. Çünkü araştırmacının bilgisi dâhilinde planlı kayıp veri tasarımı söz konusu ise kayıplık araştırmacının kontrolünde olacağından MAR ve MNAR teşhisi konulabilir.

Yöntem geliştiren bilim adamları MCAR mekanizmasını test etmek için birkaç yöntem önermiştir (Chen & Little, 1999; Diggle, 1989; Dixon, 1988; Kim & Bentler, 2002; Little, 1988; Muthén, Kaplan, & Hollis, 1987; Park & Lee, 1997; Thoemmes & Enders, 2007). Bu yöntemler Tek Değişkenli t-Testi ve Little 'nin MCAR Testidir.

3.1.2.1. Tek Değişkenli *t*-Testi Karşılaştırmaları

T-testi, veri setinin MCAR varsayımına uygun kayıplık olduğunu test etmek için kullanılan en basit değerlendirme yöntemidir. Tek değişkenli *t*-esti karşılaştırmaları belirli bir değişken üzerinde kayıp ve kayıp değer içermeyen olayları ayırıp veri setindeki diğer değişkenlerdeki grup ortalama farklarını analiz eder. MCAR mekanizması gözlenen verili olayın ortalamasının kayıp verili olayla aynı olması gerektiğini vurgular.

t-testi yaklaşımında göz önünde bulundurulması gereken bazı problemler vardır; Birincisi, süreci otomatikleştiren yazılım paketiniz (yani, SPSS Kayıp Değer Çözümleme modülü) olmadıkça, test istatistiklerini türetmek çok yavaş olabilir. İkincisi, t-test istatistiği değişkenler aralarındaki korelasyonları hesaba katmaz, dolayısıyla bir kayıp veri setinin, verideki kaybın tek bir neden olsa bile, bir veri setindeki değişkende ortalama farkları üretmesi mümkündür.

Bir test güvenilirliği dengelemek için, Cohen'in (1988) *t* testi uygulamaları kullanılabilir. MAR ve MNAR mekanizmaları eşit ortalamalı kayıp veri alt grupları türetebildiği için, ortalama karşılaştırmalarının kesin bağlayıcı bir MCAR testini sağlamayacağını belirtmekte yarar vardır.

3.1.2.2. Little 'nin MCAR Testi

Little (1988) veri setindeki her değişkenin ortalama farklarını değerlendiren *t* testi yaklaşımının çok değişkenli bir yöntemidir. Tek değişkenli *t* testinin aksine, Little'ın MCAR testi tüm veri setine uygulanan genel bir MCAR testidir ve Little'ın MCAR Testi, bazı istatistik yazılım paketlerinde (ör., SPSS Kayıp Veri Çözümleme modülü) mevcuttur. Little'ın test istatistiği aynı *t*-testi yaklaşımı gibi, kayıp veri setinin alt gruplarındaki gibi ortalama farkları değerlendirmektedir. Test istatistiği, alt grup ortalamaları ile toplam ortalaması arasındaki standart farkların standartlaştırılmış ağırlıklı bir toplamıdır ve eşitlikte verildiği gibidir.

$$d^2 = \sum_{j=1}^{J} n_j \left(\hat{\mu}_j - \hat{\mu}_j^{(ML)}\right)^T \hat{\Sigma}_j^{-1} \left(\hat{\mu}_j - \hat{\mu}_j^{(ML)}\right) \quad (3.2)$$

burada n_j, kayıp veri j'deki vakaların sayısıdır J , $\hat{\mu}_j$değişkenler için içerir. Kayıp veri durumunda j, $\hat{\mu}_j^{(ML)}$Genel en çok olabilirlik tahminlerini içerir. $\hat{\Sigma}_j$kovaryans matrisinin en çok olabilirlik tahminidir. J parametre matrislerindeki öğe sayısının kayıp veriler arasında değiştiğini gösterir. d^2İstatistiği, z skorlarının karelerinin ağırlıklı bir toplamıdır.

H_0 hipotez doğru olduğunda d^2, yaklaşık Σk_j - k serbestlik derecesine sahip ki-kare olarak dağılır, burada k_j, j modeli için tam değişkenlerin k, toplam değişkenlerin sayısıdır. Little' nin MCAR testi rasgeleliğin bulunmasında sıklıkla kullanılır ve P<0,05 olduğunda veri setinin MCAR olmadığı sonucuna varılır.

3.2. Kayıp Veri Problemini Çözen Yöntemler

Araştırmacılar yüz yıllardır kayıp veri sorunu ile mücadele etmiş ve sorunun çözümü için pek çok yöntem önermişlerdir. Bu yaklaşımlardan çoğu yaygın kullanım görürken, diğerleri neredeyse tarihteki bir dipnot konumundadır (Little & Rubin, 2002; Wilkinson, 1999). Kayıp veri problemini çözen yöntemler genellikle kayıp veri analizi olarak adlandırılır. Kayıp veri analiz yöntemleri, problemi kayıp verisi olan gözlemleri silerek veya kayıp değerler yerine değer atayarak ele almaktadır.

Kayıp değerli gözlemlerin silindiği en yaygın kullanılan yöntemler liste bazında silme ve çiftler bazında silme yöntemidir. Kayıp veriyi silmek istatistik yazılım paketlerinde uygulanan bir stratejidir ve psikoloji ile eğitim gibi aşırı yaygın disiplinlerde sık sık kullanılmaktadır (Peugh & Enders, 2004).

Değer atama, veri setindeki değişkenlerin gözlenen değerlerinin kullanılması ile kayıp değerlerin tahmin edilmesidir. En çok kullanılan değer atama yöntemleri; ortalama değer atama (Mean İmputation) regresyon değer atama (Regression İmputation) Beklenti Maksimizayon Algoritması (EM Algorithm), Çoklu Değer Atama (Multiple İmputation) yöntemleridir.

3.2.1. Silme Yöntemleri

Sosyal bilimler ve davranış bilimlerinin birçok alanında Liste ve çiftler bazında silme, en yaygın kullanılan kayıp veri analiz yöntemleridir (Peugh & Enders, 2004). İstatistiksel yazılım paketlerinde bu yöntemlerin avantajı, kolay uygulanabilir olmalarıdır. Ancak, silme yöntemlerinin ciddi kısıtlamaları vardır çoğu durumda kullanımlarını engeller. En önemlisi, bu yaklaşımlar MCAR'ı kabul eder. Bu varsayım tutmadığında veri yanlı parametre tahminleri üretebilir. Kayıp veri mekanizması MCAR olsa bile kayıp değerli gözlemlerin silinmesi örnek genişliğini küçültmekte bu da önemli ölçüde testin gücünü azaltabilmektedir.

3.2.1.1 Liste Bazında Silme

Kayıp veri içeren gözlemin çalışmadan çıkarılmasıdır. Kolay bir yöntem olması nedeniyle araştırmacılar tarafından sıklıkla uygulanır. Liste bazında silme yönteminin bir takım özellikleri vardır. Liste bazında silme yöntemi uygulanmış veri setinde kayıp değerli gözlemlerin silinmesi nedeni ile hesaplanmış standart hatalar daha az veri kullanıldığı için daha yüksek hesaplanmasına ve daha geniş güven aralıklarına neden olmaktadır (Allison, 2000). Liste Bazında silme yöntemi, doğrusal, lojistik, poisson, Cox gibi regresyon analizi çeşitlerinde MCAR varsayımı ihlal edilse bile yansız parametre tahminleri üretmektedir (Little, 1992).

3.2.1.2. Çiftler Bazında Silme Yöntemi

Yöntemin amacı ortalamalar, standart sapmalar ve korelasyon matrisi gibi özet istatistikleri ulaşılabilen tüm veri setlerine uygulayarak hesaplamaktır. Yani her değişken çifti için tüm değerleri tam olan gözlemler kullanılarak korelasyon ya da kovaryans hesaplanmaktadır. Başka bir değişken çifti için hesaba dahil olan gözlemler değişebilir. Liste bazında silme ile benzer olarak çiftler bazında silme yönteminde de kayıp veri mekanizması MCAR olmalıdır.

3.2.2. Değer Atama Yöntemleri

3.2.2.1.Ortalama Değer Atama (Mean Imputation)

Ortalama yer veya koşulsuz ortalama olarak da bilinen ortalama değer atama, elde bulunan mevcut verilerin aritmetik ortalaması ile kayıp değerlerin tamamlanması işlemidir. Kayıp değerleri ortalama ile değiştirme fikri, Wilks(1932) tarafından geliştirilmiş en eski yöntemlerden biridir. Diğer teknikler gibi, ortalama değer atama yöntemi de kayıpsız veri seti oluşturduğu için uygundur. Fakat uygulaması kolay olmasına rağmen her zaman güvenilir sonuçlar vermez. Veri setindeki kayıplık MCAR olsa bile parametre tahminleri yanlı olabilmektedir.

3.2.2.2. Regresyon Değer Atama (Regression Imputation)

Regresyon değer atama (diğer adıyla koşullu ortalama değer atama) yöntemi kayıp verilerin yerine bir regresyon denkleminden tahmin edilen sonuçları yerleştirir. Aritmetik ortalama değer atama yönteminde olduğu gibi regresyon değer atama yöntemi de yaklaşık 50 yıla yakın uzun bir geçmişi vardır (Buck, 1960). Bu yaklaşımın ardındaki temel düşünce, değişkendeki kayıp değerleri tamamlamak için kayıp değer içermeyen değişkenlerden gelen bilgiyi kullanmaktır. Yani yöntemde tahmin edilen regresyon modeli, kayıp gözlemlerin tahmin edilmesinde bir araç olarak kullanılmaktadır. Değişkenler birbiri ile ilişkili olma eğilimindedir bu yüzden gözlenmiş veriden alınan bilgi ile atanan değerleri üretmek iyi bir yoldur. Regresyon atama yönteminin sürecinde ilk adımı tam değişkenlerden kayıp değişkenleri tahmin eden bir regresyon denklemleri setini oluşturmaktır. Liste bazında silme yönteminden sonra bu denklemler tahmin edilir. İkinci adım kayıp değişkenler için tahmini değerleri türetmektir. Bu tahmini

değerler kayıp değerlerin yerini doldurur ve tam bir veri seti oluşturur. Örneklendirmek gerekirse iş performans örneğinde iş performans oranlarının IQ üzerindeki regresyonu oluşturmak için 10 tam gözlem kullanılarak regresyon denklemi tahmin edildi. Tahmin edilen regresyon denklemi eşitlik (3.3) da verilmiştir.

$$\widehat{JP_i^*} = \widehat{\beta_0} + \widehat{\beta_1}(IQ_i) = -2.065 + 0{,}123\ (IQ_i) \qquad (3.3)$$

Burada JP_i^* i. gözlem için tahmin edilen iş performans oranıdır. Böylece kayıp değerlere bu denklem yardımıyla değer atanmış veri seti tamamlanmış olur. Regresyon değer atama yöntemi çok değişkenli veri seti içinde büyük ölçüde aynıdır, ancak uygulanması biraz daha karmaşıktır. Denklemleri elde etmenin kolay bir yolu, ortalama vektör ve kovaryans matrisinin bir tahmini ile başlamaktır, çünkü bu matrislerdeki elemanlar gerekli tüm regresyon katsayılarını tanımlar. Tek değişkenli örnekte olduğu gibi Liste bazında silme yönteminden sonra $\hat{\mu}$ ve $\hat{\Sigma}$ elde edilir. Gözlemlenen değerlerin ilgili regresyon denklemlerine yerleştirilmesi ile kayıp değişkenler için tahmini değerler üretilir ve bu atanan değerler ile kayıp değer içermeyen tam veri seti oluşturulur.

3.2.2.3.Beklenti Maksimizasyonu-EM Algoritması

Beklenti maksimizasyonu-EM algoritması, kayıp veri çözümlemesinde önemli bir yöntemdir. EM algoritmasının kökenleri, 1970'lere kadar gider (Beale & Little, 1975; Finkbeiner, 1979; Dempster, Laird, & Rubin, 1977; Hartley & Hocking,1971; Orchard & Woodbury, 1972) Dempster ve arkadaşlarının 1977 yılındaki çalışması algoritmayı geliştirmede önemli bir rol oynar. EM uygulamaları temelde kayıp verili ortalama vektör ile kovaryans matrisinin tahminine dayanmaktadır. Fakat araştırmacılar algoritmayı, çok katlı modelleri ve yapısal eşitlik modellerini de kapsayan karmaşık veri sorunlarını da ele alacak biçimde geliştirmiştir. (Jamshidian & Bentler, 1999; Liang & Bentler, 2004; McLachlan & Krishnan, 1997; Muthén & Shedden, 1999; Raudenbush & Bryk, 2002).

EM Algoritmasının her tekrarı iki adımda gerçekleşir. Bu adımlar, beklenti (E-Adımı) ve maksimizasyon (M Adımı) olarak adlandırılır. Bu süreç ortalama vektör ve

kovaryans matrisinin başlangıç tahmini ile başlar. E Adımı kayıp değişkenleri gözlenen değişkenlerden tahmin eden bir regresyon eşitlikleri setini oluşturmak için ortalama vektörünü ve kovaryans matrisini kullanır. Ardından M Adımı, ortalama vektör ve kovaryans matrisinin güncellenmiş tahminlerini elde etmek ve verideki kayıp değerleri tamamlamak için standart tam veri formüllerini uygular. Algoritma güncellenmiş parametre tahminlerini, kayıp değerleri tahmin etmek için yeni bir regresyon eşitlikleri setini kurduğu bir sonraki E adımına taşır. Daha sonra M adımı, ortalama vektörü ve kovaryans matrisini yeniden tahmin eder. EM bu iki adımı $\hat{\mu}$ ve $\hat{\Sigma}$ deki elemanları artık değişmeyinceye kadar yineler. Bu noktada algoritma en çok olabilirlik tahmin edicilerine yakınsamıştır (Enders, 2003; Yuan & Bentler, 2000).

Basit bir tahmin problemi, temel görüşlerin genelde kayıp veri setini çok değişkenli çözümlemelerine dayanmaktadır. Tam veri (x) ve kayıp veri (y) içeren aşağıdaki formüller ortalama, varyans ve kovaryansın en çok olabilirlik tahminlerini türetir.

$$\widehat{\mu_y} = \frac{1}{N}\sum Y \tag{3.4}$$

$$\widehat{\sigma_y^2} = \frac{1}{N}((\sum Y^2) - \frac{(\sum Y)^2}{N} \tag{3.5}$$

$$\widehat{\sigma_{X,Y}} = \frac{1}{N}\left(\sum XY - \frac{\sum X \sum Y}{N}\right) \tag{3.6}$$

Ortalama vektör ve kovaryans matrisini tahmin için gerekli tüm bilgiyi içermektedir. Bu yeterli istatistikler E basamağında önemli bir yol oynar. E basamağının amacı kayıp değerleri atamaktır. Böylece M basamağı parametre tahminlerini türetmek için yukarıdaki eşitlikler kullanabilir. Yani E basamağı her bir vakanın yeterli istatistiğe ulaşmasını sağlar (Dempster ve ark., 1977). E basamağı gözlemlenen değişkenlerden kayıp verileri tahmin eden regresyon eşitlikleri veri setini kurmak için ortalama vektörü ve kovaryans matrisindeki parametreleri kullanır.

$$\hat{\beta}_1 = \frac{\hat{\sigma}_{X,Y}}{\hat{\sigma}^2{}_X} \tag{3.7}$$

$$\hat{\beta}_0 = \hat{\mu}_Y - \hat{\beta}_1\hat{\mu}_X \tag{3.8}$$

$$\hat{\sigma}^2_{Y/X} = \hat{\sigma}^2_Y - \hat{\beta}^2_1\hat{\sigma}^2_X \tag{3.9}$$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \tag{3.10}$$

Kayıp değeri olan iki değişkenli bir veri setinde, gerekli eşitlikler kullanılarak $\widehat{\beta_0}$ ve $\widehat{\beta_1}$ sırasıyla sabit ve regresyon katsayılarının tahminleri elde edilir. $\hat{Y}_i$ verilen bir X değeri için tahmin edilen değerdir.

3.2.2.4. Çoklu Değer Atama(Multiple İmputation)

Rubin (1987) çoklu değer atama yöntemini Bayes temellerine dayalı olarak veri kümesindeki kayıp değerlere olasılık dağılımının karşılık gelen mümkün değerlerin atanması olarak tanımlamıştır.

Markov Zinciri Monte Carlo yöntemi ile kayıp değerlere değer atanır ve tamamlanmış veri setleri oluşturulur. Böylece elde edilen tamamlanmış veri setlerine istatistiksel yöntemler uygulanarak tahminler elde edilir. Tek bir çıkarım elde etmek için her bir veri setinden elde edilen tahminler birleştirilir. Bağımlı değişken sayısı tekse doğrusal regresyon parametrelerinin sonsal dağılımları hesaplanarak kayıp değerler tamamlanır. Bu işlem m kez tekrarlanır. Çok değişkenli durumda kayıp değerli bağımsız değişken için koşullu dağılımdan Gibbs örnekleyici ile çekim yapılır.

Çoklu değer atama yöntemi genel olarak üç aşamada gerçekleşir. Birinci aşama değer atama aşamasıdır. Bu aşamada uygun bir değer atama modeli oluşturularak kayıp değer için m farklı tahmin yapılır. İkinci aşama tamamlanmış her bir veri setinin analiz edilmesidir. Son aşama ise birleştirme aşamasıdır. Bu aşamada m farklı tamamlanmış veri setinden elde edilen sonuçlar birleştirilir.

Değer atama aşaması I ve P adımından oluşmaktadır.

I-Adımı

I adımı gözlenen değişkenlerden kayıp değişkenleri tahmin eden regresyon eşitlikleri veri setini oluşturmak için ortalama vektör ve kovaryans matrisi tahminini kullanılır. Eşitlik 2.48 de verilen regresyon denklemi ile atanan değerler tahmin edilir.

$$Y_i^* = \hat{\beta}_0 + \hat{\beta}_1 x_i + z_i \qquad (3.11)$$

Burada Y_i^* , i.nci gözlemin kayıp değerli değişkeni için atanan değeri; $\hat{\beta}_0$,$\hat{\beta}_1$regresyon katsayılarını ,z_i normal dağılımlı ortalaması 0, varyansı regresyonun artık varyansına

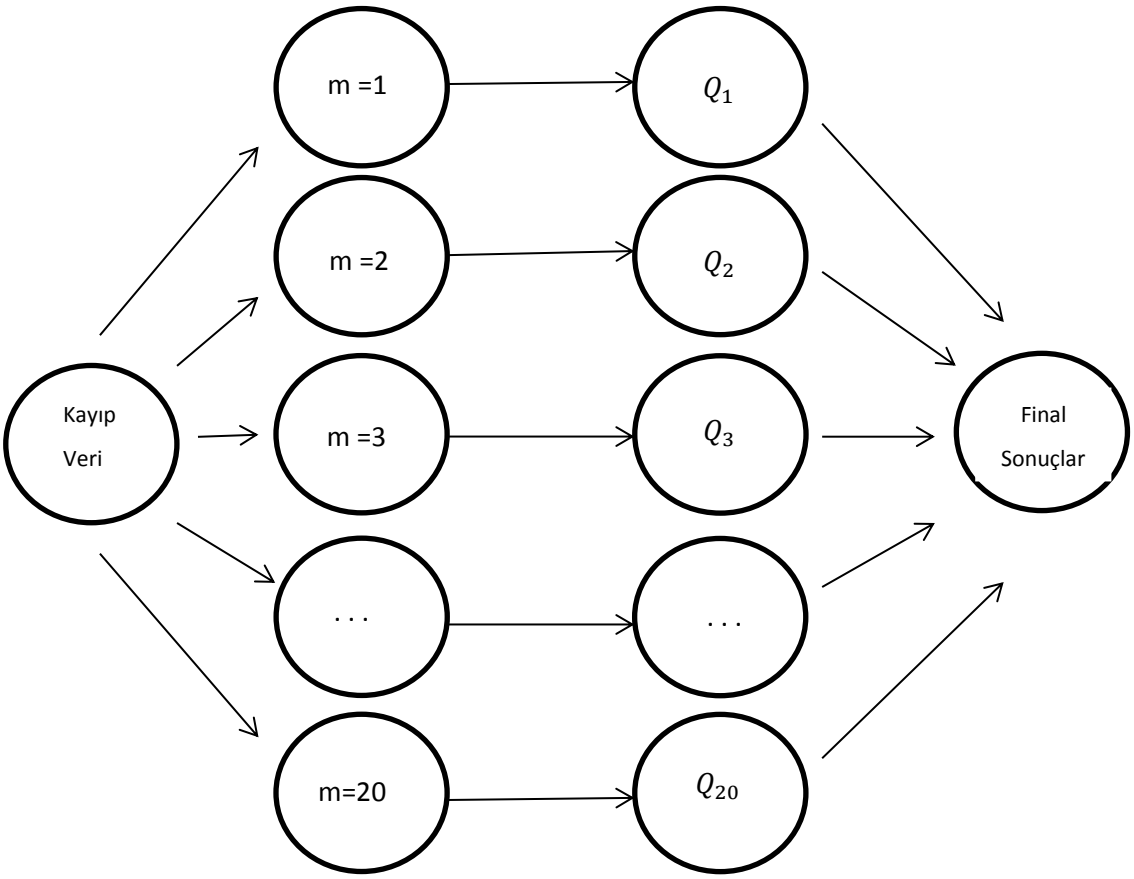
eşit olan rasgele artıklardır. Bu regresyon denklemi ile her bir kayıp değer için tahmin skorları ve bu skorlara zi artık teriminin eklenmesi ile atama değerleri elde edilir.

P-Adımı

P-Adımının amacı ortalama vektörü ve kovaryans matrisinin alternatif tahminlerini üretmektir. P adımı, ortalama vektörü ve kovaryans matrisinin sonsal dağılımını tanımlayan bir bayes analizidir. Tahmini vektörü ve matrisinin her bir elemanına artık terimin eklenmesiyle yeni bir parametre veri seti oluşturulur. Yani ortalama vektörü ve kovaryans matrisi onların sonsal dağılımından çekilir. Yeni parametrelerin sonsal dağılımından çekilmesinden sonra güncellenmiş tahminler I adımında kullanılarak yeni regresyon katsayılar seti ve atama değerleri oluşturulmuş olur. Bu yeni atama değerleri bir sonraki P adımına taşınarak parametre setleri oluşturulur (Enders,2010).

Çoklu değer atamanın ikinci aşaması analiz aşamasıdır. Değer atama aşamasında elde edilen atanmış değerlerden oluşan m farklı veri setine istenilen istatistiksel analiz ayrı ayrı uygulanır.

Yöntemin üçüncü aşaması olan birleştirme aşamasındaki amaç, atanmış veri setlerinden elde edilmiş sonuçların tek bir kümede birleştirilmesidir. Rubin (1987) çalışmasında, birleştirilmiş parametre tahminlerinin, analiz aşamasında bulunan sonuçların basit bir aritmetik ortalaması olduğuna yer vermiştir. Çoklu değer atamanın aşamaları Şekil 3.1.'de verilen grafiksel gösterimde olarak gösterilmiştir (Enders,2010).



Şekil: 3.1. Çoklu değer atama aşamasının grafiksel gösterimi

Değer atama aşaması, her bir veri setinin (örneğin, m = 20) kayıp değerlerin farklı tahminlerini içeren birden fazla kopyasını oluşturur. Adından da anlaşılacağı gibi, analizin aşamasının amacı girilen veri setlerini analiz etmektir. Bu adım da her bir veri seti için aynı istatistiksel prosedürleri uygulamaktadır. Analiz aşaması, m tane parametre tahmini verir. Birleştirme aşamasında ise ayrı ayrı elde edilmiş parametre tahminlerinden tek bir sonuç kümesi elde edilir.

4.BULGULAR VE TARTIŞMA

Bu çalışmada iki farklı uygulama yapılarak kayıp veri analiz yöntemlerinin her biri kayıp veri oluşumuna yani kayıp veri mekanizmasına göre değerlendirildi. Ayrıca farklı kayıp oranlı farklı örnek genişlikli veri setlerinde kayıp veri analiz yöntemlerin performansları karşılaştırıldı. İlk uygulama R programı ve SPSS paket programı kullanılarak simülasyon çalışması üzerinden yürütüldü. İkinci uygulama olarak gerçek veri seti olarak işsizlik verisi üzerinden yapıldı.

4.1. Simülasyon Çalışması

Simülasyon çalışmasında üç bağımsız değişkenli Cox regresyon modeli incelenmiştir. Bağımsız değişkenler standart normal dağılımdan üretilmiştir. Temel hazard fonksiyonu için ölçek ve şekil parametresi sırasıyla 0.002 ve 1 olan Weibull dağılımı kullanılmıştır. Weibul dağılımı yaşama, hayatta kalım, yıkım süreçleri gibi verilerin analizlerinde kolayca değiştirilebildiği için tercih edilen bir yöntemdir. Sansürlü zaman değişkeni, %20 sansür oranı için ölçek ve şekil parametresi 0.008 ve 1 olan Weibull dağılımından türetilmiştir. Olay zamanı ve sansür zamanının minimum değeri sağkalım zamanı olarak kaydedilir ve gerçek parametre olarak $\beta_1 = 1, \beta_2 = -3,\ \beta_3 = 2$ değerleri belirlenip örnek genişlikleri N= 50, 100 ve 300 alınarak farklı örnek genişliğinde sağkalım verisi türetilmiştir. Türetilen farklı örnek genişlikli herbir sağkalım veri setinin değerleri MAR, MCAR ve MNAR'a göre silinerek %5 , %10, %20 ve %30 kayıp oranlı veri setleri elde edilmiştir. Böylece kayıp veri analiz yöntemlerinin farklı kayıp veri mekanizmasında, farklı kayıp oranında ve farklı örnek genişliğinde karşılaştırılması sağlanmış ve veri setinde kayıp olmaması durumunda elde edilen parametreler tahminine en yakın tahmin veren yöntem bulunmaya çalışılmıştır.

Farklı kayıp veri oranına sahip her bir kayıp veri mekanizmasına göre oluşturulmuş veri setlerine ayrı ayrı kayıp veri analiz yöntemleri uygulanıp daha sonra Cox regresyon analizi ile parametre tahmini elde edilmiştir. N=50 için elde edilen sonuçlar Çizelge 4.1.'de özetlenmiştir. Yapılan tahminlerin gerçek değere (kayıp değer içermeyen veri seti tahmini) yakınlığını tespit etmek için her bir tahmin için mutlak hata hesaplanmıştır. Böylece daha rahat karşılaştırma olanağı sağlanmıştır.

Çizelge 4.1. N=50 için kayıp veri analiz yöntemleri kullanılarak tamamlanan veri setlerinin parametre tahminleri

N=50 %0 kayıp için elde edilen parametre tahminleri sırasıyla 1,475; -2,823; 2,165								
Kayıp oranı			MAR	MCAR	MNAR	Mutlak Hata MAR	Mutlak Hata MCAR	Mutlak Hata MNAR
%5	CCA	X1	1.196	1.26	1.272	0.279	**0.215**	0.203
		X2	-2.589	-2.851	-2.864	0.234	**0.028**	0.041
		X3	2.171	2.16	2.172	0.006	**0.005**	0.007
	EM	X1	-0.075	-0.103	-0.422	**1.550**	1.578	1.897
		X2	-2.274	-2.299	-1.945	**0.549**	0.524	0.878
		X3	1.702	1.645	1.442	**0.463**	0.520	0.723
	REG	X1	1.042	1.575	-0.338	0.433	**0.100**	1.813
		X2	-2.16	-2.746	-2.033	0.663	**0.077**	0.790
		X3	1.603	2.152	1.477	0.562	**0.013**	0.688
	MEAN	X1	0.412	0.488	0.333	1.063	**0.987**	1.142
		X2	-2.527	-2.693	-2.671	0.296	**0.130**	0.152
		X3	1.799	1.943	1.92	0.366	**0.222**	0.245
	MI	X1	1.3524	0.3012	0.0574	**0.122**	1.174	1.417
		X2	-2.56534	-2.4182	-2.0814	**0.257**	0.405	0.741
		X3	1.9384	1.7828	1.6002	**0.226**	0.382	0.564
%10	CCA	X1	1.827	1.087	1.11	0.352	**0.388**	0.365
		X2	-2.847	-2.901	-2.631	0.024	**0.078**	0.192
		X3	2.202	2.148	2.085	0.037	**0.017**	0.080
	EM	X1	0.901	0.548	0.71	**0.574**	0.927	0.765
		X2	-2.198	-2.169	-1.822	**0.625**	0.654	1.001
		X3	1.698	1.436	1.337	**0.467**	0.729	0.828
	REG	X1	0.718	0.852	0.073	0.757	**0.623**	1.402
		X2	-1.851	-2.504	-2.116	0.972	**0.319**	0.707
		X3	1.488	1.553	1.402	0.677	**0.612**	0.763
	MEAN	X1	0.505	0.671	0.291	0.970	**0.804**	1.184
		X2	-2.465	-2.478	-2.461	0.358	**0.345**	0.362
		X3	1.438	1.647	1.616	0.727	**0.518**	0.549
	MI	X1	0.9566	0.4802	0.5508	**0.518**	0.994	0.924
		X2	-2.0886	-2.0476	-1.996	**0.734**	0.775	0.827
		X3	1.8878	1.5064	1.4878	**0.277**	0.658	0.677
%20	CCA	X1	0.856	1.796	1.023	0.619	**0.321**	0.452
		X2	-3.392	-2.861	-3.043	0.569	**0.038**	0.220
		X3	2.695	2.111	2.5	0.530	**0.054**	0.335
	EM	X1	0.434	0.394	-0.795	**1.041**	1.081	2.270
		X2	-2.003	-1.892	-1.473	**0.820**	0.931	1.350
		X3	1.441	1.23	1.145	**0.724**	0.935	1.020
	REG	X1	0.6	0.97	0.091	0.875	**0.505**	1.384

		X2	-2.139	-2.555	-0.646	0.684	**0.268**	2.177
		X3	1.863	1.863	0.854	0.302	**0.302**	1.311
	MEAN	X1	0.553	0.456	0.102	0.922	**1.019**	1.373
		X2	-2.034	-2.055	-1.978	0.789	**0.768**	0.845
		X3	1.46	1.531	1.325	0.705	**0.634**	0.840
	MI	X1	0.6504	0.644	-0.181	**0.824**	0.831	1.656
		X2	-1.857	-1.357	-0.986	**0.965**	1.465	1.836
		X3	1.266	1.1194	1.0894	**0.899**	1.045	1.075
%30	CCA	X1	1.702	0.524	1.745	0.773	**0.049**	0.270
		X2	-3.641	-2.944	-2.711	0.818	**0.121**	0.112
		X3	2.906	2.198	2.344	0.741	**0.033**	0.179
	EM	X1	-0.577	0.746	-0.868	**0.729**	1.265	2.343
		X2	-1.783	-1.467	-1.347	**1.040**	1.356	1.476
		X3	1.344	0.821	0.927	**0.821**	1.344	1.238
	REG	X1	0.350	0.818	-0.046	1.125	**0.657**	1.521
		X2	-1.063	-1.030	-0.841	1.760	**1.693**	1.982
		X3	1.144	1.011	0.929	1.021	**0.954**	1.236
	MEAN	X1	0.495	0.640	0.019	0.980	**0.835**	1.456
		X2	-1.939	-1.740	-2.046	0.884	**0.833**	0.777
		X3	1.374	0.994	1.241	0.791	**0.771**	0.924
	MI	X1	-0.1898	0.791	-0.133	**0.684**	1.124	1.341
		X2	-0.7304	-1.0402	-7282	**0.892**	0.982	1.094
		X3	0.9868	0.9434	0.8972	**1.178**	1.221	1.267

Çizelge 4.1.'de verilen sonuçlar incelendiğinde %0 kayıp değer içeren yani tam veri setinden elde edilen değerlere (gerçek değere) 50 örnek genişliğinde %5 kayıp oranı için CCA, REG, MEAN yöntemleri en yakın değeri MCAR mekanizmasında; EM ve MI yöntemleri en yakın değerleri MAR mekanizmasında elde etmişlerdir. Tablo incelendiğinde dikkat çeken bir başka unsur ise kayıp veri mekanizması MNAR olması durumunda CCA diğer yöntemlere göre daha üstün bir performans göstermiştir. Ayrıca yöntemler gerçek değere yakınlığına göre karşılaştırıldığında en yakın tahminleri CCA ve MI yöntemleri elde etmiştir. Bu sonuçlar N= 50 örnek genişliğinde bulunan tüm kayıp oranları için geçerlidir. Fakat kayıp oranı arttıkça yöntemlerin elde ettiği tahminler gerçek değerden biraz uzaklaşmıştır. Buna rağmen her bir kayıp veri analiz yöntemi, uygun kayıp veri mekanizmasında gerçek değere yakın sonuçlar elde etmiştir.

Farklı kayıp veri mekanizmaları kullanılarak N=100 örnek genişliği için oluşturulan farklı kayıp oranlı veri setleri için Cox regresyon analizi kullanılarak parametre tahminleri ve tahminlerin gerçek değere yakınlığının tespiti için mutlak hataları elde edilmiştir ve sonuçlar Çizelge 4.2.'de verilmiştir.

Çizelge 4.2. N=100 için kayıp veri analiz yöntemleri kullanılarak tamamlanan veri setlerinin parametre tahminleri

N=100 %0 kayıp için elde edilen parametre tahminleri sırasıyla 1,495; -3,157; 2,166								
Kayıp oranı			MAR	MCAR	MNAR	Mutlak Hata MAR	Mutlak Hata MCAR	Mutlak Hata MNAR
%5	CCA	X1	1.406	1.432	1.424	0.089	**0.063**	0.071
		X2	-3.094	-3.16	-3.15	0.063	**0.003**	0.007
		X3	2.074	2.142	2.104	0.092	**0.024**	0.062
	EM	X1	1.54	0.624	0.695	**0.045**	0.871	0.801
		X2	-3.223	-2.375	-2.155	**0.066**	0.782	1.002
		X3	2.55	1.553	1.52	**0.384**	0.613	0.646
	REG	X1	0.659	1.518	1.546	0.836	**0.023**	0.051
		X2	-2.049	-3.081	-2.508	1.108	**0.076**	0.649
		X3	1.481	2.102	1.746	0.685	**0.064**	0.420
	MEAN	X1	0.906	1.533	0.958	0.589	**0.038**	0.537
		X2	-2.444	-3.08	-2.767	0.713	**0.077**	0.390
		X3	1.629	2.09	1.804	0.537	**0.076**	0.362
	MI	X1	1.448	0.7732	1.1624	**0.047**	0.721	0.332
		X2	-2.945	-2.2562	-2.1846	**0.212**	0.900	0.972
		X3	1.929	1.5612	1.7842	**0.237**	0.604	0.381
%10	CCA	X1	1.349	1.488	1.473	0.146	**0.007**	0.022
		X2	-3.275	-3.069	-3.277	0.118	**0.088**	0.120
		X3	2.323	2.135	2.213	0.157	**0.031**	0.047
	EM	X1	0.673	0.629	0.571	**0.822**	0.866	0.924
		X2	-2.391	-2.201	-2.122	**0.766**	0.956	1.035
		X3	1.505	1.264	1.266	**0.661**	0.902	0.900
	REG	X1	0.972	1.059	0.448	0.523	**0.436**	1.047
		X2	-1.747	-2.32	-1.811	1.41	**0.837**	1.346
		X3	1.573	1.689	1.089	0.593	**0.477**	1.077
	MEAN	X1	0.903	0.939	0.884	0.592	**0.556**	0.611
		X2	-2.314	-2.458	-2.441	0.843	**0.699**	0.716
		X3	1.581	1.603	1.433	0.585	**0.563**	0.733
	MI	X1	1.469	0.5378	0.4842	**0.026**	0.9572	1.010
		X2	-2.6484	-1.0174	-1.2582	**0.5086**	2.1396	1.898
		X3	1.2862	0.8356	1.0218	**0.8798**	1.3304	1.144
%20	CCA	X1	1.424	1.485	1.48	0.071	**0.010**	0.015
		X2	-3.245	-3.136	-3.243	0.088	**0.021**	0.086
		X3	2.236	2.169	2.196	0.07	**0.003**	0.030
	EM	X1	0.756	0.654	0.636	**0.739**	0.841	0.859
		X2	-2.075	-2.072	-2.026	**1.082**	1.085	1.131
		X3	1.427	1.215	1.158	**0.739**	0.951	1.008
	REG	X1	0.739	0.925	0.113	0.756	**0.570**	1.382
		X2	-1.861	-1.985	0.17	1.296	**1.172**	3.327

		X3	1.271	1.284	0.165	0.895	**0.882**	2.001
	MEAN	X1	0.869	0.952	0.854	0.626	**0.543**	0.641
		X2	-2.158	-2.449	-2.178	0.999	**0.708**	0.979
		X3	1.47	1.549	1.207	0.696	**0.617**	0.959
	MI	X1	0.9508	0.644	0.6098	**0.5442**	0.851	0.885
		X2	-2.9296	-2.3578	-1.985	**0.2274**	0.799	1.172
		X3	1.4624	1.2194	1.1692	**0.7036**	0.946	0.996
%30	CCA	X1	1.613	1.443	1.557	0.118	**0.052**	0.062
		X2	-3.282	-3.152	-3.315	0.125	**0.005**	0.158
		X3	2.009	2.285	2.376	0.157	**0.119**	0.210
	EM	X1	0.769	0.63	0.558	**0.726**	0.865	0.937
		X2	-2.292	-2.097	-1.974	**0.865**	1.060	1.183
		X3	1.318	1.315	0.908	**0.848**	0.851	1.258
	REG	X1	0.579	0.928	0.891	0.916	**0.567**	0.604
		X2	-2.74	-2.88	-2.177	0.417	**0.277**	0.980
		X3	1.847	1.907	1.77	0.319	**0.259**	0.396
	MEAN	X1	0.837	0.857	0.749	0.658	**0.638**	0.746
		X2	-1.786	-2.071	-2.026	1.371	**1.086**	1.131
		X3	1.187	1.248	0.885	0.979	**0.918**	1.281
	MI	X1	1.446	0.915	0.9424	**0.049**	0.580	0.552
		X2	-2.8138	-2.3472	-2.1922	**0.3432**	0.809	0.964
		X3	1.922	1.5812	1.5298	**0.244**	0.584	0.636

Çizelge 4.2.'de verilen bilgilere göre N=50 örnek genişliğinde verilen kayıp veri mekanizması ile kayıp veri yöntemleri arasındaki ilişki N=100 olması durumunda değişmemiştir. Fakat elde edilen mutlak hatalar aynı kayıp oranları için N=50'ye göre genel olarak bir miktar küçülmüştür. Yani gerçek değere daha yakın tahminlerde bulunulmuştur. Kayıp veri mekanizması MAR olması durumunda en iyi tahminler, MI yöntemi sonrası uygulanan Cox regresyon analizinden elde edilmiştir. Kayıp veri mekanizması MCAR olması durumunda tüm kayıp oranları için CCA yöntemi gerçek değere yakın tahminler de bulunmuştur.

Farklı kayıp veri mekanizmaları kullanılarak N=300 örnek genişliği için oluşturulan farklı kayıp oranlı veri setleri için Cox regresyon analizi kullanılarak parametre tahminleri ve tahminlerin gerçek değere yakınlığının tespiti için mutlak hataları elde edilmiştir ve sonuçlar Çizelge 4.3.'de verilmiştir.

Çizelge 4.3. N=300 için kayıp veri analiz yöntemleri kullanılarak tamamlanan veri setlerinin parametre tahminleri

N=300 %0 kayıp için elde edilen parametre tahminleri sırasıyla 1,082; -3,451; 2,171								
Kayıp oranı			MAR	MCAR	MNAR	Mutlak Hata MAR	Mutlak Hata MCAR	Mutlak Hata MNAR
%5	CCA	X1	1.172	1.167	1.078	0.090	**0.085**	0.004
		X2	-3.556	-3.449	-3.467	0.105	**0.002**	0.016
		X3	2.25	2.175	2.176	0.079	**0.004**	0.005
	EM	X1	1.136	0.95	0.911	**0.054**	0.132	0.171
		X2	-3.481	-3.067	-3.065	**0.030**	0.384	0.386
		X3	2.205	1.967	1.666	**0.034**	0.204	0.505
	REG	X1	0.458	0.592	0.713	0.624	**0.490**	0.369
		X2	-2.839	-2.905	-2.578	0.612	**0.546**	0.873
		X3	1.562	1.661	1.384	0.609	**0.510**	0.787
	MEAN	X1	1.124	1.045	0.942	0.042	**0.037**	0.140
		X2	-3.373	-3.386	-3.157	0.078	**0.065**	0.294
		X3	2.105	2.109	1.752	0.066	**0.062**	0.419
	MI	X1	1.165	0.622	0.725	**0.083**	0.460	0.356
		X2	-3.2802	-2.477	-2.367	**0.170**	0.973	1.084
		X3	2.251	1.708	1.375	**0.080**	0.462	0.795
%10	CCA	X1	0.89	0.988	0.879	0.192	**0.094**	0.203
		X2	-3.743	-3.549	-3.031	0.292	**0.098**	0.420
		X3	2.246	2.194	2.01	0.075	**0.023**	0.161
	EM	X1	0.906	0.839	0.751	**0.176**	0.243	0.331
		X2	-3.018	-3.004	-2.845	**0.433**	0.447	0.606
		X3	1.94	1.939	1.397	**0.231**	0.232	0.774
	REG	X1	0.212	0.639	0.305	0.870	**0.443**	0.777
		X2	-1.716	-2.566	-1.525	1.735	**0.885**	1.926
		X3	1.212	1.66	0.808	0.959	**0.511**	1.363
	MEAN	X1	0.902	1.001	0.773	0.180	**0.081**	0.309
		X2	-3.039	-3.049	-2.932	0.412	**0.402**	0.519
		X3	1.863	1.936	1.477	0.308	**0.235**	0.694
	MI	X1	1.225	0.74	0.534	**0.143**	0.342	0.548
		X2	-2.982	-2.842	-2.746	**0.469**	0.608	0.705
		X3	1.9042	1.7526	1.693	**0.266**	0.418	0.477
%20	CCA	X1	0.925	1.288	1.375	0.157	**0.206**	0.293
		X2	-3.15	-3.669	-3.701	0.301	**0.218**	0.250
		X3	2.464	2.366	2.372	0.293	**0.195**	0.201
	EM	X1	0.788	0.738	0.761	**0.294**	0.344	0.321
		X2	-2.895	-2.763	-2.751	**0.556**	0.688	0.700
		X3	1.851	1.729	1.102	**0.320**	0.442	1.069
	REG	X1	0.079	0.482	0.429	1.003	**0.600**	0.653
		X2	-1.874	-2.419	-1.332	1.577	**1.032**	2.119
		X3	1.07	1.336	0.776	1.101	**0.835**	1.395
	MEAN	X1	0.739	0.749	0.725	0.343	**0.333**	0.357
		X2	-2.653	-2.967	-2.847	0.798	**0.484**	0.604
		X3	1.516	1.743	1.218	0.655	**0.428**	0.953

	MI	X1	0.884	0.3594	0.411	**0.198**	0.722	0.670
		X2	-2.9194	-2.402	-2.312	**0.531**	1.049	1.138
		X3	1.9596	1.6814	1.579	**0.211**	0.489	0.591
%30	CCA	X1	0.652	1.321	1.574	0.43	**0.239**	0.492
		X2	-3.794	-3.755	-3.827	0.343	**0.304**	0.376
		X3	1.835	2.385	2.402	0.336	**0.214**	0.231
	EM	X1	0.744	0.559	0.74	**0.338**	0.523	0.342
		X2	-2.705	-2.364	-2.661	**0.746**	1.087	0.790
		X3	1.808	1.517	0.945	**0.363**	0.654	1.226
	REG	X1	0.087	0.371	0.316	0.995	**0.711**	0.766
		X2	-1.079	-1.292	-0.894	2.372	**2.159**	2.557
		X3	1.038	1.393	0.754	1.133	**0.778**	1.417
	MEAN	X1	0.738	0.873	0.641	0.344	**0.209**	0.441
		X2	-2.493	-2.867	-2.746	0.958	**0.584**	0.705
		X3	1.3	1.529	1.073	0.871	**0.642**	1.098
	MI	X1	0.8538	0.3494	0.447	**0.228**	0.732	0.635
		X2	-2.8654	-2.066	-2.151	**0.585**	1.385	1.300
		X3	1.9264	1.5844	1.483	**0.244**	0.586	0.687

Çizelge 4.3.'de verilen örnek genişliği dikkate alındığında genel olarak kayıp veri analiz yöntemleri gerçek değere yakın tahminler yaparak iyi bir performans göstermiştir. Bu örnek genişliğinde MAR mekanizmasında EM ve MI yöntemleri benzer tahminlerde bulunmuş ve bu tahminler gerçek değere oldukça yakın tahminler olarak tespit edilmiştir. Kayıp veri mekanizması MCAR olması durumunda ise en iyi performans tüm kayıp oranlarında CCA yönteminin kullanılması sonucu elde edilmiştir.

Sonuç olarak liste bazında silme yöntemi (CCA), kayıp oranlarının artması durumunda gerçek değerden biraz uzakta tahminler yapsa da özellikle büyük örnek genişlikli veri setlerinde daha iyi bir performans göstermiştir. CCA yöntemi ayrıca kayıp veri mekanizması MNAR olması durumunda da diğer yöntemlere göre iyi bir performans göstermiştir.

EM algoritması sonucu uygulanan Cox regresyon analizi tahminleri, kayıp oranının artışından etkilenerek gerçek değerden farklı bulunmuştur. Fakat yöntem büyük örnek genişliğinde tüm kayıp oralarında daha iyi performans göstermiştir.

Regresyon değer atama yöntemi sonucu uygulanan Cox regresyon tahminleri, kayıp oranı arttıkça gerçek değerden uzaklaşmış fakat örnek genişliğinin artması bu yöntemde performansı etkilememiştir.

Ortalama değer atama (MEAN) yöntemi sonrası uygulanan Cox regresyon analizi tahminleri, büyük örnek genişliği ve küçük kayıp oranlarda gerçeğe yakın bulunurken örnek genişliği küçüldükçe yöntemin performansı da düşmüştür.
Çoklu değer atama (MI) sonrası uygulanan Cox regresyon analizi tahminleri, küçük örneklemlerde kayıp oranı arttıkça gerçek değerden uzak bulunurken örnek genişliği büyüdükçe kayıp oranının artışından etkilenmemiştir.

4.2. Gerçek Veri Seti Olarak İşsizlik Verisi

Yapılan simülasyon çalışması ile kayıp veri analiz yöntemleri farklı senaryolarda incelenmiş ve yöntemlerin benzerlikleri ve farklılıkları açıkça ortaya konulmuştur. Simülasyon çalışmasında elde edilen bulguları desteklemek amacıyla gerçek veri seti olarak R programı içerisinde yer alan McCall (1996) tarafından elde edilen işsizlik veri seti kullanılmış ve örnek genişliği 200 olarak belirlenmiştir. Veri setinde, sağkalım süresi olarak işsizlik süresi ve kişinin tam zamanlı işe başlaması da olay olarak kabul edilmiş ve durum değişkeni olarak ele alınmıştır. İşsizlik süresini açıklamak için kullanılan değişkenler, yaş, işsizlik sigortası talebi, en son çalıştığı iş yerindeki haftalık kazancı ve görev süresidir. Çizelge4.4.'de işsizlik verisinde bulunan değişkenler tanıtılmıştır.

Çizelge 4.4. İşsizlik verisi değişkenleri

Değişken ismi	**Tanımı**
İşsizlik süresi	İki haftalık periyot sayısı
Durum	Tam zamanlı işe girerse 1 diğer durumda 0
Yaş	20-61 arası değişen yaşlar
İşsizlik Sigortası	Talep varsa 1 diğer durumda 0
Kazanç	2.708- 7.604 birim arası değişen haftalık kazanç
Görev süresi	En son çalıştığı iş yerindeki görev yılı

Veri setinde bulunan kişilerin %30,5'i incelenen süreç içerisinde tam zamanlı işe başlamış ve %69,5'i hala işsiz konumundadır yani sansürlüdür. Kişilerin işsizlik süresini etkileyen faktörleri belirleyebilmek için veri setine Cox regresyon analizi uygulanmış ve sonuçlar Çizelge 4.5.'de verilmiştir.

Çizelge 4.5. Cox regresyon analizi

Parameter	B	Standart Hata	Wald Testi	Sd	p değeri	Exp(B)
Yaş	-.013	.015	.689	1	.406	.987
Sigorta	-1.142	.282	16.364	1	.000	.319
Kazanç	.277	.257	1.161	1	.281	1.319
Görev Süre	.020	.029	.487	1	.485	1.020

Çizelge 4.5.'de verilen sonuçlara göre işsizlik süresi üzerinde yaş, önceki işindeki haftalık kazancı ve önceki görev süresinin önemli etkilerinin olmadığı %95 güvenle bulunmuştur. Buna karşın sigorta taleplerinin olmasının işsizlik süresi üzerinde önemli etkisinin olduğu tespit edilmiştir ($p<0,05$). Buna göre işsizlik sigortası talep etmeyenlere göre talep edenlerin işsiz kalma süresi 3,135 (1/exp(-1,142)) kat artmaktadır. Çalışmada amaç gerçek parametre değerine yakın tahminler yapan kayıp veri analiz yöntemlerinin belirlenmesidir. Bu amaçla Çizelge 4.5.'de verilen sonuçlar orijinal sonuçlar olarak değerlendirilerek referans olacaktır. Kayıp veri analiz yöntemleri incelenebilmesi için veri setinde %20 kayıp olacak şekilde bazı değerler sırasıyla MAR, MCAR ve NMAR varsayımına göre silinmiş ve her bir kayıp değerli veri setine ayrı ayrı kayıp veri analiz yöntemleri uygulanarak eksiksiz veri setleri oluşturulmuş ve Cox regresyon analizi ile parametre tahmini yapılmıştır. Elde edilen regresyon katsayısı tahminlerinin kayıp değer içermeyen veri setinden elde edilen regresyon katsayısına yakınlığını göstermek amacıyla Çizelge 4.6. oluşturulmuştur.

Çizelge 4.6. Regresyon Katsayısı Tahminleri

MAR						
Parametreler	Orijinal	CCA	MEAN	EM	REG	MI
Yaş	-0.013	-0.025	-0.012	**-0.013**	-0.012	-0.012
Sigorta	-1.142	**-1.333**	-1.091	-1.103	-1.103	-1.104
Kazanç	0.277	0.343	0.390	**0.297**	0.333	**0.286**
Görev Süre	0.020	0.025	0.022	0.018	0.018	**0.020**
MCAR						
Parametreler	Orijinal	CCA	MEAN	EM	REG	MI
Yaş	-0.013	-0.020	-0.012	-0.013	**-0.013**	-0.012
Sigorta	-1.142	-1.121	-1.119	-1.153	**-1.149**	-1.126
Kazanç	0.277	0.393	**0.272**	0.371	0.335	0.245
Görev Süre	0.020	0.002	**0.020**	0.017	0.019	0.019
MNAR						
Parametreler	Orijinal	CCA	MEAN	EM	REG	MI
Yaş	-0.013	-0.006	-0.011	-0.011	-0.011	**-0.012**
Sigorta	-1.142	**-1.273**	-1.075	-1.075	-1.075	-1.075
Kazanç	0.277	0.025	0.017	0.046	0.015	**0.060**
Görev Süre	0.020	**0.017**	0.025	0.024	0.025	0.024

Çizelge 4.6.'de verilen sonuçlara göre her bir kayıp veri mekanizması için ayrı ayrı yapılan analizlere göre bütün yöntemlerde sigorta değişkenin işsizlik süresi üzerinde önemli etkisi olduğu %95 güvenle tespit edilmiştir. Ayrıca orijinal, yani kayıp değer içermeyen veri setinden elde edilen Cox regresyon katsayısına genellikle yakın tahminler elde edilmiştir. Bununla beraber tablo içinde koyu olarak belirtilen tahminler orijinal regresyon katsayısına en yakın tahminlerdir. Gerçek veri seti uygulamasında elde edilen sonuçlar simülasyon çalışmasından elde edilenlere tutarlıdır. Buna göre MAR mekanizmasında en yakın tahminleri MI ve EM yöntemleri; MCAR mekanizmasında REG ve MEAN yöntemleri; MNAR mekanizmasında CCA ve MI yöntemleri elde etmiştir.

5. SONUÇ VE ÖNERİLER

İstenilen olayın gerçekleşmesine kadar geçen zamanın modellenmesini sağlayan sağkalım analizi, sansürlü verileri de dikkate alarak işlem yaptığı için pek çok kullanım alanı mevcuttur. Sağkalım analizinde en yaygın kullanılan yöntem ise, sağkalım süresini etkileyen faktörleri belirleyen Cox regresyon analizidir. Çoğu araştırmada karşılaşılan kayıp değer, sağkalım analizi çalışmalarında da ortaya çıkmaktadır. Bu durum araştırmacılar için önemli bir problemdir. Çünkü istatistiksel analizler ve istatistiksel bilgisayar programları, veride kayıp değer olmaması durumu için tasarlanmıştır. Bu nedenle pek çok araştırmacı karşılaştıkları kayıp değer problemini o gözlemi silerek gidermeye çalışmaktadır. Yapılan bu uygulama veri setinin azalmasına ve istatistiksel gücün küçülmesine neden olmaktadır. Bu nedenle çok sık karşılaşılan kayıp veri probleminin giderilmesi önem kazanmaktadır.

Bilim insanlarının yaptığı araştırmalar neticesinde bu problemi çözen pek çok yöntem geliştirilmiştir. Bu çalışmada en yaygın kullanılan Ortalama değer atama, Regresyon değer atama, Beklenti Maksimizasyonu, Çoklu Değer Atama ve Liste Bazında silme yöntemleri ayrıntılı olarak incelenmiştir. Kayıp veri analiz yöntemlerinin performansları, farklı kayıp veri mekanizmalarında (MAR, MCAR, MNAR), örnek genişliklerinde (N=50, 100, 300) ve kayıp oranlarında (%5, %10, %20, %30) karşılaştırılmıştır. Buna göre liste bazında silme yöntemi (CCA), kayıp oranlarının artması durumunda gerçek değerden biraz uzakta tahminler yapsa da özellikle büyük örnek genişlikli veri setlerinde daha iyi bir performans göstermiştir. CCA yöntemi ayrıca kayıp veri mekanizması MNAR olması durumunda da diğer yöntemlere göre iyi bir performans göstermiştir. EM algoritması sonucu uygulanan Cox regresyon analizi tahminleri, kayıp oranının artışından etkilenerek gerçek değerden farklı bulunmuştur. Fakat yöntem büyük örnek genişliğinde tüm kayıp oralarında daha iyi performans göstermiştir. Regresyon değer atama yöntemi sonucu uygulanan Cox regresyon tahminleri, kayıp oranı arttıkça gerçek değerden uzaklaşmış fakat örnek genişliğinin artması bu yöntemde performansı etkilememiştir. Ortalama değer atama (MEAN) yöntemi sonrası uygulanan Cox regresyon analizi tahminleri, büyük örnek genişliği ve küçük kayıp oranlarda gerçeğe yakın bulunurken örnek genişliği küçüldükçe yöntemin performansı da düşmüştür. Çoklu değer atama (MI) sonrası uygulanan Cox regresyon

analizi tahminleri, küçük örneklemlerde kayıp oranı arttıkça gerçek değerden uzak bulunurken örnek genişliği büyüdükçe kayıp oranının artışından etkilenmemiştir.

Ayrıca genellikle klinik çalışmalarda yaygın olarak kullanılan Cox regresyon analizi, ekonomi alanında uygulanmış ve gerçek veri seti olarak işsizlik verisi ele alınmıştır. Veri setinde, sağkalım süresi olarak işsizlik süresi ve kişinin tam zamanlı işe başlaması da olay olarak kabul edilmiş ve durum değişkeni olarak ele alınmıştır. Veri setinde bulunan kişilerin %30,5'i incelenen süreç içerisinde tam zamanlı işe başlamış ve %69,5'i hala işsiz konumundadır yani sansürlüdür. Kişilerin işsizlik süresini etkileyen faktörleri belirleyebilmek için veri setine Cox regresyon analizi uygulanmış veişsizlik süresi üzerinde yaş, önceki işindeki haftalık kazancı ve önceki görev süresinin önemli etkilerinin olmadığı %95 güvenle bulunmuştur. Buna karşın sigorta taleplerinin olmasının işsizlik süresi üzerinde önemli etkisinin olduğu tespit edilmiştir ($p<0,05$). Kayıp veri analiz yöntemleri incelenebilmesi için veri setinde %20 kayıp olacak şekilde bazı değerler sırasıyla MAR, MCAR ve MNAR varsayımına göre silinmiş ve her bir kayıp değerli veri setine ayrı ayrı kayıp veri analiz yöntemleri uygulanarak eksiksiz veri setleri oluşturulmuş ve Cox regresyon analizi ile parametre tahmini yapılmıştır. bütün yöntemlerde sigorta değişkenin işsizlik süresi üzerinde önemli etkisi olduğu %95 güvenle tespit edilmiştir. Ayrıca simülasyon çalışması ile tutarlı bir şekilde MAR mekanizmasında en yakın tahminleri MI ve EM yöntemleri; MCAR mekanizmasında REG ve MEAN yöntemleri; MNAR mekanizmasında CCA ve MI yöntemleri elde etmiştir. Bu çalışma ile araştırmacılara kayıp veri olma olasılığı ile ölçülen değişkenler arasındaki ilişkiyi tanımlayan kayıp veri mekanizması belirlenebiliyor ise o oluşuma uygun yöntemin seçilmesinin daha doğru sonuç elde edeceği bulunmuştur.

KAYNAKÇA

1. Alkan, N., Terzi, Y., Cengiz, M. A., & Alkan, B. B. (2013). Comparison of Missing Data Analysis Methods in Cox Proportional Hazard Models. Turkiye Klinikleri Journal of Biostatistics, 5(2).
2. Allison, T., Puce, A., & McCarthy, G. (2000). Social perception from visual cues: role of the STS region. Trends in cognitive sciences, 4(7), 267-278.
3. Ata, N., Karasoy, D., & Sözer, M. T. (2008). Orantısız hazardlar için tabakalandırılmış Cox regresyon modeli ve meme kanseri hastaları üzerine bir uygulama. Turkiye Klinikleri Journal of Medical Sciences, 28(3), 327-332.
4. Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. Journal of school psychology, 48(1), 5-37.
5. Beale, E. M. L., & Little, R. J. A. (1975). Missing values in multivariate analysis. Journal of the Royal Statistical Society, Series B, 37, 129–145
6. Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics*, 89-99.
7. Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. Journal of the Royal Statistical Society: Series B (Methodological), 22(2), 302-306.
8. Chen, H. Y., & Little, R. J. (1999). Proportional hazards regression with missing covariates. *Journal of the American Statistical Association*, *94*(447), 896-908.
9. Cohen, S. (1988). Perceived stress in a probability sample of the United States.
10. Cox, D.R, Regression Models and Life-Tables, Journal of the Royal Statistical Society, 1972 Series B, Vol.34,No: 2, 187-220.
11. Custer, M., Sharairi, S., & Swift, D. (2012). A Comparison of Scoring Options for Omitted and Not-Reached Items through the Recovery of IRT Parameters When Utilizing the Rasch Model and Joint Maximum Likelihood Estimation. Online Submission.
12. Demir, E., & Parlak, B. (2012). Türkiye'de eğitim araştırmalarında kayıp veri sorunu. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 3(1), 230-241.
13. Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39(1), 1-22.

14. Diggle, P. J. (1989). Testing for random dropouts in repeated measurement data. *Biometrics*, 1255-1258.
15. Dixon, J. W., & Ooi, B. T. (1988). Indirect current control of a unity power factor sinusoidal current boost type three-phase rectifier. *IEEE Transactions on industrial electronics*, *35*(4), 508-515.
16. Efron, B., & Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, *70*(350), 311-319
17. Enders, C. K. (2003). Using the expectation maximization algorithm to estimate coeffi cient alpha for scales with item-level missing data. Psychological Methods, 8, 322–337
18. Enders, C. K. (2010). Applied missing data analysis. Guilford press.
19. Finch, H., & Margraf, M. (2008). Imputation of categorical missing data: A Comparison of multivariate normal and multinomial methods. Retrieved November, 12, 2009.
20. Finkbeiner, C. (1979). Estimation for the multiple factor model when data are missing. Psychometrika, 44, 409–420
21. Garson, N. G. (2015). The Swaziland Question and a road to the sea 1887-1895 (Doctoral dissertation).
22. Graham, J. W., Hofer, S. M., & MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research*, *31*(2), 197-218.
23. Halley, E. (1693) An estimate of the degrees of mortality of mankind, drawn from the curious tables of births and funerals in the City of Breslaw; with an attempt to ascertain the price of annuities upon lives. Phil. Trans., 17, 596–610
24. Hartley, H. O., & Hocking, R. R. (1971). The analysis of incomplete data. Biometrics, 27, 783–808.
25. İnceoğlu, F. (2013). Sağkalım analiz yöntemleri ve karaciğer nakli verileri ile bir uygulama (Master's thesis, İnönü Üniversitesi).
26. Jamshidian, M., & Bentler, P. M. (1999). ML estimation of mean and covariance structures with missing data using complete data routines. Journal of Educational and behavioral Statistics, 24(1), 21-24.
27. , J. D., & Prentice, R. L., (1980). The Statistical Analysis of Failure time data. Wiley,219 s, New York.

28. Karasoy, D. (2008). Cox Regresyon Modeli ve Akciğer Kanseri Verileri İle Bir Uygulama. İstatistikçiler Dergisi: İstatistik ve Aktüerya, 1(1), 16-23.
29. Kim, K. H., & Bentler, P. M. (2002). Tests of homogeneity of means and covariance matrices for multivariate incomplete data. *Psychometrika*, *67*(4), 609-623.
30. Klein, J. P., & Moeschberger, M. L. (2006). Survival analysis: techniques for censored and truncated data. Springer Science & Businss Media.
31. Kunst, F., Ogasawara, N., Moszer, I., Albertini, A. M., Alloni, G. O., Azevedo, V., ... & Borriss, R. (1997). The complete genome sequence of the gram-positive bacterium Bacillus subtilis. *Nature*, *390*(6657), 249.
32. Lawless, J. F. (2011). Statistical models and methods for lifetime data (Vol. 362). John Wiley & Sons.
33. Lee, E. T., & Wang, J. (2003). Statistical methods for survival data analysis (Vol. 476). John Wiley & Sons.
34. Leung, K-M., Elashoff, R.M. and Afifi, A.A., 1997. Censoring issues in survival analysis. Annu. Rev. Public Health, 18, 83-104.
35. Liang, J., & Bentler, P. M. (2004). An EM algorithm for fitting two-level structural equation models. *Psychometrika*, *69*(1), 101-122.
36. Little Roderick, J. A., & Rubin Donald, B. (1987). Statistical analysis with missing data. Hoboken, NJ: Wiley.
37. Little, R. J., & Rubin, D. B. (2002). Statistical analysis with missing data. Wiley. New York.
38. Little, R. J. (1992). Regression with missing X's: a review. *Journal of the American Statistical Association*, *87*(420), 1227-1237.
39. Little, W. C., Constantinescu, M., Applegate, R. J., Kutcher, M. A., Burrows, M. T., Kahl, F. R., & Santamore, W. P. (1988). Can coronary angiography predict the site of a subsequent myocardial infarction in patients with mild-to-moderate coronary artery disease?. Circulation, 78(5), 1157-1166.
40. Marubini, E. Ve Valsecchi, M., "Analysing Survival Data from Clinical Trials and Observational Studies", John Wiley and Sons ,USA,1-250, (2004).
41. McLachlan, G. J., & Krishnan, T. (1997). Wiley series in probability and statistics. The EM algorithm and extensions.
42. McCal, B.P .(1996) Unemployment insurance rules, joblessness, and part-time work Econometrica: Journal of the Econometric Society, 647-682.

43. Muthén, B., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics, 55*(2), 463-469.
44. Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, *52*(3), 431-462.
45. Nelson, H. E. (1982). National Adult Reading Test (NART): For the assessment of premorbid intelligence in patients with dementia: Test manual. Nfer-Nelson.
46. Orchard, T., & Woodbury, M. A. (1972). A missing information principle: Theory and applications. Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability, 1, 697–715.
47. Özdamar, K. (2001). SPSS İle Biyoistatistik, Kaan Kitapevi. Baskı. Eskişehir, 315-368.
48. Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. Review of educational research, 74(4), 525-556.
49. Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical linear models:Applications and data analysis methods (Vol. 1). Sage.
50. Rubin, D. B. (1987). Statistical analysis with missing data. Wiley.
51. Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581-592.
52. Sinharay, S., Stern, H. S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. Psychological methods, 6(4), 317.
53. Terzi, Y., Cengiz, M. A., & Bek, Y. (2005). Cox Regresyon Modelinde Oransal Hazard Varsayımının Artıklarla İncelenmesi ve Akciğer Kanseri Hastaları Üzerinde Uygulanması. Türkiye Klinikleri Journal of Medical Sciences, 25(6), 770-775.
54. Terzi, Y., (2003). Sansürlü Veriler İçin Sağkalım Analizi ve Gerçek Verilere Uygulanması. Doktora Tezi, Ondokuz Mayıs Üniversitesi Fen Bilimleri Enstitüsü, Samsun, 167.
55. Thoemmes, F., & Enders, C. K. (2007, April). A structural equation model for testing whether data are missing completely at random. In annual meeting of the American Educational Research Association, Chicago, IL.
56. Yuan, Y. C. (2000). Multiple imputation for missing data: Concepts and new development. In Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference (Vol. 267).

57. Wilkinson, L. (1999). Task Force on Statistical Inference. Statistical methods in psychology journals. American Psychologist, 54(3), 594-604.
58. Wilks, S. S. (1932). Moments and distributions of estimates of population parameters from fragmentary samples. *The Annals of Mathematical Statistics*, *3*(3), 163-195.

ÖZGEÇMİŞ

KişiselBilgiler

AdSoyad	Özden İŞÇİMEN
DoğumTarihi	11.05.1992
DoğumYeri	BAFRA
E-posta Adresi	iscimen.ozdeny@gmail.com

1. EğitimBilgileri

Lisans	Sinop Üniversitesi Fen Edebiyat Fakültesi İstatistik Bölümü

2. İş Deneyimi

Haziran 2016- Ekim 2018-	Lcw satış danışmanlığı Samsun Büyükşehir Belediyesi Anakent Ltd.Şti