KAYIP VERİ DURUMUNDA

SAĞLAM KESTİRİM

ROBUST ESTIMATION IN CASE OF

MISSING DATA

ONUR TOKA

Hacettepe Üniversitesi

Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin

İSTATİSTİK Anabilim Dalı İçin Öngördüğü

YÜKSEK LİSANS TEZİ

olarak hazırlanmıştır.

2012

Fen Bilimleri Enstitüsü Müdürlüğü'ne,

Bu çalışma jürimiz tarafından **İSTATİSTİK ANABİLİM DALI'nda YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Başkan :……

Prof. Dr. Öniz Toktamış

Üye (Danışman) :……

Doç. Dr. Meral ÇETİN

Üye :……

Prof. Dr. Olcay ARSLAN

ONAY

Bu tez Hacettepe Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliği'nin ilgili maddeleri uyarınca yukarıdaki jüri üyeleri tarafından …../...../…….. tarihinde uygun görülmüş ve Enstitü Yönetim Kurulunca …../…../…….. tarihinde kabul edilmiştir.

Prof. Dr. Fatma SEVİN DÜZ

Fen Bilimleri Enstitüsü Müdürü

Anneme

Ve

Babama...

KAYIP VERİ DURUMUNDA SAĞLAM KESTİRİM

Onur Toka

ÖZ

Bu çalışmadaki amaç, hem aykırı değer hem de kayıp veri bulunduran veri kümesinin kayıp değerlerine ve parametre kestirimlerine aykırı değerlerden en az etkilenecek şekilde ulaşabilmektir.

Çalışmada kayıp veri ile ilgili tanımlamalar, kayıp veri yükleme yöntemleri ve kayıp verinin tarihsel gelişimi ile ilgili bilgiler verilmiştir. Sağlam istatistik ve sağlam kestirim ile ilgili bilgiler verilerek çalışmada kullanılan sağlam kestirimler açıklanmıştır. Kayıp veri durumunda sağlam kestirim için yapılan çalışmalar aktarılmıştır.

Uygulamada iki veri kümesi ele alınmıştır. İlk örnekte belli oranda bozuluma sahip veri kümelerinin rasgele olacak şekilde silinmiş kayıp değerleri klasik ve sağlam yöntemlerle elde edilmiştir. Kayıp değerleri yüklenen veri kümesinin ortalama vektörü ve kovaryans matris kestirimleri için sağlam yöntemlerle çözümleme yapılmıştır. İkinci uygulamada ise regresyon veri kümesinden rasgele şekilde silinmiş kayıp değerler için klasik ve sağlam veri yükleme (imputasyon) yöntemleri kullanılarak kayıp veri kestirimleri karşılaştırılmıştır.

Sonuç olarak, veri kümesinin dağılımında bozulumların olması, aykırı değerlerin bulunması kayıp değerlerin ve parametrelerin kestirimlerinde sorun yarattığı görülmüştür. Bu sorunu çözmek için sağlam yöntemler kullanılan kayıp veri yükleme yöntemlerin kullanılmasının daha etkin sonuçlar getirdiği gösterilmiştir.

ANAHTAR KELİMELER: Kayıp veri, kayıp veri yükleme yöntemleri, sağlam veri yükleme yöntemi, ER algoritması, sağlam kestiriciler, kayıp veri düzenekleri, kayıp veri analizi, kısmi sağlam regresyon.

Danışman: Doç. Dr. Meral ÇETİN, Hacettepe Üniversitesi, İstatistik Bölümü.

ROBUST ESTIMATION IN CASE OF MISSING DATA

Onur Toka

ABSTRACT

The aim of this study is to find missing values and parameter estimations, which both have been least-influenced by outliers when the data sets have missing values and outliers at the same time.

In this study, definitions and historical developments of missing data and missing data imputation methods have been explained. Robust approach and robust estimators have been introduced briefly and then robust estimators which are utilized in this study have been explained. The most significant papers in the literature about robust estimators, missing data imputation methods and parameter estimations in the case of missing data and outliers have been given.

For the application, two data sets have been debated. In the first example, randomly removed missing values from contaminated data sets have been obtained by employing classical and robust methods. After imputating the missing parts of the data sets, mean vectors and covariance matrices of these sets have been estimated by using robust methods. For the second example, estimations of randomly removed missing values from contaminated regression data sets, have been obtained by using classical and robust imputation methods and thus, these missing value estimations have been compared.

In conclusion, it has been shown that contaminations or the existence of the outliers in the data sets affects the reliability of the estimation of missing values and parameters. It has been shown that robust imputations methods give more consistent results than classical ones.

KEYWORDS: Missing data, missing data imputation methods, robust data imputation, ER algorithm, robust estimators, missing data mechanism, missing data analysis, partial robust M-regression.

Advisor: Associate Prof. Dr. Meral ÇETİN, Hacettepe University, Department of Statistics.

TEŞEKKÜR

Hayatın her alanında yaşanan sıkıntıların akademik hayatta belli bir sınırda kalmasını sağlayan, problemleri omuzlamaya yardımcı olan kişilerdir danışmanlar. Ben, danışman tanımlamalarından hangisi yaparsam yapayım Doç. Dr. Meral ÇETİN'i bu tanımlamaların içine sığdıramam. Kendisiyle çalışmaktan büyük bir mutluluk duyuyorum. Teşekkürler Hocam.

Bütün ömrümü geçirebileceğim Beytepe'nin merkezinde yer alan İstatistik Bölümü ise hayatımdaki en güzel günlerin şahididir. Bu güzel günleri paylaştığım bütün hocalarıma, çalışma arkadaşlarıma tek tek teşekkür ederim.

Zorlukları çekmek için onlara göğüs germeyi öğrenmek gerekir. Benim bu konuda örneğim ise babam Mahir TOKA'dır. Yaşanan güzel günler, birlikte çekilen zorlukların meyvesidir baba. Benim ilk meyvem olan bu çalışmanın yarısı senindir. Ancak itiraf etmeliyim ki babamın ve benim zorlukları aşmamızdaki en büyük koz, annem Gülüzar TOKA'dır. En büyük hayat, mutluluk ve sabır kaynağım anneciğim, bu çalışmanın diğer yarısı da senindir. İkinize de teşekkürler.

Yaş büyüdükçe zorluklar artıyor. Bu zorlukları beraberce paylaşmak için yola çıktığım, sabır taşım, hayatımın en önemli parçası, eşim; Gülnaz TOKA. Hayatıma şimdiye kadar verdiğin ve şimdiden sonra vereceğin anlam için, desteklerin için teşekkür ederim.

Dostlar nasıl birer kardeşse, kardeşim de büyük bir dosttur benim için. Olcay TOKA, kardeşim, ailemizin diğer parçası, desteklerin için teşekkür ederim. Bu küçük ailenin hep yanında olan destekçileri; teyzem Gülender KURTUL, eniştem Satılmış KURTUL ve yeğenim ÖZGÜN KURTUL… Birliktelik güzel şeydir. Bizimle birlikte olduğunuz için teşekkürler.

Çoğu insan arkasında birileri oldukça hayattan güç alır. Az sayıda bir kısım ise arkasında olan büyükleriyle değil, yanında omuz omuza duran dostlarıyla, yakınlarıyla birlikte mücadele eder hayattaki zorluklarla… Hayatımın bu anına kadar yan yana durduğum dostlara ancak bugün teşekkür edebiliyorum. Değerli dostlarım, düşündaşlarım… Bana zor günlerde bir olabilmenin ne olduğunu gösterdiğiniz için hepinize teşekkür ediyorum.

Son olarak benim için anlamlı olan bu sayfayı buraya kadar okuyan kıymetli insanlar, sizlere de teşekkür ederim.

İÇİNDEKİLER DİZİNİ

ÇİZELGELER DİZİNİ

ŞEKİLLER DİZİNİ

KISALTMALAR DİZİNİ

EKK	:En Küçük Kareler
ML	:En Çok Olabilirlik
TRK	:Tamamen Rasgele Kayıp
RK	:Rasgele Kayıp
ROK	:Rasgele Olmayan Kayıp
EM	:Beklenti En Büyükleme
ECM	:Koşullu Beklenti En Büyükleme
OVYH	:Ortalama Veri Yükleme Hatası
OMS	:Ortalama Mutlak Sapma
OMGS	:Ortalama Göreli Mutlak Sapma
LTS	:En Küçük Kareler Kestirimi
MCD	:Minimum Kovaryans Determinant
OGK	:Ortogonolleştirilmiş Gnanadesikan–Kettenring
SD	:Stahel – Donoho
BS	:Bisquare S
MVE	:En Küçük Hacim Elipsoit
ER	:Beklenti Sağlamlaştırma
FSA	:İleri Seçim Algoritması
FS	:İleri Seçim
TBS	:t-Bisquare S
PLS	:Kısmi En Küçük Kareler
LAD	:En Küçük Mutlak Sapma
LADimpute	:En Küçük Mutlak Sapma ile Veri Yükleme
PRM	:Kısmi Sağlam M-Regresyonu
EM-PRM	:EM Algoritması ile Kısmi Sağlam M-Regresyonu
EM-PLS	:EM Algoritması ile Kısmi En Küçük Kareler
TBA	:Temel Bileşenler Analizi
TB	:Temel Bileşen

RobPCA :Sağlam Temel Bileşenler Analizi

RobPLS :Sağlam Kısmi En Küçük Kareler

LLSimpute :Yerel En Küçük Kareler ile Veri Yükleme

RLSP :TB'ler ile Sağlam En Küçük Kareler

TBR :Temel Bileşenler Regresyonu

YEM : Yapısal Eşitlik Modelleri

BİRİNCİ BÖLÜM

1. GİRİŞ

Bilimsel alanda yaşanan gelişmeler yüzyıllardır süregelmektedir. İnsanoğlunun hayatta kalmak adına başlattığı çalışmalar, gün be gün gelişerek hayatı daha kolay hale getirmenin birer anahtarı olmuştur. Geçmiş zamanda meraklar sonucunda ortaya çıkan keşifler ve elde edilen bilgiler, evrimsel süreç içerisinde kendi yollarını bulmuş, akıl ve düşünce ile birleşmiştir. Bu süreç, bize Dünya'yı ve Dünya'nın imkânlarını daha yakından tanıma, daha fazla bilgi sahibi olma fırsatını vermiştir. Bilgiye sahip olma duygusu, insanoğlunun vazgeçilmez isteklerindendir ve bu nedenle bilgi günümüzün sahip olunabilecek en önemli unsurudur. Bilgiye ulaşmak için yapılan çalışmalarda araştırıcıların başlangıç noktaları, geçmişteki deneyimlerdir. Kaynaksız bir kitap olmayacağı gibi yeni bir şey için düşünsel ya da deneysel bir birikimin olmaması da mantıksızdır. Yeni bir şeyi ortaya çıkartmak için eskiden yararlanmak en geçerli yöntemlerden biridir. Geleceği ve olacağı kestirebilmek adına faydalanılan istatistik, geçmiş birikimlerle gelecek için çalışmalar yapmak amacıyla kullanılan, kendine özgü yöntemleri sayesinde hemen hemen her bilim dalının uygulamalarında karşılaşılabilen bir disiplindir, bilim dalıdır.

İstatistik, geçmişteki birikimler yoluyla toplanmış olan tüm bilgiyi çözümleyerek daha sonraki süreçte alınacak kararlara yardımcı olacak sonuçları verir. Sürecin gerçekleşme ihtimali ile ilgili bilgiler sağlar. Aynı zamanda deneysel düzenlemelerle elde edilen sonuçların gerçekliğini ve doğruluğunu araştırır. Yeni yöntemler oluşturarak bilim dallarının birçok alanında uygulanır ve uygun modeller tanımlar. İstatistiğin gerek diğer bilim dallarıyla olan ilişkisinden gerekse sonuç değerlendirmelerinden dolayı önemli olduğu görülmektedir. Dolayısıyla istatistikte yapılan her inceleme sonucunda elde edilecek bilgiye güvenebilmek için sağlıklı ve güvenilir veriyle çalışmak, veriye uygun istatistiksel çalışmalar yapmak gerekmektedir.

Elde edilmiş bir veri kümesinde geçmiş bilgilerin kaybı, yok olmaları ve bazı imkânsızlıklardan elde edilememesi büyük bir problemdir. Aynı şekilde deneysel düzenlemelerde kayıt altına alınmayan, maddi gerekçelerle tekrarlanamayan bilginin eksikliği de büyük bir problemdir. Bu noktadan itibaren istatistiksel

yöntemlerdeki bilgi kayıpları göz önüne alınarak kayıp değerleri kestirebilmek için yapılabilecek çeşitli çalışmalar üzerinde durulmuştur. İşte bu süreç sonunda kayıp veri durumunda istatistiksel yöntemlerin nasıl ele alınması gerektiğine dair sorulara cevap verilmeye başlanmıştır. Veri kümelerini tamamlamadan kayıp veriden kurtulma çalışmaları yeni yöntemlerle yerini model tabanlı veri yükleme, çoklu veri yükleme, Bayesci yöntemler yardımıyla veri yükleme gibi gelişmiş yöntemlere bırakmıştır.

Diğer taraftan, belirli bir modele uyan veride meydana gelen bozulumların kitle kestirimlerini etkilememesi amacıyla da çeşitli yöntemler geliştirilmiştir. Aykırı değerler yüzünden bozulum yaşayan bir veri kümesinin, istatistiksel kestirimleri hem kayıp veriyi hem de kestirilen parametreleri etkilemektedir. Veri yapısının aykırı değerle bozulmuş olması durumunda klasik istatistiksel yöntemlerin yerine daha etkili sonuçlar verdiği bilinen sağlam (robust) kestirim yöntemleri uygulanmaktadır. Sağlam kestirim yöntemleri ile verinin içinde bulunan aykırı değerlerin etkisi azaltılarak istatistiksel çözümlemeler yapılır.

Bu çalışmadaki amaç, hem aykırı değer hem de kayıp veri bulunduran veri kümesinin, aykırı değerlerden en az etkilenecek şekilde kayıp değerlerine ve parametre kestirimlerine ulaşabilmektir. Bu amaç doğrultusunda ikinci bölümde kayıp veri ile ilgili tanımlamalar, kayıp veri yükleme yöntemleri ve kayıp verinin tarihsel gelişimi ile ilgili detaylı bir aktarım yapılacaktır. Üçüncü bölümde sağlam istatistik ve sağlam kestirim ile ilgili bilgi verilecektir ve uygulamada kullanılan sağlam kestiriciler açıklanacaktır. Üçüncü bölümün devamında kayıp veri durumunda sağlam kestirim ile ilgili çalışmalar ve bu konu ile ilgili gelişen süreç paylaşılacaktır. Dördüncü bölümde ise, aykırı ve kayıp değerleri bulunan veri kümesinin ortalama ve kovaryans matris kestirimleri ile ilgili bir uygulama verilecektir. Ayrıca aykırı değerleri bulunan bir regresyon veri kümesinde rasgele şekilde oluşturulacak kayıp veri için klasik ve sağlam veri yükleme yöntemleri kullanılarak kayıp veri kestirimlerinin karşılaştırıldığı ikinci bir uygulama verilecektir.

İKİNCİ BÖLÜM

2. KAYIP VERİ (MISSING DATA) ve KAYIP VERİ YÖNTEMLERİ İÇİN GENEL BİLGİLER

Bu bölümde ilk olarak kayıp veri ve kayıp verinin yapısı hakkında genel bilgiler verilecektir. Daha sonra ise kayıp veri düzeneklerinden bahsedilecektir.

2.1. Veri Matrisi (Data Matrix)

İstatistiksel çözümlemelerde toplanan veriyi daha kolay ifade edebilmek için veri matrisi kullanılır. Veri matrisi, p değişkenin bulunduğu sütunlardan ve n gözlemin bulunduğu satırlardan oluşan değerler topluluğu olarak kabul edilir ve en genel haliyle Şekil 2.1'deki gibidir:

$$\boldsymbol{Y} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \dots & y_{np} \end{bmatrix}$$

Şekil 2.1. Veri Matrisi Gösterimi

Veri matrisinde satırlar; gözlemleri, durumları, sütunlar ise gözlemlere ait olan özellikleri yani değişkenleri temsil eder. Satırlardaki her gözlem, sütunlardaki özellikleri ifade eden sayısal değerle temsil edilerek veri kümesini oluşturur. Veri matrisleri genellikle alfabenin harfleri ile temsil edilirken matrisin herhangi bir gözlemi i indisiyle, herhangi bir değişkeni ise j indisiyle gösterilir. Bu gösterimler matematiksel anlatımlarda ve eşitliklerde kolaylık sağlamak amacıyla kullanılmaktadır. Gösterimlerde Y matrisinin y_{ij}. elemanı, i. gözlemin j. değişken için aldığı değeri ifade etmektedir. Y matrisinin diğer bir gösterim şekli ise (y_{ij})'dir.

2.1.1. Tamamlanmış Veri (Complete Data):

Tamamlanmış veri, tam veri tanımlamasıyla da kullanılmaktadır. Şekil 2.1'deki veri matrisinde y_{ij} değerlerinin tamamen elde edilmiş olması ya da kayıp veri çözümlemeleriyle tamamlanmış olması durumunda Y matrisi tamamlanmış veri matrisidir. Bu durumdaki veri matrisleri, Little ve Rubin (2002) tarafından açıklanan dikdörtgen yapı biçimine uyduğundan istatistiksel yöntemler uygulanıp elde edilen sonuçlar yorumlanabilmektedir. Dikdörtgen biçiminde veri yapısının ne olduğu kayıp veri ile ilgili tanımlamaların yapıldığı bölümde açıklanacaktır.

2.1.2. Tamamlanmamış Veri (Incomplete Data):

Şekil 2.1'deki veri matrisinin bir veya birden fazla y_{ij} değerinin elde edilememiş, kayıp olduğu durumda Y matrisi tamamlanmamış veri matrisidir. Tamamlanmamış veri, Little ve Rubin (2002)'in dikdörtgen biçimine uymadığından standart istatistiksel yöntemlerle çözümlemesi yapılamaz. Tamamlanmamış veri matrisleri için çeşitli çalışmalar yapılmış ve yöntemler önerilmiştir. Bu yöntemler kayıp veri çözümlemeleri başlığı altında incelenecektir.

2.2. Kayıp Veri Tanımı

Little ve Rubin (2002), veri matrisinin şekilsel olarak dikdörtgen biçimli bir görünüme sahip olduğunu, standart istatistiksel yöntemlerin dikdörtgen veri biçimini çözebilecek şekilde geliştirildiğini açıklamışlardır. Bu dikdörtgen biçim, veri matrisinde bulunan her gözleme ait bütün değişkenlerin, değer olarak ifade edildiği anlamına gelir. Bu durumda veri matrisinin yapısı bir dikdörtgene benzer olacaktır. Eğer herhangi bir eksiklik varsa dikdörtgen biçimi bozulacaktır.

Veri matrisindeki bazı değerlerin herhangi bir nedenle gözlenememesi durumunda kayıp veri sorunu ile karşılaşılır. Araştırmacılar, özellikle uygulamalı istatistikte karşılaşılan kayıp veri sorununu kullanacağı yönteme uygun olacak şekilde geçici çözümlerle aşmaya çalışmaktadırlar.

Kayıp veri için en basit anlatım, planlanan veri kümesi ile elde edilen veri kümesi arasındaki farklılıktır (Longford, 2006). Bu tanımın çok kapsayıcı olduğunu düşünen, özellikle sosyal bilimlerdeki ve eğitim bilimlerindeki araştırmacılar, gözlemin tümünün ya da herhangi bir değişkenin tüm değerlerinin elde edilememesinin kayıp veri olarak isimlendirilemeyeceğini belirtmişlerdir. Örneğin Allison (2000), veri matrisindeki bazı gözlemlerin birkaç değişkenine karşılık gelen değerlerin elde edilememiş olmasını kayıp veri olarak tanımlamıştır. Allison, bütün gözlemler için bir değişkenin elde edilmemiş olmasını kayıp veri olarak adlandırmamış, gizli (latent) değişken ya da gözlenmemiş değişken olarak adlandırmıştır. Aynı şekilde bazı gözlemler için bütün değişkenlerin kayıp olması durumunu gözlemin elde edilememesi sorunu olarak açıklamıştır. Allison (2000)'un tanımlama olarak yapmış olduğu bu ayrım uygulamada önemsenmemektedir. Çünkü kayıp veri çözümlemeleri için kullanılan yöntemler, elde edilememiş gözlemi ya da değişkeni bazı varsayımlar altında açıklayabilmektedir.

2.3. Kayıp Veri İle İlgilenme Nedeni

İstatistik bilimi, elde bulunan bilgiden en yüksek düzeyde yararlanarak çözümler elde etmektir. Uygun bir çözümlemeyi seçmek kadar tüm verinin anlattıklarını en sağlam bilgiye dönüştürmek önemlidir. Kayıp değerleri bulunan veri matrisleri, istatistiksel çözümlemelerde kullanılan paket programlarda sorunlar çıkartmaktadır. Bu durum bazı gözlemlerin elde edilebilecek olmasına rağmen sonuçlarda kullanılmamasına neden olmaktadır. Bu da sonuçlara bilgi kaybı olarak döner. Araştırmacılar, elde ettiği ve elde edebileceği tüm bilgiyi kullanarak daha iyi çözümlemeler yapabileceği tüm durumları incelerler. Dolayısıyla kayıp veri durumunda kullanılacak istatistiksel yöntemler eldeki veriden daha fazla bilgi sağlama sanatıdır.

Kayıp veri ile istatistiksel çalışmaların birçoğunda karşılaşılmaktadır. Örneğin, yapılan bir anket çalışmasında insanların kişisel, özellikle maddi bilgilerini saklaması, soruya cevap vermek istememesi büyük sıkıntıdır. Yine aynı şekilde endüstriyel bir süreçte teknik ve maddi kısıtlar nedeniyle süreç içinde elde edilememiş veri için tekrarın olmaması da problemdir. Kayıp veri durumunda istatistiksel çözümlemelerle yukarıda aktarılan problemleri aşmak için ilgilenilir.

2.4. Kayıp Verinin Tarihsel Gelişimi

İstatistiksel yöntemlerin gelişmesi kayıp veri sorunu için çeşitli çalışmaların yapılmasına neden olmuştur: Özellikle çok değişkenli çalışmalarda veri kaybının bulunması ve analizin yapılamaması gibi sorunlar, araştırmacıları kayıp verinin çözümüne yönlendirmiştir. Her ne kadar kayıp veriyi çözümleme çalışmaları, 1930'larda başlamış olsa da kayıp verinin bir merak ve ilgi alanı olması Rubin (1976)'in kayıp veri düzeneklerini sınıflandırmasıyla artmıştır. Little ve Rubin (1987)'in çalışması da kayıp veri çözümlemeleri için önemli bir kaynak olmuştur. Bu nedenle çalışmada, tarihsel gelişim anlatılırken Rubin (1976)'in çalışması kayıp veri için dönüm noktası olarak görülecektir.

İlk çalışmalar deney tasarımları üzerinde gerçekleşmiş ve genellikle tek kayıp değerin bulunması durumu incelenmiştir. Allan ve Wishart (1930), deneysel düzenler üzerinde bir kayıp veri bulunması durumunda çeşitli çözüm çalışmaları yapmışlardır. Rasgele blok deney düzeni ve latin kare deney düzeni içinde elde edilememiş kayıp değerin en küçük kareler (EKK) kestirimini elde etmişlerdir.

Yates (1933), bu kayıp değerin kestirilme durumunu incelemiş ve devamında (Yates,1936) yarı etkensel (quasi-factorial) tasarımlarda tamamlanmamış latin kare ve tamamlanmamış rasgele blokları açıklamıştır. Wilkonson (1958), Yates (1933)'in standart hatalardaki hesaplama hataları için yeni eşitlikler vermiştir. Wilkinson (1957), kayıp veri durumunda kovaryans çözümlemesi ile ilgili yöntem önermiştir.

Wilks (1932), parçalı örneklemlerde (fragmentary samples) kitle parametrelerinin dağılımları ve momentleri ile ilgili çalışmasında ortalama ve kovaryans kestirimlerinin nasıl bulunacağı konusunda ilk çalışmayı yapmıştır. Bu çalışma, günümüze çiftler bazında veri silme çözümlemesi olarak isimlendirilen yöntemin başlangıcıdır. Wilks (1932) çalışmasını iki değişken arasında yapmıştır. Lord (1955), eldeki kayıp veri kümesinin normal dağılımlı olduğunu düşünerek en çok olabilirlik (maximum likelihood-ML) denklemlerinin kolayca oluşturulabileceği özel bir durumu ele almış ve kitlenin parametrelerini kestirmeye çalışmıştır. Edgett (1956), aynı şekilde üç değişkenli normal dağılımlı kitle parametrelerinin ML kestirimlerini bağımlı değişkende örneklem kaybı bulunması durumunda elde etmiştir. Nicholson (1957), Edgett (1956)'in yöntemini p değişken bulunması durumunda incelemiştir. Anderson (1957), Lord ve Edgett'in yöntemleriyle de örtüştüğünü belirttiği çalışmasında, iki veya üç değişkenli normal dağılımlı kitle parametrelerinin ML kestirimlerini belli kayıp veri modelleri altında kesin çözüm eşitlikleriyle elde etmeye çalışmıştır.

Regresyon kestirimleri ile ilgili çalışmalar Wilks (1932) çalışmasıyla başlamaktadır. Kitlenin ortalama ve kovaryans kestirimleri ile ilgili yukarıda bahsedilen çalışmalarda, regresyon parametrelerinin kestirimleri de elde edilmiştir. Ancak bu çalışmalarda bazı varsayımlar kullanılmıştır: Verinin oluştuğu kitle normal dağılmaktadır, en fazla üç değişken bulunmaktadır ve kayıp yapıları özel şekillerde belirlenmektedir. Dear (1959), kayıp gözlemlerin yerine değişken ortalamalarını koyma fikrinden yola çıkarak, tüm bağımsız değişkenlerden elde edilen ortalamanın kayıp değerleri doldurduğunda birkaç kayıp değeri bulunan veri kümesinin sorunlarının çözüleceğini belirtmiştir. Glasser (1964), kayıp veri oranının ve bağımsız değişkenler arasındaki ilişkinin artmasının etkinliği azalttığını belirtmiştir. Walsh (1959), bilgisayarların çözümleme hızlarının artmasından sonra birçok gelişmenin olacağını, özellikle işlem fazlalıklarından dolayı çözümlenmeyen

problemlerin aşılabileceğini belirtmiştir. Yapmış olduğu çalışmada bağımsız değişkenin kayıp değerleri için ayrı ayrı regresyon eşitlikleri oluşturarak veri kümesini tamamlamayı ve daha sonra tamamlanmış veri kümesi üzerinden çoklu regresyon yapılmasını önermiştir. Haitovsky (1968), kayıp veri bulunan gözlemleri silerek elde edilen veri kümesine EKK uygulamış, Lord (1955)'un normal eşitlikler üzerinden uyguladığı yöntem ile karşılaştırmış ve kayıp verinin göz ardı edildiği durumun diğer yönteme göre daha iyi sonuçlar verdiğini gösteren bir benzetim çalışması yapmıştır.

Orchard ve Woodbury (1972), ilk defa kayıp veri ilkeleri tanımını kullanmışlar ve çok değişkenli normal dağılıma sahip veri kümesini tam veri kısmı ve tamamlanmamış veri kısmı olarak ayırıp elde edilecek parametre değerleri için eşitlikler oluşturmuşlardır. Kayıp verinin elde edilmesi için olabilirlik fonksiyonu üzerinden çalışma yapılmış ve en büyük olabileceği noktalardaki değerler kayıp gözlemlere yüklenmiştir. Dempster v. d. (1977), Orchard ve Woodbury (1972)'nin Beklentiyi En Büyükleme (Expectation Maximization – EM) algoritmasının başlangıç teorisini ortaya attıklarını belirtmişlerdir. Little ve Rubin (2002), EM algoritmasından önceki çalışmaların sürekli matris tersi alma gerektirdiğini, EM algoritmasının bu gereksinime ihtiyaç duymadığından üstün olduğunu açıklamışlardır.

Buck (1960), iki ve üç değişkenli durumlardaki kitle parametrelerinin kestirimlerini genelleştirerek k değişken bulunması durumunda parametre kestirimlerinin elde edilmesi için bir yöntem önermiştir. Buck (1960) gözlenmemiş durumlardaki her değişken için ayrı ayrı regresyon eşitlikleri kurmuştur ve kovaryans matrisinde oluşan yanlılık içinde bir düzeltme yöntemi göstermiştir.

Hartley (1956), tek bir kaybın kestirimi için, genel iteratif olmayan bir yöntemin, birden fazla şekilde uygulanması ile ilgili bir öneri getirmiştir. Üç farklı değeri deneyip en küçük artık kareler toplamını veren değeri kayıp gözeye yükleyen bu yöntem çok rağbet görmemiştir. Healy ve Westmacott (1956), önemli bir iteratif yöntem tanımlamışlardır. Yöntemin adımları, bir başlangıç değeri ile bütün kayıp gözlemlerin yüklenmesi, tam veri çözümlemesinin yapılması, kayıp veri için değerler kestirilmesi, kestirilmiş değerlerin başlangıç değerleriyle yer değiştirmesi ve kayıp değerler ile yerine konulan kestirimler arasındaki farkın önemseyecek

kadar küçülmesini sağlamaktır. Bu iteratif yöntem ve 1950'lerde başlayan kayıp veri durumunda kitle parametrelerinin ML ile kestirim çalışmaları, önemli bir yöntem olan EM algoritmasının kapılarını açmıştır. Pearce (1965) ve Preece (1971) bu yöntemin yakınsaklığının yavaşladığı durumları hızlandırmak için çalışmalar yapmışlardır. Ancak çalışmaların sonucunda elde edilen yöntemlerin, artık kareler toplamının monoton azalan özelliğini bozdukları Jarrett (1978) tarafından gösterilmiştir (Little ve Rubin, 2002).

Afifi ve Elashoff (1966), önceki araştırmalarda ortalama, kovaryans matrisi, ilişki matrisi ve regresyon modelleri ile ilgili yapılan çalışmaları açıklamışlardır. Tanımlamış oldukları kayıp veri yapısına göre çözümlemenin basitleştirilmesi ile ilgili bilgiler vermişlerdir. Afifi ve Elashoff (1967), basit doğrusal regresyonda bağımsız veya bağımlı değişkende kayıp olması durumunda EKK kestirimini elde etmişler ve kayıp veriye göre sınıflandırılmış yöntemlerin karşılaştırılmasını yapmışlardır. Afifi ve Elashoff (1969a), önceki çalışmalarında kullanmış oldukları kestiricilerin etkinliğini, koşullu ve koşulsuz asimptotik dağılımlarını tanımlamışlardır. Afifi ve Elashoff (1969b), doğrusal regresyonda kayıp verinin yapısına göre elde edilen kestiricilerin yan değerlerini ve küçük örneklem durumlarındaki etkinliklerini araştırmışlardır. Aynı çalışmada, küçük örneklemlerde asimptotik etkinliğin değişkenler arasındaki ilişki miktarı ve kayıp verinin yapısına göre değiştiğini göstermişlerdir.

Rubin (1976), kayıp veriyi üç durumda incelemiştir. Bu inceleme sonunda çözümlemenin hangi durumda yapılabileceğini veya yapılamayacağını göstermiştir. Bu durumlar kayıp veri düzenekleri olarak isimlendirilmiştir.

Little ve Rubin (1987), kayıp veri yapısının nasıl olabileceği konusunda önceki çalışmaları inceleyerek örnekler vermişlerdir. Kayıp veri çözümlemesi bu çalışmalardan sonra daha fazla ilgi çeker ve üzerinde çalışılır hale gelmiştir. Süreç içerisinde kayıp veri yapısı ve düzenekleri varsayımlarda kullanılırken karıştırıldıkları ya da aynı olayı anlatıyormuş gibi yanlış anlaşılmalara sebep olmuştur.

Enders (2010), kayıp veri yapısı ve kayıp veri düzenekleri ile ilgili farkı ayırt etmenin gerektiğini belirtmiştir. Kayıp veri yapısı, gözlenmiş değerler ile kayıp değerlerin veri kümesi üzerinde nasıl bir düzene sahip olduğu ile alakalıdır. Kayıp

veri düzenekleri ise kayıp veri olasılığı ve elde edilmiş değerler arasındaki mümkün ilişkileri tanımlar. Kayıp veri yapısı kaybın neden olduğunu açıklamaz, sadece verideki boş gözelerin nerede olduğunu gösterir. Kayıp veri düzenekleri ise verinin neden kayıp olduğuyla ilgili kesin bir açıklama veremese de veri kümesi ve kayıp değer arasındaki matematiksel eşitliği gösterir. İlerleyen başlıklarda iki tanımlama ile ilgili anlatımlar yapıldığında farklılıkları daha detaylı olarak görülecektir.

2.5. Kayıp Veri Gösterimi ve Kayıp Veri Gösterge Matrisi

Kayıp veri gösterge matrisi, veri matrisinden yararlanılarak oluşturulmaktadır. Kayıp veri gösterge matrisi, tasarım matrisinde gözlenen veriye "1", kayıp veriye "0" değeri konularak elde edilir (Yazıcı, 2005). Böylece K kayıp veri gösterge matrisi, Şekil 2.2'deki gibi bir matris olmaktadır.

$$K = \begin{bmatrix} k_{11} & k_{12} & \dots & k_{1p} \\ k_{21} & k_{22} & \dots & k_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ k_{n1} & k_{n2} & \dots & k_{np} \end{bmatrix}$$

Şekil 2.2. Kayıp Veri Gösterge Matrisi

Bu matris, gözlenmemiş veriyi belirtir ve kayıp veri düzeneklerinin matematiksel gösteriminde de kullanılır.

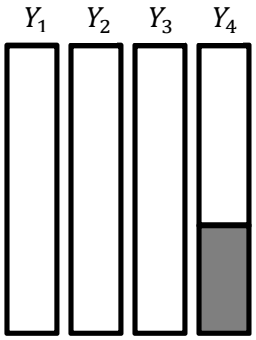
2.6. Kayıp Veri Yapıları

$Y = (y_{ij})$,kayıp verisi bulunmayan $(n \times p)$ dikdörtgen biçimli veri kümesi olsun. Veri kümesindeki kayıpları belirleyebilmek için gösterge matrisi Şekil 2.2' deki elde edildiğinde kayıp veri yapısı kolaylıkla görülür. Araştırmacılar birçok şekilde kayıp veri yapısı ile karşılaşabilirler. Yapılan çözümleme ve veri toplama tekniklerine göre karşılaşılabilecek kayıp veri yapıları altı sınıfta toplanabilir. Kayıp verinin oluştuğu kısımlar taranarak gösterildiğinde veri yapıları aşağıdaki şekilde sınıflandırılır.

2.6.1. Tek Değişkenli Yapı (Univariate Pattern)

Tek değişkenli yapı, kayıp veriyle ilgili istatistiksel çözümlemelerin başladığı yapıdır. Veri kümesindeki değişkenlerden sadece bir tanesinde kaybın oluştuğu

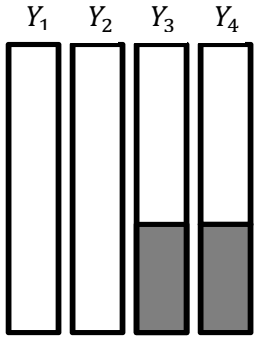
durumdur. Şekil 2.3, dört değişkenli bir veri kümesi örnek olarak alındığında üç değişkenin değerlerinin elde edildiği ancak dördüncü değişkenin bazı değerlerinin elde edilemediği tek değişkenli yapıyı göstermektedir. Araştırmacıların bu kayıp yapısı ile en çok karşılaştıkları çalışmalar genellikle deneysel çalışmalardır (Enders, 2010).



Şekil 2.3. Tek Değişkenli Yapı

2.6.2. Tek Cevapsızlık Yapısı (Unit Nonresponse Pattern)

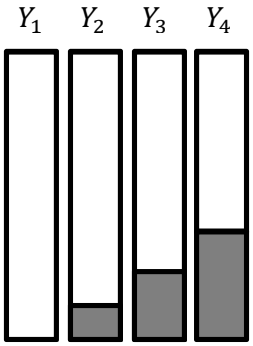
Genellikle anket çalışmalarında karşılaşılabilecek bir yapıdır. Tek cevapsızlık yapısında aynı gözlemlerin bazı değişkenlerdeki değerleri elde edilemez. Örneğin, Şekil 2.4, dört değişkenli bir veri kümesini gösterdiğinde, elde edilecek veri kümesindeki Y_1, Y_2 değişkenleri kayıpsızdır (Enders, 2010). Ancak Y_3, Y_4 değişkenlerinde kayıplar bulunmaktadır ve bu kayıplar aynı gözlemlerden oluşmaktadır. Örneğin anket çalışmalarında deneklerin belli soruları cevaplamayı reddetmesi sonucunda oluşabilecek bir yapıdır. Aynı şekilde endüstriyel bir çalışmada örneklemlerin toplandığı Y_1, Y_2 değişkenlerinin maliyet konusunda sorun yaratmadığı ve Y_3, Y_4 değişkenlerinin belli bir gözlem sayısından sonra fazla maliyetli olacağı düşünüldüğünde Şekil 2.4 yapısının ortaya çıkacağı açıktır.



Şekil 2.4. Tek Cevapsızlık Yapısı

2.6.3. Tekdüze Yapı (Monotone Pattern)

Tekdüze kayıp veri yapısı genellikle uzun süreli (longitudinal) çalışmalarda karşılaşılabilecek bir yapıdır. Her değişken değiştiğinde elde bulunan gözlem sayısının sürekli azalması olarak aktarılabilir ve yapısı, dört değişkenli bir veri kümesi örnek olarak alındığında Şekil 2.5 gibi gösterilebilir:



Şekil 2.5. Tekdüze Yapı

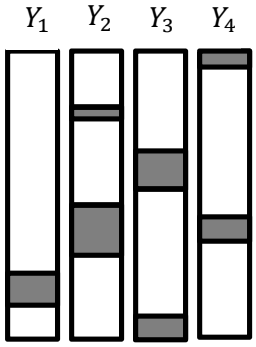
Enders (2010), bu konuyla ilgili yeni bir ilacın deneme çalışmalarını örnek olarak vermiştir. İlacı kullandıkça yan etkiye uğramış gözlemlerin çalışmadan çıkarılması sonucu elde edilecek kayıp veri yapısı tekdüzedir.

Bu veri yapısı, çoklu veri yükleme, model tabanlı veri yükleme gibi kayıp veri çözümlemelerin kullanımı ile ilgili birçok araştırmaya sebep olmuştur. Enders (2010), bu araştırmanın sebebi olarak kayıp verinin belli bir düzende olduğunu, matematiksel karışıklığın bu şekilde azaldığını ve iteratif kestirimler için gerekli varsayımların kolaylaştığını belirtmiştir. Schafer (1997), verinin normal dağıldığı durumlar altında tekdüze kayıp veri ile ilgili detaylı bir çalışma yapmıştır.

2.6.4. Genel Yapı (General Pattern)

Genel yapı olarak kabul edilen kayıp veri yapısı kayıpların rasgele olmasıdır. Sık olarak karşılaşılabilir ve dört değişkenli bir veri kümesi için Şekil 2.6 gibi gösterilir. Buradaki kayıplar, gelişigüzel bir şekilde olabilir ama sistemli olarak bir kayıp yapısı olup olmadığına dikkat edilmelidir. Örneğin, Y_2 değişkenindeki kayıp olma eğiliminin Y_1 değişkeni ile ilişkili olup olmadığı önemlidir.

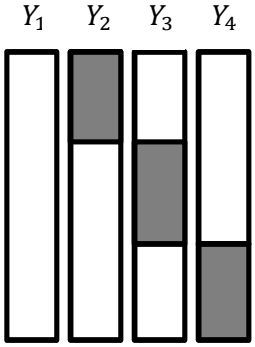
Bu ilişki kayıp verinin yapılarından değil kayıp verinin düzeneklerinden bilinebileceği unutulmamalıdır (Enders, 2010).



Şekil 2.6. Genel Yapı

2.6.5. Dosya Eşleşme Yapısı (File Matching Pattern)

Dosya eşleşme yapısı, değişkenler arasındaki ilişkiler tam veri olmadan elde edilebilecek şekilde özel olarak belirlenir. Böylece ikili ilişkiler üzerinden çözümlemeler yapılabilecek şekilde gözlem sayısı parçalı olarak azaltılabilir. Örneğin, dosya eşleşme yapısı, Şekil 2.7'de dört değişkenli bir veri kümesi için gösterilmiştir:



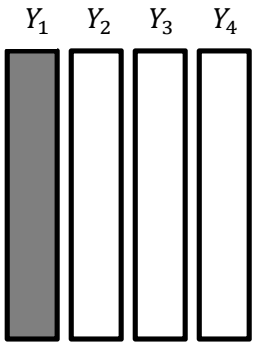
Şekil 2.7. Dosya Eşleşme Yapısı

Şekil incelendiğinde, değişkenlerin herhangi biri tamamen gözlenmiş; diğer değişkenler ise birbirleriyle gözlem sayısı kadar gözlenememiştir. Enders (2010), bu durumun planlı bir şekilde de tasarlanabildiğini belirtmiştir. Graham v. d. (1996), üç formlu anket çalışması da planlı bir tasarım örneğidir. Bu tasarımlarla genellikle çok fazla soruların bulunduğu anketlerde çalışılırken karşılaşılabilir.

2.6.6. Gizli Değişken Yapısı (Latent Variable Design)

Gizli değişken yapısında, değişkenlerden biri örneklemin tamamında gözlenmemiştir. Dört değişkenin bulunduğu bir veri kümesinde gizli değişken yapısı, Şekil 2.8'deki gibi gösterilebilir. Bu yapı için yapısal eşitlik modelleri (YEM) üzerinden çeşitli kayıp veri yöntemleri önerilmektedir. Açıklayıcı faktör

çözümlemesi (Confirmatory Factor Analysis), değişkenler arasındaki ilişkileri açıklamak için faktör yük değerlerini kullanır. Bu yük değerleri tamamen kayıptır. Yük değerleri gizli değişken olarak kabul edilebilir ve yük değerlerini bulmak için kayıp veri yükleme yöntemleri kullanılabilir (Enders, 2010).



Şekil 2.8. Gizli Değişken Yapısı

2.7. Kayıp Veri Düzenekleri (Missing Data Mechanism)

Kayıp veri yapısı, düzenek çeşitlerinden hangisine uyuyorsa yöntem, düzeneğin izin verdiği koşullarda çözüme kavuşur. Aksi halde, kayıp veri çalışması, düzenek tarafından uygun olmayan yöntemle çözümlenebilir ve yanlı kestirimler elde edilebilir. Herhangi bir kayıp verinin nasıl oluştuğunu incelemeden uygun olmayan kayıp veri yöntemleri kullanımı çok iyi performans sağlamaz. Diğer bir deyişle, kayıp veri düzeneklerini göz önünde bulundurmadan yapılacak bir çalışmada daha fazla bilgi elde edebilmek adına kullanılan kayıp veri çözümlemesi deyim yerindeyse "kaş yapayım derken göz çıkartmak" gibi olacaktır. Bu yüzden kayıp veri düzeneklerinin varsayımlarına göre çözümlemeler yapmak gerekmektedir.

$Y = (y_{ij})$ veri matrisi ile $K = (k_{ij})$ kayıp veri gösterge matrisi ile tanımlanmıştı. θ bilinmeyen parametreleri göstersin. $Y_{göz}$, Y veri kümesinin gözlenmiş kısmını; Y_{kay} ise Y veri kümesinin kayıp kısmını betimlesin. Kayıp veri düzenekleri, Y verildiğinde K'nın koşullu dağılımı ile tanımlanır. Bu koşullu dağılım veride $f(K|Y, \theta)$ ile gösterilmektedir (Little ve Rubin, 2002).

Rubin (1976), kayıp veri düzeneklerinin kayıp veri kuramında çok önemli olduğunu belirtmiş ve kayıp veri düzeneklerini üçe ayırmıştır. Günümüzde birçok çalışma (Donders v. d., 2006; Enders, 2010; Allison, 2000) ile Rubin (1976)'nin teorik çalışması daha açıklayıcı ve örneklendirilebilir hale getirilmiştir.

2.7.1. Tamamen Rasgele Kayıp (TRK) (Missing Completely At Random-MCAR)

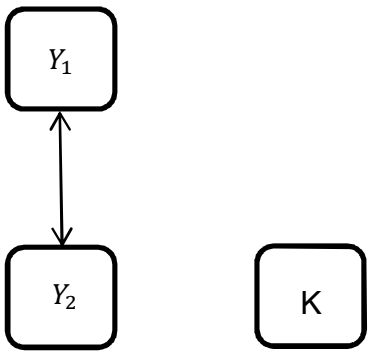
Tamamen rasgele kayıp düzeneği araştırmacıların kayıp veri çözümlemelerinde kolaylarına gelen bir düzenektir. Ama aynı zamanda varsayımları göz önünde bulundurulmadan yapılan çalışmalarda araştırmacıları yanlış çözümlere götürebilir. TRK düzeneğinde herhangi bir kayıp veri, ne diğer değişkenlerle ne de olduğu değişken ile ilişkilidir. Çok güçlü bir varsayımdır.

Donders v. d.(2006), verinin TRK olduğu bilindiğinde, kayıp veri içermeyen değişkenler kümesinin, hedef kitlenin rasgele örneklemi olduğunu belirtmiştir. Bu nedenle, kayıp veri incelemeleri için TRK varsayımında birçok yöntem yansız sonuçlar vermektedir (Satıcı, 2009).

Y verildiğinde K'nın koşullu dağılımı TRK için Eş. 2.1'deki gibi gösterilir:

$$f(k|y,\theta) = f(k|\theta) \tag{2.1}$$

İki başlı oklar istatistiksel birlikteliği anlatmak üzere, Şekil 2.9, TRK durumunun iki değişkenli veri kümesindeki görsel ifade biçimidir. Burada, K kayıp veri gösterge matrisi, Y_2 değişkenindeki kayıplar yardımıyla oluşturulsun. K'nın istatistiksel ilişkisi bulunan herhangi bir değişken olmadığı görülebilir.



Şekil 2.9. TRK Düzeneğinin İlişki Gösterimi

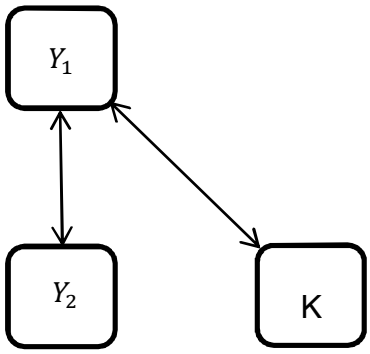
2.7.2. Rasgele Kayıp (RK) (Missing At Random -MAR)

Herhangi bir değişkende kayıp veri olması olasılığı, modeldeki diğer değişkenlerle ilişkili ve kendi değişkenindeki gözlenen değerlerle ilişkili değilse kayıp veri rasgele kayıptır. Enders (2010), RK durumunun aslında kayıp veri olasılığı ile bir veya

daha fazla değişkenin arasındaki sistematik ilişkiyi açıkladığını belirtmiştir. Bu durumda kayıp veri RK olacaktır. Genellikle veri RK olduğu durumda, veri yükleme gibi basit yöntemler yanlı sonuçlar vermektedir. Bununla birlikte, model tabanlı veri yükleme ve çoklu veri yükleme gibi daha ayrıntılı yöntemler yansız sonuçlar vermektedir (Satıcı, 2009; Donders v. d., 2006). Y ve $Y_{göz}$ verildiğinde K'nın koşullu dağılımı RK için aşağıdaki gibi gösterilir:

$$f(k|y,\theta) = f\left(k\middle|y_{göz},\theta\right) \tag{2.2}$$

İki başlıklı oklar istatistiksel birlikteliği anlatmak üzere, Şekil 2.10, iki değişkenli durum için RK durumunun görsel ifade biçimidir. K kayıp veri gösterge matrisi, Y_2 değişkenindeki kayıplardan oluşturulsun. K'nın istatistiksel ilişkisi bulunan değişkenin sadece Y_1 olduğu görülebilir.



Şekil 2.10. RK Düzeneğinin İlişki Gösterimi

TRK, RK'den daha kısıtlayıcı bir kayıp veri düzeneği olduğu görülmelidir. TRK durumunda kayıp veri olasılığı herhangi bir değişkene bağlı değilken, RK durumunda kayıp veri olasılığının diğer değişkenlerle ilişkili olduğu görülmektedir. RK, ihmal edilebilir kayıp olarak düşünülerek çeşitli çözümlemelerde sıkça kullanılmıştır. Rubin (1976) ise örnekleme dayalı olarak yapılan çözümleme tekniklerinin TRK olması durumunda geçerli olabileceğini göstermiştir.

Kayıp verinin TRK varsayımını sağlaması geleneksel yöntemler için uygundur. Ancak TRK'nin kısıtlayıcı, zor bir varsayım olması ve model tabanlı veri yükleme yöntemlerinin önerilmesi, geleneksel yöntemlerin terkedilmesine neden olmuştur. Bu süreçte TRK ve RK varsayımları altında yansız kestirimler elde edebilen ML kestirimleri ve çoklu veri yükleme yöntemlerinin rolü büyüktür (Wilkinson,1999).

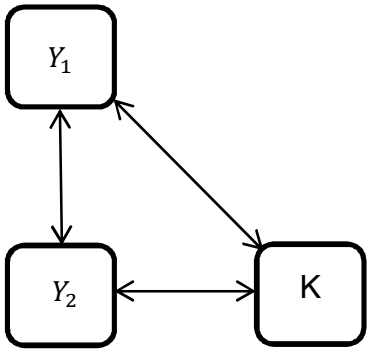
2.7.3. Rasgele Olmayan Kayıp (ROK) (Missing Not At Random - MNAR)

Herhangi bir değişkendeki kayıp verinin olasılığı modeldeki diğer değişkenlerle ve aynı zamanda kendi değişkeniyle de ilişkiliyse kayıp veri ROK'dir. RK ve TRK olmayan verinin ROK olduğu da söylenebilir. ROK düzenekleri altında kayıp veri çözümlemesi mevcut yöntemleri kullanmaya uygun değildir.

Şekil 2.11, ROK durumunun iki değişken bulunduğu durumda görsel ifade biçimidir. Şekil incelendiğinde, K kayıp veri gösterge matrisi, Y_2 değişkenindeki kayıplardan oluşturulsun. K'nın istatistiksel ilişkisi bulunan değişkenlerin Y_1 ve Y_2 olduğu görülebilir.

Kayıp verinin uyduğu düzeneğe göre veri silme, veri yükleme veya model tabanlı yöntemleri kullanma durumları ortaya çıkar. Kayıp veri çözümlemeleri için kullanılan yöntemler ileride daha detaylı anlatılacak ve kayıp veri düzenekleri ile olan ilişkileri açıklanacaktır. Burada çözümleme yöntemlerine değinme amacı, TRK, RK ve ROK ile ilgili örneklerin açıklanabilmesidir.

TRK veri düzeneğinde liste bazında veri silme (listwise deletion), çiftler bazında veri silme (pairwise deletion) ya da veri yükleme yöntemleri kullanılabilir. RK veri düzeneğinin varsayımlarından dolayı veri yükleme yöntemleri yanlı sonuç verirken model tabanlı yöntemlerdeki sonuçlar daha etkindir.



Şekil 2.11. ROK Düzeneğinin İlişki Gösterimi

Konu ile ilgili örneklendirme yapabilmek için Enders (2010)'in çalışmasındaki örnek burada verildi. Belirli bir dönem için iş performansları ve zeka düzeyleri ile ilgili bilgi Çizelge 2.1'deki gibi elde edilmiştir. Performans puanları, TRK, RK ve ROK şeklinde oluşturulmuştur ve tartışmalar yapılmıştır. Kayıp veri düzeneklerine göre ayrı ayrı araştırma konusuymuş gibi kullanılmıştır.

Çizelge 2.1. Zekâ Puanı ile TRK, RK ve ROK Düzeneklerinden Oluşturulmuş İş Performansları

Sıra	Zekâ Puanı	İş Performansları			
		Tam Veri	TRK	RK	ROK
1	78	9	—	—	9
2	84	13	13	—	13
3	84	10	—	—	10
4	85	8	8	—	—
5	87	7	7	—	—
6	91	7	7	7	—
7	92	9	9	9	9
8	94	9	9	9	9
9	94	11	11	11	11
10	96	7	—	7	—
11	99	7	7	7	—
12	105	10	10	10	10
13	105	11	11	11	11
14	106	15	15	15	15
15	108	10	10	10	10
16	112	10	—	10	10
17	113	12	12	12	12
18	115	14	14	14	14
19	118	16	16	16	16
20	134	12	—	12	12

İlk olarak, iş performansları ile ilgili değerler TRK olacak şekilde rasgele silinmiştir. Yani silinmiş olan 5 veri ne iş performansı değişkeniyle ne de zekâ düzeyi ile ilişkidir. Bu şirketin iş performansı ile ilgili çözümlemeler, geriye kalan 15 veri üzerinden elde edilebilir.

Aynı örnek şirkette işe alınacak olan 20 kişi için incelensin. Bu incelemede işe alma kararının deneme süresindeki iş performansları dikkate alınmadan zekâ puanları üzerinden verildiği varsayılsın. Kayıp veri, RK olacak şekilde elde edilmiştir. Kayıp verinin RK durumunda iş performansı ile ilgili bilginin bulunmadığı 5 kişi en düşük zekâ puanındaki kişilerden oluşmaktadır. Kayıp veri RK olduğundan gözlenemeyen iş performansları kendi değişkenlerindeki değerlere değil; zekâ düzeyi değişkenine bağlı olacaktır.

Son olarak, ROK düzeneği de aynı çizelgeden kontrol edildiğinde hem iş

performanslarına bağlı hem de zekâ düzeylerine bağlı olduğu varsayılsın. Örneğin işyerindeki 20 kişiden iş performansı düşük olanların işine son verileceği düşünülsün. Performansı en düşük olan 5 kişinin iş performansının elde edilemediğinden her ne kadar zekâ düzeyleri kontrol edilse de aynı zamanda iş performanslarıyla da ilişkili olduklarından verinin yapısı ROK'dir.

Kayıp veri düzenekleri TRK ile RK arasında bazı farklılıklar vardır. Bu farklılıklar yapılacak çözümlemelerde önemli etkilere sahip olmaktadır. Veri düzeneğinin TRK veya RK olması arasındaki farklılıklar anlatımlarda aktarılmış olsa da genel haliyle Çizelge 2.2'de görülebilir.

Çİzelge 2.2. TRK ve RK Düzenekleri Arasındaki Farklılıklar

	TRK	RK
1	TRK düzenekler içerisinde en kısıtlayıcı olanıdır.	TRK'dan daha az kısıtlayıcıdır.
2	TRK düzenekleri geleneksel veri yükleme yöntemlerinde yansız sonuçlar verebilmektedir.	RK, ML kestirim ve çoklu veri yükleme gibi yöntemlerde yansız sonuçlar verebilmektedir.
3	TRK, Little tarafından geliştirilen bir testle test edilebilir bir düzenektir (Little'ın TRK Testi).	RK için test edilebilir bir yöntem geliştirilememiştir.
4	TRK, kısıtlayıcı özelliğinden dolayı birçok kayıp veri düzeneğinde sağlanamaz.	RK kayıp veri düzeneğini elde etmek daha kolaydır.

TRK düzeneği, test edilebilen tek düzenektir. Dixon (1988), kayıp verinin alt gruplarını bağımsız t testi ile karşılaştırmıştır. Little (1988b), t testi yaklaşımını genişleterek veri kümesindeki her değişkenin ortalama farklarını hesaplamıştır. Test istatistiği alt grupların ortalamaları ile bütün ortalamaların arasındaki standartlaştırılmış farkların ağırlıklandırılmış toplamıdır. RK için test istatistiği geliştirilememiştir. Test istatistiğinin eşitliği, anlatımlı örnekleri ve RK durumu için çeşitli şekillerde planlı kayıp veri yapıları Enders (2010)'den takip edilebilir.

2.8. Kayıp Veri Çözümlemeleri

Kayıp veri çözümlemeleri ile ilgili yapılan ilk çalışmalar, tek kayıp değer ya da tek değişkende birkaç kayıp değer çözümlemelerinden başlamıştır. Ancak bilgisayarların hızları ve programların kolaylığı sayesinde günümüzde birçok çözüm yolu geliştirilmiştir. Kayıp veri düzeneğinin bu çözümlemelerdeki rolü önemlidir. Yanlı kestirimler elde etmemek için kayıp verinin hangi değişkenlere bağlı ya da bağımsız olduğunu bilmek gerekmektedir. Little ve Rubin (2002), kayıp veri yöntemleri ile ilgili genel bir sınıflandırma yapmışlardır. Bu çalışmada Little ve Rubin (2002) sınıflandırması baz alınmıştır. Ancak, başlıkların altında geliştirilen ve başka kaynaklardan elde edilen diğer yöntemlerde eklenmiştir.

2.8.1. Veri Silme Yöntemleri

Kayıp veri sorununu aşmanın, veri yapısını dikdörtgen biçimine taşımanın en basit biçimi, günümüzde bile kullanılan veri silme yöntemleridir. Ancak bu yöntemlerin TRK varsayımını sağlaması gerekmektedir. Tam durum çözümlemelerinde ise kayıp gözlemlerin bulunduğu değişkenlere ait gözlemler silinebilir ya da ikişerli ilişkiler üzerinden hesaplamalar yapılarak sonuçlar elde edilir. Özellikle eğitim ve sosyolojik araştırmalarda kayıp veri bulunması durumunda başvurulan yöntemlerdir. Bu çözümlemeler anlatılırken Atalay (2003) çalışmasındaki basit örnekten faydalanılacaktır ve veri kümesi Çizelge 2.3'deki gibidir. Veri silme yöntemleri, liste bazında veri silme ve çiftler bazında veri silme olmak üzere ikiye ayrılmaktadır.

2.8.1.1. Liste Bazında Veri Silme (Listwise Deletion – Complete-Case Analysis):

Kayıp verinin bulunduğu gözlemlerin silinip geriye kalan gözlemler üzerinden çözümlemelerin yapıldığı bir yöntemdir. Kayıp verinin bulunduğu gözlemler çıkartıldıktan sonra herhangi bir değişiklik yapılmadan istatistiksel çözümlemeler yapılabildiğinden uygulama olarak kolay bir yöntemdir. Tek değişkenli istatistiklerde karşılaştırmalar kolayca yapılabilir olmasına rağmen, kayıp değerlerin bulunduğu gözlemler silinirken bilgi kaybına sebep olmaktadır. Little ve Rubin (2002), bu bilgi kaybının kesinliği etkileyeceğini ve kayıp veri düzenek varsayımının TRK olmaması durumunda yanlı kestirimler elde edilebileceğini belirtmiştir. Ayrıca Baygül (2007), liste bazında veri silme yöntemi kullanılırken

örneklemdeki gözlem sayısı azaltılacağından standart hataların artacağını belirtmiştir. Kayıp veri düzeneğinin RK ya da ROK olması durumunda yanlı kestirimler elde edilecektir.

Çizelge 2.3. Veri Silme Yöntemleri için Örnek Veri Kümesi

Gözlem	Y_1	Y_2	Y_3
1	13	23	21
2	14	22	17
3	15	-	11
4	16	18	-
5	17	17	12
6	-	20	8
7	-	20	15

Çizelge 2.3'deki veri kümesi göz önünde bulundurulsun. Gözlenen değerlerin 21 tanesinden 4 tanesinin kayıp olduğu görülmektedir. Veri, TRK durumu altında elde edilmiştir.

Liste bazında veri silme yöntemi uygulandığında veri kümesinde sadece birinci, ikinci ve beşinci gözlemler üzerinden istatistiksel çözümlemeler yapılacaktır. Üçüncü, dördüncü, altıncı ve yedinci gözlemler kayıp veri bulundurduklarından çözümleme dışında bırakılmıştır. Gözlenmiş olmasına rağmen çözümleme dışında kalan değerlerin oranı %38'dir. Dolayısıyla kayıp değerlerin yaratmış olduğu sorunlardan kurtulmak için uygulanan bu yöntem gözlemlerin büyük bir kısmını da çözümleme dışında bırakmaktadır.

Birçok araştırmacı, liste bazında veri silme yönteminin uygulanabilmesi için kaybedilecek olan bilginin kabul edilebilir düzeyde olması gerektiğini belirtmiştir.

2.8.1.2. Çiftler Bazında Veri Silme (Pairwise Deletion – Available-Case Analysis):

Liste bazında veri silme yöntemindeki bilgi kaybının fazlalığı göz önünde bulundurulduğunda, çiftler bazında veri silme yöntemi veri kümesini ikili değişkenler üzerinden düşünerek bilgi kaybını azaltmaya çalışır. Ancak ikili değişkenlerde gözlenmiş olan gözlemler dâhil edilirken diğerleri gözlem kümesinden çıkartıldığı için yine bilgi kaybı söz konusudur. Her ne kadar liste bazında veri silme yöntemine göre daha etkili bir yöntem olsa da veri kümesinde elde edilmiş bilgileri kaybederek çözümleme yapmak akıllıca değildir.

Glasser (1964), bu yöntemin liste bazında veri silmeden daha etkili bir yöntem olmasına rağmen ilişki matrisini olması gereken değerlerde sınırlandıramadığını belirtmiştir. Enders (2010) ise her değişkenin farklı örneklem sayısı bulundurmasının ANOVA ve regresyon çözümlemesi gibi yöntemlerin kullanılacağı durumlarda problemli olduğunu belirtmiştir. Çiftler bazında veri silme yöntemi, liste bazında veri silme yönteminde olduğu gibi TRK veri düzeneği varsayımı altında etkili kestirimler verebilir.

Çizelge 2.3'deki örnek, çiftler bazında veri silme ile ele alındığında birinci ve ikinci değişken için birinci, ikinci, dördüncü ve beşinci gözlemler üzerinden işlem yapılacaktır. Birinci ve üçüncü değişken için yapılacak kestirimler, birinci, ikinci, üçüncü ve beşinci gözlemler üzerinden yapılacaktır. İkinci ve üçüncü değişkenler baz alındığında ise üçüncü ve dördüncü gözlem hariç diğer gözlemler üzerinden çözümlemeler yapılacaktır. Yapılan işlemler dikkate alındığında ilk altküme için dört gözlem çifti, ikinci altküme için dört gözlem çifti ve üçüncü altküme için beş gözlem çifti üzerinden çözümlemeler yapılarak birleştirilecektir. Görüldüğü gibi altkümelerin gözlem sayıları birbirlerinden farklı olabilmektedir. Liste bazında veri silme yönteminde gözlenmesine rağmen çözümleme dışında kalan değerlerden oluşan bilgi kaybı, çiftler bazında veri silme yöntemi ile azaltılmıştır.

Veri silme yöntemlerinde kayıp veri problemini çözümleyebilmek aslında tam veri yöntemini elde ederek istatistiksel çözümlemeleri yapabilecek durumu elde etmek için kullanılabilecek özel amaçlı, geçici (ad-hoc) yöntemlerdir. Wilkinson (1999), liste bazında ve çiftler bazında veri silme yöntemlerinin uygulamada kullanılabilecek en kötü yöntemler olduğunu belirtmiştir. Bu iki yöntem, daha sonra bahsedilecek olan veri yükleme, çoklu veri yükleme, EM algoritması gibi detaylı çalışmaların yanında özellikle veri kayıpları büyük olduğu durumlarda etkisiz kalmaktadır.

2.8.2. Ağırlıklandırma Yöntemleri

Ağırlıklandırma yöntemleri liste bazında veri silme yöntemlerinde ortaya çıkan yanlılık problemlerini çözebilmek amacıyla farklı ağırlıklarla ortalama ve varyans kestirimlerini oluşturmayı amaçlamıştır. Bu yöntemin ortaya çıkmasındaki düşünce, sonlu kitle anketleri için rasgeleleştirme çıkarımlarında yapılan ağırlıklandırmalardır.

Anket veri kümesindeki kayıtlar, genellikle örneklem ağırlıkları ile ilişkilidir. Ağırlıklandırılmış yöntemlerde gözlenen değerlere verilen ağırlıklar sadece bu gözlemlerin kendilerini değil aynı zamanda kayıp veriyi de temsil etmeleri esasına dayanır (Satıcı, 2009). Geleneksel yöntemlerdeki yanlılık sorununa çözüm bulmaya çalışmış olmasına rağmen kayıp veri çözümlemeleri içinde ağırlıklandırılmış yöntemin uygulaması çok azdır. Little ve Rubin (2002) çalışmasından daha detaylı açıklamalara ulaşılabilir.

2.8.3. Veri Yükleme (Imputation) Yöntemleri

Kayıp değerlerin bulunduğu durumda veri silme yönteminin bilgi kaybına sebep olması araştırmacıları kayıp değerlerin bulunmasına yönlendirmiştir. Veri yükleme yaparak kayıp değerlerin yerlerinin doldurulması da standart istatistiksel yöntemler için gerekli olan dikdörtgen biçimi tamamlayacağından araştırmacılar tarafından ilgi çekici olarak görülmüştür.

Kayıp olan verinin yerine yükleme yapılırken dikkat edilmesi gereken durumlar vardır. Veri kaybının oranı, kayıp veri yapısı, kayıp veri düzenek varsayımı koşullarına göre hangi yöntemlerin kullanılacağı belirlenmelidir. Bu varsayımlar göz önüne alınmadan yapılacak çözümlemelerin vereceği sonuçlar güvenilir olmayacaktır. Veri kümelerindeki kayıp oranlarına göre veri yapısından kayıp veri düzeneklerine kadar birçok şekilde kayıp veri oluşabilir. Little ve Rubin (2002), veri yüklemenin fark edilemeyebilir tehlikeleri olarak tanımladıkları varsayımlar için Dempster ve Rubin (1983)'in anlatımlarını aktarmışlardır:

"Veri yükleme fikri hem cezbedici hem de tehlikelidir. Kolaylığı ve sonucunda tam-veri kümesinin bulunmasından dolayı cezbedicidir. Problemi çözdüğü zaman ise gerçek veri ile yüklenen veri arasında büyük bir yanlılık olması durumu da tehlikelidir."

Veri yükleme yöntemleri oldukça fazladır. Çalışmanın bu bölümünde en çok kullanılan veri yükleme yöntemleri ile ilgili bilgiler verilecektir.

2.8.3.1. Ortalama ile Veri Yükleme (Mean imputation)

Aritmetik ortalama yükleme (arithmetic mean imputation, mean substitution), koşulsuz ortalama yükleme (unconditional mean imputation) olarak da aktarılabilen bir veri yükleme yöntemidir. Birçok araştırmacı, Wilks (1932)'e atıf

yaparak bu yöntemi kullanmaya başlamıştır. Veri matrisindeki kayıplara, ait olduğu değişkenin elde edilen gözlemleri üzerinden hesaplanan ortalamanın yüklenmesidir. Ortalama yükleme veriyi tam veri haline getirdiği için kullanıma elverişli olduğu düşünülebilir. Ancak çözümlemelerde kayıp veri düzeneğinin TRK olduğu durumlarda bile parametre kestirimlerine zarar vermektedir. Bu durumu engellemek için Dear (1959), değişkenlerin hepsinin genel ortalamasını yükleme (over all mean imputation) gibi bir yöntem önermiştir. Ama bütün kayıp gözlemlerin tek bir veri ile doldurulması varyansları küçültmektedir ve verinin dağılımını bozmaktadır. Dağılımın merkezindeki bir veriyi kayıp değerlerin yerine koymak verinin değişkenliğine etki etmek olduğu bir gerçektir.

Ortalama veri yükleme ve ağırlıklandırma ile ilgili daha fazla örnek ve açıklama David v. d.(1983) çalışmasında görülebilir (Little ve Rubin,2002). Little ve Rubin (2002), bu sorunları çözümleyebilmek için varyans, kovaryans formüllerinde düzeltme eşitlikleri vermiş olsalar da bu düzeltmeler sonucunda elde edilen kestirimler, çiftler bazında veri silme kestirimleriyle aynı olmuştur (Enders, 2010).

2.8.3.2. Regresyon İle Veri Yükleme (Regression Imputation)

Doğrusal regresyon, bağımlı değişkenin bağımsız değişkenlerce doğrusal bir eşitlikle açıklanması yöntemidir. Elde bulunan değişkenlerden sadece bir tanesinde kayıp veri bulunduğu varsayılsın. Bu durumda tam veri olan değişkenlerin bağımsız değişken, kayıp veri bulunduran değişkenin bağımlı değişken olarak kabul edilmesi bir çözümleme getirebilir. Gözlenmiş değerler üzerinden elde edilen regresyon modeli kullanılarak gözlenmemiş değerlerin bulunması ile yöntem tamamlanmış olur. Regresyon yöntemi ile veri yüklemenin en basit hali bu şekildedir.

Buck (1960)'ın yöntemi, kayıp verinin genel yapısı için bu basit regresyon veri yükleme yöntemini genişletmiştir. Tam veri durumundan ortalama vektörü ile kovaryans matrisi kestirilir. Bu kestiricileri kullanarak her kayıp veri yapısı için kayıp değerlerin doğrusal regresyonda EKK kestiricileri elde edilir. Kayıp gözlemlerin bulunduğu değişkenler için farklı regresyon eşitlikleri oluşturmak da mümkündür. Çözümlemeler TRK varsayımı altında etkilidir.

Buck (1960)'ın yönteminde varyans ve kovaryanslar, gerçek değerlerinden daha aşağıda kestirilmektedir (Little ve Rubin,2002). Çünkü değerler verinin dağıldığı

doğru üzerindeki noktalardan elde edilmektedir. Bu durum yine merkezileştirmeyi getirdiğinden varyans ve kovaryans kestirimlerinde etkin kestirim elde edilememektedir. Kayıp verinin her değişkende bulunması durumu da aynı mantıkla bulunmasına rağmen, biraz daha karışık haldedir. Daha açıklayıcı olması açısından kayıp değerleri bulunan üç değişkenin olduğunu düşünülsün.

Değişkenlerin kayıp veri içeren değişkenleri ve bu durumda kullanılabilecek regresyon modelleri Çizelge 2.4'deki gibidir. Çizelge incelendiğinde elde edilmiş gözlemler üzerinden yapılacak olan kestirimler sonucunda kayıp değerlerin nasıl bulunacağı kolayca belirlenebilmektedir. Burada önemli bir problemle karşılaşılmaktadır. Kayıp veri için elde edilen değerler diğer değişkenlerle aralarında tam bir ilişki olacak şekilde doldurulmaktadır. Dolayısıyla ilişkiler ve R^2 istatistikleri artacaktır (Enders,2010).

Çizelge 2.4. Kayıp Veri için Regresyon Eşitlikleri

Kayıp Verili Değişkenler	Regresyon Eşitlikleri	
Y_1	$\hat{y}_1 = \beta_0 + \beta_1 y_2 + \beta_2 y_3$	
Y_2	$\hat{y}_2 = \beta_0 + \beta_1 y_1 + \beta_2 y_3$	
Y_3	$\hat{y}_3 = \beta_0 + \beta_1 y_1 + \beta_2 y_2$	
Y_1 ve Y_2	$\hat{y}_1 = \beta_0 + \beta_1 y_3$	$\hat{y}_2 = \beta_0 + \beta_1 y_3$
Y_1 ve Y_3	$\hat{y}_1 = \beta_0 + \beta_1 y_2$	$\hat{y}_3 = \beta_0 + \beta_1 y_2$
Y_2 ve Y_3	$\hat{y}_2 = \beta_0 + \beta_1 y_1$	$\hat{y}_3 = \beta_0 + \beta_1 y_1$

2.8.3.3. Stokastik Regresyon ile Veri Yüklemesi (Stochastic Regression Imputation)

Regresyon ile veri yüklemeleri yapılabilmesi için doğrusal regresyon doğrusunun üzerindeki değerlerin kayıp veri yerine kullanılması nedeniyle verinin dağılışları etkilenmektedir. Aynı zamanda kovaryansların beklenenden daha az olacak şekilde kestirilmesi gibi sorunlarla karşılaşılmaktadır. Stokastik regresyon yöntemi ile bu sorunlara çözüm bulmak istenmiştir. Elde edilmiş değerler üzerinden kurulan regresyon modeline sıfır ortalamada normal dağılan artıklar eklenmiştir. Enders (2010), stokastik regresyon ile veri yükleme yönteminin kayıp veri düzeneği

varsayımı RK olan veride bile yansız parametre kestirimi veren bir yöntem olduğunu belirtmiştir.

Örnek olarak yine üç değişkenli (Y_1, Y_2 ve Y_3) ve tek değişkende (Y_3) kayıp olan bir veri matrisi düşünülsün. Bu durumda regresyon eşitliği Eş. 2.3'deki gibi olurken stokastik regresyon eşitliği ise Eş. 2.4'deki gibi olmaktadır. z olarak eklenen değişken sıfır ortalamaya ve elde edilmiş gözlemler üzerinden bulunan varyans ile normal dağılmaktadır.

$$\hat{y}_3 = \beta_0 + \beta_1 y_1 + \beta_2 y_2 \tag{2.3}$$

$$\hat{y}_3 = \beta_0 + \beta_1 y_1 + \beta_2 y_2 + z \tag{2.4}$$

Enders (2010), RK varsayımı altında yansız parametre kestirimleri üretebilmesi özelliğinden dolayı stokastik regresyon ile veri yükleme yönteminin, ML ve çoklu veri yükleme yöntemlerine alternatif olduğunu belirtmiştir. Little (1992), bağımsız değişkenlerde kayıp veri bulunması durumunda regresyon ile veri yükleme yöntemini kullanmanın, kestirimde oluşturduğu yanlardan dolayı yanlış olacağını belirtmiştir. Stokastik regresyon ile veri yükleme yönteminde ise kestirimlerdeki yan sorunu, eklenen rasgele artık terimleri tarafından çözümlenebilmektedir.

2.8.3.4. Deste Yardımıyla (Deck) ile Veri Yükleme

En basit şekilde tanımlanmak istenirse kayıp değerlerin, benzer durumda olan gözlenmiş değerler yardımıyla doldurulmasıdır. Bu yöntem daha çok sosyal bilimlerde ve eğitim bilimlerinde kullanılmaktadır. Dolayısıyla anket çalışmalarında kayıp veri çözümlemeleri sırasında epeyce karşılaşılabilen bir yöntemdir.

Deck, ismini cevapsızlık durumunda elde edilmiş gözlemlerle karşılaştırmalı bilgisayar kartları destesinden almıştır (Little ve Rubin, 2002). Cold deck ve hot deck olarak ikiye ayrılabilir. Özellikle hot-deck (anlık oluşan deste) ile veri yüklemenin birçok farklı çalışması görülmektedir (Örneğin, ikili kovaryanslar üzerinden sıralı hot-deck, basit rasgele örnekleme ile hot-deck, düzeltilmiş hücreler üzerinden hot-deck gibi.).

Cold-deck (Eskiden oluşmuş deste) ve hot-deck yöntemlerini birbirinden ayıran fark ise cold-deck yönteminde veri yüklemesi için kullanılan deste, önceki çalışmalardan elde edilmiştir. Örneğin önceden yapılmış benzer bir çalışmanın veri

kümesi deste olarak alınabilir. Hot-deck yöntemi ise o anda yapılan araştırmada elde edilmiş gözlemler ve değişkenlerin yardımıyla kayıp veri probleminin çözümlenmesidir.

Enders (2010), hot-deck ile veri yükleme yönteminin birkaç çeşidinin olduğunu belirtmiş, bu yöntemi en basit haliyle diğer gözlem değerleriyle kayıp değerleri tamamlayabilmek olarak tanımlamıştır. Bu genel açıklama daha sonra daraltılarak hot-deck ile veri yükleme yöntemi sayesinde kayıp değerlerin daha etkili elde edilebileceği düşünülmüştür. Yani, her kayıp veri için elde edilen veri destesinden bilgi, rasgele de seçilebilir, düzeltilmiş olarak da seçilebilir ya da ikili ilişkiler üzerinden de seçilebilir. Bu şekilde yöntem araştırmaya uygun olarak değiştirilmeye elverişli hale gelmiştir. Hot-deck ile veri yükleme yönteminin çeşitlendirilmesi ile ilgili detaylı bilgilere Marker v. d. (2002) çalışmasından ulaşılabilir.

Deck yöntemleri kayıp veriyi elde etmeyi araştırmacıya bıraktığından tercih edilebilir olmasına rağmen götürüleri de vardır. Yanlış tercihlerin yanlış değerleri elde etmeye sebep olması önemli bir sıkıntıdır. Ayrıca Satıcı (2009), desteler için uygun büyüklüğün belirlenmesinin tartışma konusu olduğunu belirtmiştir.

2.8.3.5. En Yakın Komşu Yöntemi (Nearest Neighbor)

Hot-deck yöntemine benzerdir. Hatta bazı kaynaklarda hot-deck yönteminin farklı bir uygulaması olarak görülmektedir. En yakın komşu yöntemi, aslında bir sınıflandırma yöntemidir. Kayıp değerlerin çözümlenmesinde bu sınıflandırma mantığı kullanılmaktadır.

En genel haliyle gözlemler arasındaki birim uzaklıkları, kovaryans değerleri üzerinden elde edilir. Daha sonra kayıp verili gözlemin kayıp değeri, en yakın gözlemin elde edilmiş değeriyle yüklenir. Birimler arasındaki uzaklıklar için maksimum sapma, Mahalanobis gibi uzaklık yöntemleri kullanılabilir (Little ve Rubin, 2002). Uzaklık tanımında dikkatli olunmalıdır. Böylece katkıda bulunan değişkenlerin göreli önem düzeyleri uygun bir şekilde yansıtılmış olur. Elde edilmiş gözlemlerin tam kayıtlı olmasına gerek yoktur. Bazı değişkenler, yerine koymada ya da farkın hesaplanmasında göz ardı edilebilir veya bunların katkıları çok küçük olabilir (Satıcı, 2009).

2.8.3.6. Son Gözlemi İleri Taşıma (Last Observation Carried Forward)

Son gözlemi ileri taşıma yöntemi özellikle uzun süreli çalışmalarda karşılaşılabilecek yöntemdir. İsminden de anlaşılacağı gibi tekrarlı gözlemlerde kayıp olan değere bir önceki değeri yükleme ile veriyi tamamlar. Genellikle sağlık çalışmalarında ve klinik denemelerinde kullanılmaktadır. Bu yöntem, son elde edilen değeri ileriye taşıyarak belli aralıklardaki zaman içerisinde durumun hiç değişmediğini varsaymaktadır. Ama deneysel çalışmalar bu durumun geçerli olmadığını göstermiştir (Deneysel çalışmalar için, Cook v. d., 2004; Liu ve Gould, 2002; Mallinckrodt v. d., 2001). Özellikle kayıp veri oranının fazla olması çalışmayı olumsuz yönde etkileyecektir (Enders, 2010).

Son gözlemi ileri taşıma yönteminin tek getirisi uzun zamanlı çalışmalardaki kayıp gözlemleri elde edebilmenin en kolay yolu olmasıdır. Ancak TRK varsayımı altında bile son gözlemi ileri taşıma yönteminin yanlı sonuçlar vermesi, 1. tip hata oranını aşırı derecede arttırması ve uzun süreli çalışmaların yapısı gereği ölçülü ya da abartılmış sonuçlarla karşılaşılması önemli götürülerdir. (Molengberghs v. d., 2007; Baygül, 2007).

2.8.4. Model Tabanlı Yöntemler

Gözlenmiş değerler üzerinden bir model ve o model altında olabilirlik ve sonsal dağılımları elde ederek parametrelerin kestirildiği kayıp veri yöntemleridir. Bu yöntemlerin en önemli getirisi esneklikleridir. Geçici yöntemlerden kaçınarak model varsayımları altında çözümleme yaparlar. Ayrıca veri matrisinde gözlenememiş değerleri de hesaba katarak varyans kestirimleri verir. Model üzerinden hesaplamalar yapılması planlandığında birçok model söz konusu olabilir. Özellikle Lord (1955) çalışmasıyla başlayan ML kestirimleri, Dempster v. d. (1977) ile teorisi ortaya konulan EM algoritması, Rubin (1987b) çalışmasında önerilen çoklu veri yükleme (multiple imputation), model tabanlı yöntemlerdir.

2.8.4.1. ML ile Kayıp Veri Çözümlemesi

Kayıp veri çözümlemelerinde kitle parametrelerinin kestirilebilmesi için ML çok fazla önem ve ilgi gören bir yöntem olmuştur. Normal dağılıma sahip bir kitleden geldiği bilinen örneklemin, logaritması alınmış olabilirlik fonksiyonu aşağıdaki gibidir:

$$\log L = \sum_{i=1}^{n} \left\{ -\frac{k}{2}\ln(2\pi) - \frac{1}{2} ln|\Sigma| - \frac{1}{2}(y_i - \mu)^T \Sigma^{-1}(y_i - \mu) \right\} \quad (2.5)$$

Eş. 2.5'teki $(y_i - \mu)^T \Sigma^{-1}(y_i - \mu)$ ifadesi, ortalamalar arasındaki standart uzaklığı belirleyen Mahalanobis uzaklığıdır. Uzaklık değerinin sıfıra yakın olması gözlemin dağılımın merkezine yakın olduğunu açıklar. Verinin elde edildiği kitle normal dağıldığında ve parametre kestirimleri bilindiğinde en büyük log-olabilirlik değerine sahip olan değer kitlenin konum değerine en yakınıdır. Örneklemden kitle parametrelerinin kestirilmesi için örneklemin log-olabilirliği kullanılabilir. Log-olabilirliğin en büyük olduğu nokta verinin merkezini göstereceğinden açıklayıcı olacaktır. Örneklem log-olabilirliğinin ilk türevi sıfıra eşitlendiğinde fonksiyonun en büyük olduğu noktayı verir ki bu değer parametrenin konum kestirimine en yakın değerdir. ML kestirimleri için log-olabilirliğin kullanılarak elde edilebilecek ikinci kestirimler ise standart hatalardır. Standart hatalarla ilgili yorumlamalar örneklemin log-olabilirlik fonksiyonunun ikinci türevinden elde edilebilir.

Schafer ve Graham (2002), ML kestirimlerinin kayıp veri çözümlemeleri için çok önemli olduğunu, özellikle araştırmacıların en çok kullandıkları yöntemlerin ML kestiricileri yardımıyla oluşturulan yöntemler olduğundan bahsetmiştir. Bunun en büyük nedeni ise ML, RK veri düzeneğinde bile yansız kestiriciler vermektedir. Ayrıca TRK varsayımı altında geleneksel yöntemlerin hepsinden daha iyi sonuçlar vermektedir.

Enders (2010), ML'nin veri kümesindeki gözlenmiş değerleri kullanarak istatistiksel gücü en üst düzeye çıkardığını ve bu yüzden de etkili ve kullanıma elverişli olduğunu belirtmiştir. Kitlenin normal dağıldığı kabul edilirse kayıp veri durumunda i. gözlem için log-olabilirlik Eş. 2.6'daki gibidir:

$$\log L_i = \sum_{i=1}^{n} \left\{ -\frac{k_i}{2}\ln(2\pi) - \frac{1}{2} ln|\Sigma_i| - \frac{1}{2}(y_i - \mu_i)^T \Sigma_i^{-1}(y_i - \mu_i) \right\} \quad (2.6)$$

Tam veri durumunda elde edilen Eş. 2.5'teki log-olabilirlik fonksiyonundan tek farkının i indisinin bulunmasıdır. i. indis, i. gözlem için log-olabilirliğin sadece değişkenlere ve elde edilmiş gözlemlerin parametrelerine bağlı olduğunu göstermektedir. Yani, i. gözlemin herhangi bir değişkene ait değeri gözlenmediği durumda, log-olabilirlik değerleri kayıp veriye ait değişkenin bulunmadığı

parametre kestirimleri üzerinden hesaplanacaktır.

Özellikle tek ve iki değişkenli örnekler için gösterilmesi ve anlaşılması kolay olmasına rağmen çok değişkenli durumlarda ML kestirimleri daha karışıktır ve genellikle parametre değerlerini kümesini tanımlayabilmek için iteratif algoritmalar kullanılır. ML'de iteratif algoritmanın amacı, log-olabilirlik fonksiyonunun tepe noktasına hızlıca çıkabilmektir. Parametre kestirimleri için öncelikle bir başlangıç değeri elde edilmelidir. Bu değerler elde edildikten sonra geriye tepe noktasına ulaşmak için zincirsel adımları kurabilmek kalır. Her adım optimizasyon sürecinin bir iterasyonudur. İlk adımda algoritma olasılık fonksiyonundaki başlangıç değerleriyle değişir ve log-olabilirlik hesaplanır. Her adımda amaç, log-olabilirlik değerler yönünde parametre değerlerini düzenleyebilmektir. Farklı algoritmalar, adımlarında farklı yöntemleri kullanabilir. Log-olabilirlik fonksiyonun tepe noktasına ulaşmak için elde edilmiş algoritma adımları ilk birkaç adımda hızlıca değişkenlik gösterirken belli bir yerden sonra değişkenliği azalacaktır. Bu değişkenlikler arasındaki farklar çok azaldığında yakınsaklık sağlanır ve ML kestirimlerine yaklaşılır (Enders, 2010). Bu konu ile ilgili daha detaylı bilgilere, Schafer (1997), Little ve Rubin (2002) çalışmalarından, örnekli açıklamalarsa Enders (2010) çalışmasından elde edilebilir.

2.8.4.2. EM Algoritması

EM algoritması, tamamlanmamış veri problemlerinde parametreleri hesaplamak için ML kestirimini içeren yinelemeli bir yöntemdir. EM algoritması, elde edilmiş değerlere ve kayıp veriye göre oluşturulan olabilirlik fonksiyonundan yola çıkar (Little ve Rubin, 2002). Genel olarak EM algoritması, kestirilmiş değerleri kayıp değerlerin yerine koyar, kayıp veriyi kullanarak parametre kestirir ve daha uygun parametreyi buluncaya dek algoritmayı yineler. EM algoritması iki adımdan oluşmaktadır:

E-adımı (Beklenti adımı): Tam veri log-olabilirlik fonksiyonunun koşullu dağılımının beklenen değerinden hesaplanır. Parametrelere başlangıç değeri yüklenmesi bu adımda yapılmaktadır.

M-adımı (En büyükleme adımı): Kestirilen kayıp veri değerini kabul ederek oluşan tam veri modeli üzerinden ML kestirimi hesaplanmaktadır. M-adımı

sonucunda ortaya çıkan kestirimler, EM algoritmasının çıktısını oluşturmaktadır (Dempster v.d., 1977). İteratif yöntem bir E bir de M adımından oluşmaktadır. M adımında bir önceki iterasyondaki M adımı kestiriminden daha büyük log-olabilirlik değeri elde edilmektedir. Log-olabilirlik değerinin en yüksek noktaya çıktığı düşünülen ve kitle parametrelerinin kestirimlerinin arasındaki farkın önemsizleştiği durumda iteratif çözümleme biter. M adımında elde edilen son parametre kestirimleri geçerli değerlerdir.

EM algoritması kayıp değerlere uygun değerler yüklemek için kullanılabilmesine rağmen daha çok kitle parametreleri için kestirimler elde ederken kullanılır. Verinin normal dağıldığı varsayımı altında E adımında bulunan yeterli istatistiklerin beklenen değerleri ile M adımında parametrelerin ML kestirimlerinin elde edilmesi kolaydır. Verinin normal dağılmadığı durumlarda ise özel kestirim yöntemleri kullanılarak sorun giderilebilir ancak EM algoritması ile ilgili paket yazılımlar genellikle normallik varsayımına dayanır (Atalay, 2003).

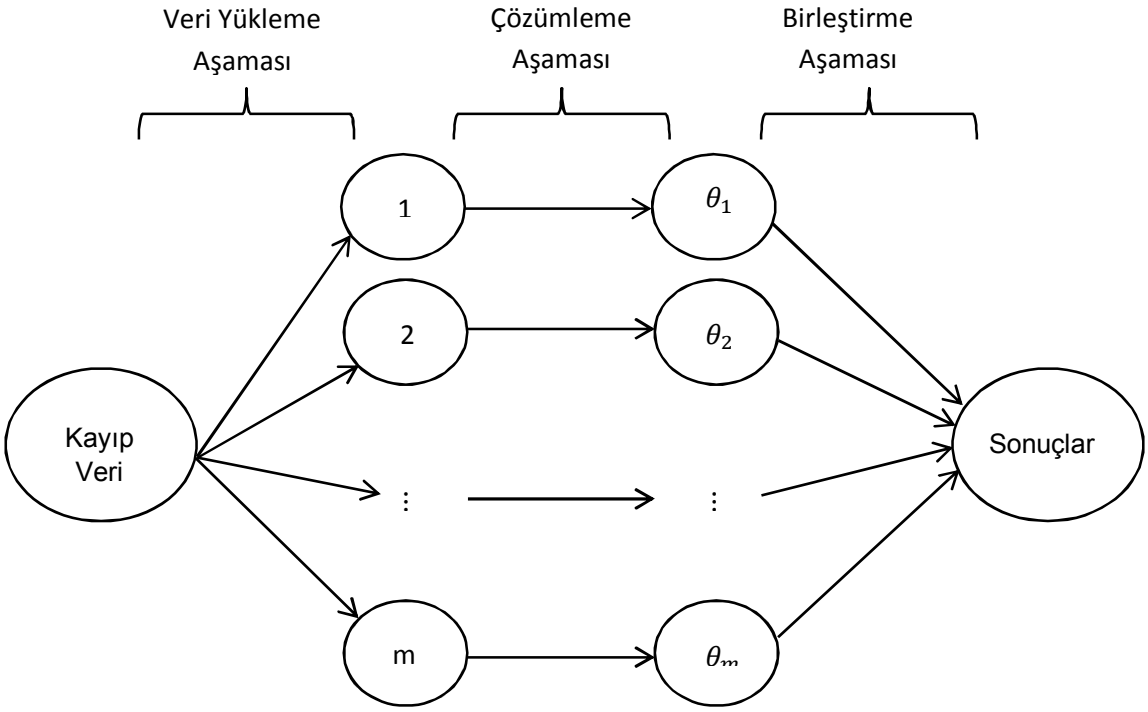
EM algoritması, karmaşıklıktan dolayı ML kestiriminin doğrudan hesaplanamadığı durumlarda uygulanmaktadır. Ayrıca doğrudan gözlenemeyen ve $y_{göz}$ ile ilişkisi olan Y'lerin gözlenen değerlere bağlı kayıp gözlemlerin (y_{kay}) tamamlanması temeline dayanmaktadır (Prescher, 2003; Yazıcı, 2005). EM algoritması ve uzantıları birçok kez geliştirilmiş ve hemen hemen her veri yapısında kullanılabilir uygunluktadır. EM algoritmasının yöntemi ve işleyişi ile ilgili daha detaylı bilgiye Yazıcı (2005) ve McLaachlan ve Krishnan (1997) çalışmalarından ulaşılabilir.

2.8.4.3. Çoklu Veri Yükleme (Multiple Imputation)

Çoklu veri yükleme yöntemi başlıca üç adımdan oluşmaktadır. Aşamaların genel gösterimi Şekil 2.12'deki gibidir.

Veri yükleme aşaması (imputation phase) , m sayıda veri kümesini kopyalar ve her birisi kayıp veri yöntemleri ile kestirilir. Burada, stokastik regresyon ile veri yükleme yönteminin iteratif bir versiyonu uygulanır. Ama matematiksel altyapısı Bayesci yöntemlere dayanmaktadır. Çözümleme aşaması, istatistiksel yöntemlerle m sayıdaki tam küme için çözümlemeler yapılmasıdır. Sonuç olarak parametre kestirimlerinin elde edildiği m küme elde edilir. Son aşama ise birleştirme aşaması (pooling phase) olarak adlandırılır ve sonuçların hepsi bu aşamada tek bir küme

olarak birleştirilir. Rubin (1987a) parametre kestirimleri ve standart hataların birleştirilmesi için basit bir eşitlik vermiştir. Örneğin ortalama kestirimi, birleştirilen parametre kestirimlerinin ortalamasının alınmasıyla; standart hatalar ise biraz daha karışık olsa da aynı mantıkla elde edilmektedir.



Şekil 2.12. Çoklu Veri Yükleme Aşamaları

Birçok veri yüklemesi yapılmış küme oluşturulmasının, kestirim çözümlemelerinin yapılmasının ve birleştirilmesinin zaman kaybına neden olabileceği düşünülebilir. Ancak, çözümlemelerin paket programlarda yapılabiliyor olması zaman kaybını en aza indirgemiştir. Büyük örneklemlerde çoklu veri yükleme yönteminin sonuçları ML kestirimi kullanan yöntemlerin sonuçlarıyla asimptotik olarak benzerdir.

2.8.4.4. Bayesci Veri Yükleme Yöntemleri

Bayesci çıkarım ile çeşitli yöntemler birleştirilerek de kayıp veri yöntemleri elde edilmiştir. Little ve Rubin (2002), kayıp veri düzeneğini ihmal edilebilir olması durumunda Bayesci çıkarımlar kullanılabileceğini belirtmiştir. Kayıp veri düzeneğinin ihmal edilmesi için düzenek RK olmalı ve Eş. 2.7'de gösterildiği gibi θ ve ψ parametreleri önsel olarak bağımsız olmalıdır. Burada ψ, kayıp veri düzeneğinin dağılımı için bilinmeyen parametredir.

$$p(\theta, \psi) = p(\theta)p(\psi) \quad (2.7)$$

Bayesci çıkarımlar genellikle iteratif benzetim yöntemlerinde kullanılmaktadır. Veri arttırma (data augmentation) yöntemi örneklem büyüklüğünün küçük olduğu durumlarda kullanışlı olmaktadır.

Tanner ve Wong (1987) tarafından önerilen yöntemde EM algoritmasının ve çoklu veri yüklemenin özellikleri birleştirilmiştir. I ve P adımlarından oluşmaktadır (Imputation ve Posterior). Enders (2010), Bayesci çıkarımda, kayıp verinin, elde edilmiş değerlere bağlı olan koşullu dağılımdan, I adımı için oluşturulan ortalama vektörü ve kovaryans kestiriminden yüklendiklerini açıklamıştır. P adımı ise bilinen sonsal dağılımların elde edildiği Bayesci çözümlemeler ile tamamlanmaktadır.

Gibbs Örneklemesi, kayıp verinin yapısal şeklinin genel kayıp olması durumunda değişkenlerin birleşik fonksiyonundan yararlanarak kayıp veri çözümlemelerinde kullanılabilir. Bu yöntem, ML kestirimleri için koşullu beklenti en büyükleme (Expectation - Conditional Maximization-ECM) algoritmasına benzer olmakla birlikte adımları değişkenlerden oluştuğundan anlaşılması ECM algoritmasına göre daha basittir (Little ve Rubin, 2002).

2.8.4.5. Sağlam Veri Yükleme Yöntemi (Robust Imputation)

Kayıp veri çözümlemesi yapılırken aykırı değerin etkilerini en aza indirgemek gerekmektedir. Ardışık veri yükleme (sequential imputation), kovaryans ölçütü en küçük olacak şekilde veri yükleyen bir yöntemdir (Verboven v. d., 2007). Bu yöntem sağlam bir yöntem değildir ancak ölçütlerde değişiklik yapılarak sağlam bir yönteme dönüştürülebilir. Bu tezde ardışık veri yükleme yöntemi aktarılacak ve yöntemin sağlam veri yükleme yöntemi olması için yapılacak değişiklikler gösterilecektir.

k indisi kayıp değeri; g indisi gözlenen değeri ifade ederken veri matrisinin herhangi bir gözlemi, $x_i = \begin{bmatrix} x_k^T & x_g^T \end{bmatrix}^T$ şeklinde yazılabilir. Veri kümesinde herhangi bir değeri kayıp olmayan gözlemde bulunabileceğinden t tam-veri yapısında olan gözlemleri göstermek üzere X matrisi, elde edilmiş gözlemler ve kayıp gözlemler olarak $X = [X_t^T \quad X_k^T]^T$ şeklinde ayrılabilir. Bu durumda, $t \times p$ boyutlu kısım tamamen elde edilmiş gözlemlerden oluşurken $(n - t) \times p$ boyutlu kısım X matrisinin kayıp gözlemli parçasını oluşturur.

Ardışık veri yükleme yönteminde kullanılan kavramlar kovaryans matrisi ve determinanttır. $\bar{X}$ sütun vektörü, ortalamayı gösterdiğinde $n \times p$ boyutlu örneklemin kovaryans matrisi,

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T \tag{2.8}$$

şeklinde ifade edilir. $(p \times p)$ boyutlu karesel determinant matrisi S $(i = 1,2, \dots, p)$'yi hesaplamak için Laplace formülü kullanılır. $M_{i,j}$, i. satır ve j. sütun çıkartıldıktan sonra geriye kalan satır ve sütunlar üzerinden yapılan determinant işlemi olduğunda S'nin determinantı,

$$\det(S) = \sum_{i=1}^{p}\sum_{j=1}^{p} S_{i,j}(-1)^{i+j}M_{i,j} \tag{2.9}$$

şeklindedir. Burada, $S_{i,j}$ ise i. satır ve j. sütun elemanlarıdır.

Bu bilgiler sayesinde kayıp veri için ardışık yükleme yöntemi uygulanabilir. Kayıp verisi bulunmayan gözlemlerin oluşturduğu küme X_t şeklinde bir matristir. En az kaybı olan gözlem (x^*) elde edilerek çözümleme başlar. $\bar{x}_t = \sum_{i=1}^{t} x_i/t$ ifadesi $X^* = [X_t^T \quad x^*]^T$'nun satır ortalamalarının kestirimi olarak kullanılırsa, $x^* = \begin{bmatrix} x_k^{*T} & x_g^{*T} \end{bmatrix}^T$ şeklinde ifade edilebilen gözlemin kayıp değerleri, X^*'ın kovaryans determinantını en küçük yapacak şekilde yüklenir. x_k^* için determinantın minimize edilmesini sağlayan eşitlik Eş. 2.10'daki gibidir (Wynn,1970):

$$D(x^*) = (x^* - \bar{x}_t)^T\big(cov(X_t)\big)^{-1}(x^* - \bar{x}_t) \tag{2.10}$$

Eğer Eş. 2.10'daki $(cov(X_t))^{-1}$ matrisi C olarak gösterilirse,

$$C = \begin{bmatrix} C_{k,k} & C_{g,k} \\ C_{k,g} & C_{g,g} \end{bmatrix} \tag{2.11}$$

şeklinde yazılabilir. Buradan,

$$\frac{\partial D(x^*)}{\partial x_k^*} = 0 \tag{2.12}$$

minimizasyon probleminin çözümü,

$$x_k^* = (\bar{x}_t)_k - \left(C_{k,k}\right)^{-1} C_{k,g}\left(x_g^* - (\bar{x}_t)_g\right) \tag{2.13}$$

şeklindedir. Eş. 2.12'deki x^*'ın kayıp kısmı, $((\bar{x}_t)_k^T)$ ve gözlenen kısmı, $((\bar{x}_t)_g^T)$ olmak üzere $\bar{x}_t = \left[(\bar{x}_t)_k^T \quad (\bar{x}_t)_g^T\right]^T$'dur. C matrisi için $C_{k,k}$, x^*'ın kayıp değişkenlerine ait kısımdır. $C_{k,g}$, satırlarda x^*'ın kayıp değişkenlerini sütunlarda ise x^*'ın gözlenmiş değişkenlerini gösterir ve C'nin bir parçasıdır. x^*'ın kayıpları kestirildiğinde tam veri kümesi $X_t = [X_t^T \quad x^*]^T$ şeklinde güncellenir ve en az kaybı olan yeni bir x^* ile adımlar tekrarlanır.

Sağlam veri yükleme yönteminde ise ardışık veri yükleme yöntemindeki ortalama ve kovaryans matrisi yerine sağlam kestirimler tercih edilir. Sağlam kestirimler için X_t üzerinden aykırılık ölçütü (outlyingness measure) tanımlanır. x_i gözleminin aykırılık ölçütü Eş. 2.14'teki gibidir (Stahel, 1981; Donoho, 1982):

$$aykırılık_i = maks_{v\epsilon B}\left(\frac{\left|x_i^T v - ortanca(x_j^T v)\right|}{ort.mut.sapma(x_j^T v)}\right) \tag{2.14}$$

Burada B, sıfır olmayan vektörleri içerir. $ortanca\left(x_j^T v\right)$, $v\epsilon B$ için ortanca değeridir. $ort.mut.sapma\left(x_j^T v\right) = ortanca(\left|x_l^T v - ortanca(x_j^T v)\right|)$'dir. Sağlam veri yükleme yöntemi için α aykırı değerlerin oranı olmak üzere, sağlam veri yükleme yöntemi için başlangıç noktası, X_t tam veri kümesinin $(1-\alpha)$ oranındaki en küçük aykırılık ölçüsüne sahip olan gözlemlerin sağlam ortalama ve sağlam kovaryans matrisinden elde edilir. X_t'nin ortalama ve kovaryansı için sağlam kestirimler oluşturulduktan sonra en az kayıp veriye sahip x^* yukarıda açıklanan ardışık veri yükleme yöntemi kullanılarak çözümlenir.

Gözlemin kayıp değerleri yüklendikten sonra aykırılığı Eş. 2.14'ten belirlenebilir. Ancak bu yöntemde B kümesi başlangıçtan itibaren değiştirilmemektedir. Dolayısıyla ortanca ve en küçük mutlak sapma sayesinde, kayıp değerler Eş. 2.13 kullanılarak ardışık veri yükleme yöntemi ile yüklenir ve Eş. 2.14'ten x^* için aykırılık ölçüsü belirlenir. Eğer aykırılık ölçüsü, belirlenmiş olan aykırılık değerini geçerse çözümlemelere dâhil edilip edilmeyeceğine karar verilebilir (Branden ve Verboven, 2009).

2.8.5. Diğer Veri Yükleme Yöntemleri

Yukarıdaki bölümlerde kayıp veri için yapılan başlıca ayrımlar açıklanmıştır. Kayıp veri çözümlemeleri, özellikle kayıp veri yapısına göre uygulanması gerektiğinden aralarında yapılacak kıyaslamada en iyi sonuç verecek yöntemden bahsetmek zordur.

Atalay (2005), evrimsel (evolutionary) algoritmalar ve sinir ağları (neural networks) için yapılan uygulamaların veri yükleme için çözümlemeler getirebileceğini aktarmıştır. Ağaç tabanlı (tree-based) ve entropi tabanlı (entropy-based) yöntemler, özellikle çalışma alanı olan araştırmacılar tarafından veri yükleme yöntemleri olarak önerilmektedir.

Kayıp veri çözümlemelerinde yöntemlerin fazla olması veri yapılarındaki farklılıklardan kaynaklanmaktadır. Bu farklılıklar olduğu sürece çeşitli yöntemlerin birleştirilmesiyle ve önerilmiş yöntemlerin değiştirilmesiyle kayıp veri sorunu aşılmaya çalışılacaktır.

2.8.6. Veri Yükleme Yöntemleri İçin Ölçütler

Gerçek değerlerin bilinmesi durumunda kullanılabilecek ölçütler, benzetim çalışmalarında sıklıkla kullanılmaktadırlar. Veri yükleme yöntemlerinin performanslarını ölçmek için, ortalama veri yükleme hatası (OVYH), ortalama mutlak sapma (OMS) ve ortalama mutlak göreli sapma (OMGS) ölçütleri kullanılabilir.

n gözlem ve $(n-m)$ kayıp veri için y_i gerçek değeri, y_i^* ise i. kayıp veri için yükleme yapılan değeri gösterdiği düşünüldüğünde ortalama veri yükleme hatası,

$$OVYH = \sqrt{\frac{1}{(n-m)} \sum_{i=1}^{(n-m)} (y_i^* - y_i)^2} \tag{2.15}$$

biçimindedir. Ortalama mutlak sapma,

$$OMS = \frac{1}{(n-m)} \sum_{i=1}^{(n-m)} |y_i^* - y_i| \tag{2.16}$$

ile verilir. Ortalama mutlak göreli sapma ise,

$$OMGS = \frac{1}{(n-m)} \sum_{i=1}^{(n-m)} |(y_i^* - y_i)/y_i| \qquad (2.17)$$

ile ifade edilir (Atalay, 2005).

2.8.7. Veri Yükleme Yazılım Paket Programları

Veri yükleme yöntemleri için "Texas Üniversitesi Kayıp ve Tamamlanmamış Veri konusundaki Yardım sayfasından (General FAQ#25,2003)" elde edilmiş bazı programlar Atalay (2005)'da verilmiştir. Çizelge 2.5'de programlar, çözümleme varsayımları ve yöntemleri verilmiştir.

Çizelge 2.5. Veri Yükleme Yazılım Paket Programları

Yazılımlar	Veri Yükleme Yöntemleri	Varsayımlar
Amelia	Çoklu Veri Yükleme	TRK
SAS Temel -Proc Standart	Ortalama ile Veri Yükleme	TRK
SAS/IML Çoklu Veri Yükleme	Çoklu Veri Yükleme	RK
Paul Allison SAS Çoklu Veri Yükleme Makrosu	Çoklu Veri Yükleme	RK
SAS EM_COVAR.SAS EM Programı	Kovaryans Matrisinde Boopstrap ile EM Algortiması	RK
SPSS Temel	Ortalama ile Veri Yükleme	TRK
SPSS "Missing Value Modülü"	EM Algoritması	RK
AMOS	ML Kestirimi	RK
SOLAS	Çoklu Veri Yükleme, Regresyon ile Veri Yükleme	TRK ve RK
SAS/IML Örüntü Karışımı Model Programları	Çoklu Veri Yükleme	RK

Şimdiye kadar verilen bölümlerde, veri yükleme yöntemleri, kayıp veri düzenekleri, kayıp veri yapıları açıklanmıştır. Bu çalışmanın amacı sağlam yöntemlerin kayıp veri çözümlemelerindeki kullanımıdır. Ayrıca, aykırı değerlerin, klasik kestiriciler kullanıldığında yüklenen kayıp verinin değerlerini ve kitle parametrelerini nasıl etkilediğini incelemektir. Bu nedenle bundan sonraki bölümde sağlam istatistik, sağlam kestirim yöntemleri özetlenecektir.

ÜÇÜNCÜ BÖLÜM

3. Sağlam Yaklaşım ve Kayıp Veri Durumunda Sağlam (Robust) Kestirim

İstatistiksel çalışmalarda kayıp veri probleminin nasıl aşılacağı bir önceki bölümde incelendi. Veri kümelerinde kayıp veri gibi karşımıza çıkacak diğer problem de aykırı değerlerdir. Bu problemin aşılabilmesi için önerilen sağlam yaklaşım, aykırı değer sorununa ve modeldeki bozulmalara cevap verir. Bu bölümde sağlam yaklaşım ve sağlam yöntemler üzerinde kısa bir bilgi vererek kayıp veri durumunda sağlam kestirim konusu aktarılacaktır.

3.1. Sağlam Yaklaşım (Robust Approach) ve Özellikleri

Uygulamalarda klasik istatistiksel yöntemlerin kullanılabilmesi için çeşitli varsayımlar vardır. Rasgelelik, bağımsızlık, normallik gibi varsayımların sağlanamaması durumunda parametrik dağılımda kısmi değişmeler, veri kümesinin çoğunluğuna uymayan aykırı değerler ortaya çıkabilir. Sağlam yaklaşım, aykırı değerlerde ve parametrik dağılımdaki sapmalarda klasik yönteme bir alternatif olarak ortaya çıkmıştır. Sağlam sözcüğünü istatistik literatürüne kazandıran Box (1953), sağlamlığı varsayımlardaki sapmalara karşı sağlam sonuçlar veren istatistiksel yöntemler için kullanmıştır. Huber (1964) ve Hampel (1971) sağlamlık ile ilgili temel çalışmaları vermişlerdir. Huber (1981), incelenen bir dağılım biçiminin varsayılan modelden küçük sapmalar göstermesini dağılımsal sağlamlık (distrubitional robustness); varsayımlardan küçük sapmaların kestirimleri etkilemediği durumu ise kestiricilerin sağlamlığı olarak açıklamıştır (Candan, 1995).

Hampel v. d. (1986), sağlam yaklaşımın kullanılması için parametrik modelden sapmaların, büyük hataların ortaya çıkmasından kaynaklanan sapmaların, yuvarlama ve gruplama sonucunda ortaya çıkan sapmaların, modelin yaklaşık kestiriminden dolayı ortaya çıkan sapmaların, dağılımsal ve bağımsızlık varsayımlarının yaklaşık olarak sağlanmamasından kaynaklanan hataların olması gerektiğini açıklamıştır.

Verinin büyük çoğunluğunu en iyi biçimde temsil edecek yapıyı belirlemek, veri yapısındaki sapmaları ya da aykırı değerleri belirlemek, etkili gözlemler ile ilgili uyarıları vermek ve bu noktaları saptamak sağlam yöntemin amaçlarıdır (Hampel,1974; Candan,1995).

Hampel (1974), sağlam bir yöntemin bulundurması gereken özellikleri açıklamıştır. Sağlam bir yöntem, varsayılan modelde iyi etkinliğe sahip olmalı, modeldeki küçük sapmalar performanslara çok az zarar vermelidir. Büyük bozulum varlığında bile güvenilir olmalıdır, aykırı veriyi tespit edebilmelidir. Sağlam yöntemler, verinin büyük çoğunluğunu iyi bir şekilde temsil ederek olabildiğince küçük varyanslara sahip olmalıdır.

Bir kestiricinin sağlamlığını ve dirençliliğini değerlendirmede bazı sağlamlık ölçütleri kullanılmaktadır. Etki fonksiyonu, kestiricinin herhangi bir noktadaki küçük bozulum miktarına nasıl karşılık vereceği hakkında kesin bir bilgi verir. Dolayısıyla belirli bir dağılım biçimine duyarlı olan kestiriciler, küçük miktardaki bozulumlardan etkilenecektir. Duyarlılık eğrisi, eklenen gözlemin kestirici üzerindeki etkisini gösterir ve bu nedenle örneklemde oluşabilecek büyük bir hatanın etkisini tanımlamak için sıkça kullanılır. Büyük hata duyarlılığı, tek bir bozulmuş gözlemin kestirici üzerindeki asimptotik olarak en büyük etkisini ölçer. Yerel kayma duyarlılığı, gözlemlerdeki küçük değişimlerin kestirici üzerindeki ölçülebilir etkisini ortaya çıkartır. Reddetme noktası, bir kestiricinin aykırı değerleri reddedip reddetmediğini, reddediyorsa hangi uzaklıkta reddedileceğini veren bir ölçüttür.

Bozulma noktası (breakdown point), kestiricinin aykırı değere karşı direncinin olduğu sınırı vermektedir. Daha tanımsal şekliyle bozulma noktası, örneklem oranı kısıtsız değiştirildiğinde, kestiricideki değişime bir sınır koymak için gözlemlerin en büyük olası oranı olarak açıklanmıştır.

3.2. Sağlam Kestiriciler

Bu bölümde sağlam kestiricilerin genel tanımlamaları verilerek uygulamada kullanılan kestiriciler detaylı olarak açıklanacaktır.

Konum kestiricisi olarak literatürde L, R ve M kestiricileri önemli yer tutmaktadır. L-kestiriciler sıralı istatistiklerin doğrusal birleşimleridir. R-kestiriciler iki örneklem arasındaki sıra (rank) testlerinden türetilmişlerdir. M-kestiriciler, en çok olabilirlik türü kestiricilere karşılık gelir. İşlerlik, uygunluk açısından M kestiricileri daha üstündür ve çok parametreli problemlere genelleştirilebilir (Andrews v.d., 1972; Candan, 1995).

M kestiricileri, belirli bir amaç fonksiyonunu minimize etmek için kullanılır ve amaç fonksiyonundaki artıklarla ilgilenir. Huber'in M kestiricisi, Hampel'in yeniden azalan kestiricisi, Andrews'in dalga fonksiyonu (Andrew's wave), Tukey'in bisquare kestiricisi bazı örnekleridir. M-kestiricileri iteratif olarak çözülür.

S-kestiricileri, günümüzde gittikçe artan kullanım alanına sahiptir. Çok değişkenli durumlarda meydana gelen yüksek bozulma noktaları için önerilmektedir.

3.3. Çalışmada Kullanılan Sağlam Kestiriciler

Tezin uygulama bölümünde, sağlam veri yükleme yöntemi uygulanmış ve yüksek bozulma noktasına sahip kestiricilerden Minimum Kovaryans Determinant (MCD-Minimum Covariance Determinant) kestiricisi, Ortogonolleştirilmiş Gnanadesikan–Kettenring (OGK-Orthogonalized Gnanadesikan–Kettenring) kestiricisi, Stahel–Donoho (SD) kestiricisi, bisquare-S (BS) kestiricisi sağlam kestiriciler olarak, yükleme yapılan veri kümelerine uygulanmıştır.

3.3.1 Minimum Kovaryans Determinant (MCD) Kestiricisi

En Küçük Kesilmiş Kareler (LTS- Least Trimmed Square) kestiricisi, artıkların kareleri $e_{(1)}^2 \leq e_{(2)}^2 \leq \cdots \leq e_{(n)}^2$ şeklinde en küçükten en büyüğe sıralanıp ilk h artık durumuna sahip gözlemleri kullanılarak elde edilir. Buradaki h değeri örneklemin gözlem sayısına ve parametre sayısına bağlı olacak şekilde doğru seçilmelidir. Rousseeuw ve Leroy (1987)'un, h değeri için yaptıkları öneri, $\lfloor . \rfloor$ gözlem sayısının tam kısmını, α da kesilme miktarını göstermek üzere Eş. 3.1'deki gibidir:

$$h = \lfloor n(1-\alpha) \rfloor + 1 \tag{3.1}$$

MCD ise çok değişkenli konum kestirimi için LTS yönteminin genelleştirilmiş halidir. MCD konum kestirimi, kovaryans matrisinin determinantını en küçükleyen h tane gözlemin ortalamasıdır. Çalışmada MCD'nin ortalama ve kovaryans matrisi kestiricileri olarak kullanılan T, Eş. 3.2'de, C ise Eş. 3.3'te verildiği gibidir:

$$T_{MCD} = \frac{1}{h}\sum_{i=1}^{h} x_i^{(j)} \tag{3.2}$$

$$C_{MCD} = k_{tdf} k_{ködf} \frac{1}{h-1} \sum_{i=1}^{h} (x_i^{(j)} - T_{MCD})(x_i^{(j)} - T_{MCD})^T \quad (3.3)$$

Eş. 3.3'teki k_{tdf} ve $k_{ködf}$ değerleri sırasıyla tutarlılık düzeltme faktörü (Detaylı bilgi için Butler v. d., 1993) ve küçük örneklem düzeltme faktörü (Detaylı bilgi için Pison v. d., 2002) olarak açıklanır. MCD kestiricisi C kovaryans matrisi olmak üzere Eş. 3.4'ü sağladığı için uyum eşdeğişkenlik (affine equivariant) özelliğine sahiptir.

$$\det(A^t CA) = (\det(A))^2 \det(C) \quad (3.4)$$

Uyum eşdeğişkenlik, değişkenlere uygulanan doğrusal bir dönüşümün kestiriciye de aynı şekilde dönüşebilmesidir. Uyum eşdeğişkenlik özelliğini sağlayabilmek için A tekil olmayan matris ve b herhangi bir vektör olmak üzere Eş. 3.5 elde edilebilmelidir:

$$T(x_1 A + b, x_2 A + b, \dots, x_n A + b) = T(x_1, x_2, \dots, x_n) A + b \quad (3.5)$$

MCD kestiricisinin örneklem büyüklüğü sonsuza yaklaştıkça bozulma noktası %50'dir. MCD kestiricisi için en küçük kovaryansa sahip h gözleme klasik kovaryans kestirimi uygulanmaktadır. Bozulmanın olmadığı durumlarda MCD kestiricisi aritmetik ortalamaya yakınlaşır (Rousseeuw ve Leroy, 1987).

3.3.2. Ortogonalleşmiş Gnanadesikan–Kettenring (OGK) (Orthogonalized Gnanadesikan–Kettenring) Kestiricisi

MCD ve diğer uyum eşdeğişken özellikli yüksek bozulma noktasına sahip kestiriciler yüksek boyutlarda konveks olmayan optimizasyon problemlerinin çözümüdür ve ciddi hesaplamalar gerektirir. Birçok yüksek bozulma noktalı kestiriciler uyum eşdeğişkenlik özellikleri sayesinde hızlı çözümlemelerle elde edilmektedir. Gnanadesikan ve Kettenring (1972) tarafından önerilen iki değişkenli kovaryans kestiricileri temel alınarak elde edilen OGK kestiricisi Maronna ve Zamar (2002) tarafından önerilmiştir. Y_j ve Y_k raslantı değişken çiftleri için σ, standart sapma fonksiyonu olarak düşünülmüş ve S_{jk} Eş. 3.6'daki gibi önerilmiştir:

$$S_{jk} = \frac{1}{4}\left(\sigma\left(\frac{Y_j}{\sigma(Y_j)} + \frac{Y_k}{\sigma(Y_k)} \right)^2 - \sigma\left(\frac{Y_j}{\sigma(Y_j)} - \frac{Y_k}{\sigma(Y_k)} \right)^2 \right) \quad (3.6)$$

Eş. 3.6'daki $\sigma()$ için sağlam bir yöntem kullanılır ve $j = 1,2, \dots, p$ ve $k = 1,2, \dots, p$ için her bir S_{jk} hesaplanırsa kovaryans matrisinin sağlam kestirimi elde edilmiş olur.

OGK kestiriminin önemli bir götürüsü simetrik ve pozitif tanımlı kovaryans matrisi olmaması ve uyum eşdeğişken olmamasıdır. Maronna ve Zamar (2002), aşağıdaki adımlarla bu sorunu aşmışlardır:

- $X = \{x_1, x_2, \dots x_n\}$ veri kümesinin sütunları X_l $(l = 1,2, \dots, p)$ olmak üzere $y_i = D^{-1}x_i$ $(i = 1,2, \dots, n)$ şeklinde tanımlanmıştır. $D = köşegen\left(\sigma(X_1),\right.$ $\sigma(X_2),$ … , $\left.\sigma\left(X_p\right)\right)$ olduğu için $Y = \{y_1, y_2, \dots, y_n\}$ normalleştirilmiş veri kümesidir.
- Dönüşüm uygulanmış $Y = \{y_1, y_2, \dots y_n\}$ veri kümesinin sütunları Y_l $(l = 1,2, \dots, p)$ olmak üzere $u_{jk} = S(Y_j, Y_k)$ değerleri $j \neq k$ olmak üzere Eş. 3.6'dan sağlam kovaryans kestirimleriyle hesaplanır. $j = k$ için kovaryans matrisinin değerleri 1 olur.
- Y, Temel Bileşenler Çözümlemesi yardımıyla $U = E\Lambda E^T$ ayrışımı yapılır. Burada oluşturulan Λ matrisi, U'nun özdeğerleri λ_j'lerden oluşan köşegen matristir. E ise U'nun sütunlardaki özvektörler e_j'lerden oluşur.
- Adımlarda elde edilen matrisler kullanılarak $z_i(E^T y_i) = E^T D^{-1} z_i$ ve $A = DE$ şeklinde tanımlamalar yapılır. Bu durumda OGK kestiricileri sırasıyla Eş. 3.7 ve Eş. 3.8'deki gibidir. Eş. 3.7'deki $m = m(z_i) = (m(Z_1), m(Z_2), \dots, m\left(Z_p\right))$ olan sağlam ortalama fonksiyonudur. Eş. 3.8'deki Γ, dönüştürülmüş değerler olan z_j'lerin varyans kestirimi $\sigma\left(z_j\right)^2$ olmak üzere $köşegen(\sigma\left(z_j\right)^2)$ ile tanımlanır $(j = 1,2, \dots, p)$.

$$T_{OGK} = Am \tag{3.7}$$

$$C_{OGK} = A\Gamma A^T \tag{3.8}$$

Eş. 3.7 ve Eş. 3.8, $Z = \{z_1, z_2, \dots z_n\}$ dönüşüm uygulanmış değerlerinin OGK kestiricileri olduğundan dönüşümlerle gerçek koordinatlarına geçilmelidir (Todorov ve Filzmoser, 2009).

3.3.3.Stahel-Donoho (SD) Kestiricisi

Stahel (1981) ve Donoho (1982) birbirilerinden bağımsız şekilde yapmış oldukları çalışma sonucunda yüksek bozulma notasına sahip ilk uyum eşdeğişken kestiriciyi bulmuşlardır. Kestirici, uzaklıkların ağırlıklı ortalaması olarak tanımlanmıştır. Her x_i gözlemini daha açık hale getirebilmek için tek boyutlu izdüşümleri araştırılmaktadır. Bu durumda x_i'nin uzaklık ölçümü Eş. 3.9'daki gibidir ve eşitlikteki $ortanca_j(x_j v^t)$, v vektörü yönünde x_j noktalarının izdüşümünü ve payda ise bu izdüşümlerinin ortanca mutlak sapmasını gösterir:

$$u_i = maksimum_{\|v\|=1} \frac{\left|x_i v^t - ortanca_j\left(x_j v^t\right)\right|}{ortanca_k\left|x_k v^t - ortanca_j\left(x_j v^t\right)\right|} \tag{3.9}$$

u_i değerini hesaplamak için bütün mümkün boyutlar araştırılmalıdır. Boyutlar üzerinden konum kestiricisi elde edilir ve Eş. 3.10'daki gibidir:

$$T(X) = \frac{\sum_{i=1}^{n} w(u_i) x_i}{\sum_{i=1}^{n} w(u_i)} \tag{3.10}$$

Burada $w(u)$, pozitif ve $u \geq 0$ aralığı için azalan bir fonksiyondur. Dolayısıyla $uw(u)$ sınırlıdır. Çalışmada kullanılan ağırlıklar, Maronna ve Yohai (1995) tarafından kullanılan Huber ağırlıklarıdır. Aynı şekilde kovaryans kestirimleri de bulunabilir.

Donoho (1982), Eş. 3.9'u Mahalanobis uzaklığı olarak klasik ortalama ve standart sapma ile yazmıştır. Klasik kestiricilerin kullanılması Mahalanobis uzaklığının aykırı değerlere karşı duyarlı olmasına sebep olmuştur. Ortanca ve ortanca mutlak sapma kullanıldığında daha sağlam kestirici elde etmiştir. Eş. 3.9'da x_i'deki değişim u_i'yi değiştirmediğinden uyum eşdeğişkenlik sağlanmaktadır. Bozulma noktası, $n > 2p + 1$ olduğu sürece %50'dir. u_i'lerin büyük değerleri verinin genel yapısına uyan gözlemlere ait olabilir. Dolayısıyla ağırlıkların azaltılması mantıklıdır (Rousseeuw ve Leroy,1987).

3.3.4. Bisquare-S (BS) Kestiricisi

Davies (1987) tarafından önerilmiştir. p değişkenli olmak üzere $(x_1, x_2, \dots, x_n)$ gözlemli veri kümesinde (T, C)'nin S-kestirimleri, $\sigma(d_1, d_2, \dots, d_n)$'in en küçük

çözümü olacak şekilde tanımlanır. Burada $d_i = (x - T)^T C^{-1}(x - T)$ ve $\det(C) = 1$'dir. $\sigma = \sigma(z)$, $\frac{1}{n}\sum \rho(z/\sigma) = \delta$ eşitliğinin çözümü olarak tanımlanan $z = (z_1, z_2, \dots, z_n)$ veri kümesinin M-ölçek kestirimidir. Bu eşitlikler kullanılırken $\rho(0) = 0$, $\rho(\infty) = 1$ ve $\delta\epsilon(0,1)$ olup ρ azalmayan fonksiyondur. ρ fonksiyonu bisquare fonksiyonudur. T vektörünü ve pozitif tanımlı simetrik C matrisini bulmak için eşdeğişken tanım, Eş. 3.11'i sağlayacak $\det(C)$'yi en küçüklemektir. Eşitlikteki d_i ve ρ yukarıda tanımlandıkları gibidir.

$$\frac{1}{n}\sum_{i=1}^{n} \rho(d_i) = b_0 \tag{3.11}$$

Lopuhaä (1989), S kestirici ve M kestirici çözümleri arasında yakın bir ilişki olduğunu göstermiştir. Aynı zamanda (T, C) çözümünün, M-kestiricisinin ağırlıklandırılmış örneklem ortalaması ve kovaryans matrisi gibi tanımlayan eşitliklerin Eş. 3.13 ve Eş. 3.14 olduğunu göstermiştir (Todorov ve Filzmoser, 2009).

$$d_i^j = \left[\left(x_i - T^{(j-1)}\right)^T \left(C^{(j-1)}\right)^{-1} (x - T^{(j-1)}\right]^{1/2} \tag{3.12}$$

$$T^{(j)} = \frac{\sum w(d_i^{(j)}) x_i}{\sum w(d_i^{(j)})} \tag{3.13}$$

$$C^{(j)} = \frac{\sum w(d_i^{(j)}) \left(x_i - T^{(j)}\right)\left(x_i - T^{(j)}\right)^T}{\sum w(d_i^{(j)})} \tag{3.14}$$

Maronna v. d. (2006), en küçük hacimli elipsoit (MVE- Minimum Volume Elipsoide) kestiricisinin maksimum yan değerinin diğer sağlam kestiricilerden daha küçük olduğunu, bu yüzden S-kestiricisi için bir başlangıç değeri olarak kullanılabileceğini aktarmıştır.

3.4. Kayıp Veri Durumunda Sağlam Kestirim ile İlgili Önceki Çalışmalar

Kayıp veri çözümlemeleri, sağlam yaklaşım ve çalışmada kullanılacak olan sağlam kestirimler, önceki bölümlerde açıklandı. Kayıp veri için en önemli çözüm yollarının en çok olabilirlik ve çoklu veri yükleme yöntemlerinin olduğu kabul edilmektedir.

Son yarım yüzyılda bilgisayar teknolojisindeki ve sağlam istatistikteki büyük gelişmeler, aykırı değer ve kayıp veriye sahip veri kümelerinde istatistiksel çözümlemelerinin kolayca yapılabilmesini ve yeni çözümleme yöntemlerinin geliştirilmesini sağlamıştır. Çalışmanın bu bölümünde kayıp veri durumunda sağlam kestirim için yapılan önemli çalışmalar ve bu çalışmalar sonucunda elde edilen yöntemler açıklanacaktır.

3.4.1. ER Algoritması

ER algoritması, EM algoritmasının M adımında aykırı değerlere karşı dayanıklı kestirimler oluşturmak için kullanılır. Bozulmuş veri kümelerinde aykırı değerlerin kestirimleri etkilememesi için aykırı değerlere düşük ağırlıklar verilerek kestirime olan etkileri azaltılabilir. Bu düşünceden hareketle aykırı değerlerin ağırlıklarının azaltılması için Little ve Smith (1987), EM algoritmasında küçük bir değişiklik yapmışlardır. Little ve Smith, yapmış oldukları benzetim çalışması sonucunda EM algoritmasının varyansı yüksek kestirdiğini ER algoritmasının ise aykırı değerler için yapılan düzeltme nedeniyle beklenenin altında varyans kestirdiğini göstermiştir. ER algoritmasında, kestirimlerin ağırlıkları Mahalanobis uzaklığı ile ilişkilendirilmiştir.

ER algoritması için $t.$ iterasyonda μ ve Σ'nın kestirimleri $\theta^{(t)} = \left(\mu^{(t)}, \Sigma^{(t)}\right)$ ile gösterilsin. $i.$ gözlemde kayıp veri bulunuyorsa $X_{(gi)}$, i gözleminde elde edilmiş X değişkenler vektörünü oluştursun ve g_i gözlem sayılarını versin. $X_{(gi)}$ değişkenlerine ait olan $\mu_{(gi)}$ ve $\Sigma_{(ggi)}$ sırasıyla $(g_i \times 1)$ boyutlu ortalama vektörünü ve $(g_i \times g_i)$ boyutlu kovaryans matrisini göstersin. Bu durumda ER algoritmasının E adımında gözlenmiş değerler ve iterasyonun önceki adımından gelen $\theta^{(t)}$ üzerinden yeterli istatistiklerin beklenen değerleri hesaplanır:

$$E\left\{\sum_{i=1}^{n} X_{ij} | X_{(gi)}, \theta^{(t)}\right\} = \sum_{i=1}^{n} X_{ij}^{(t)}, \quad j = 1,2,\dots,p \tag{3.15}$$

$$E\left\{\sum_{i=1}^{n} X_{ij} X_{ik} | X_{(gi)}, \theta^{(t)}\right\} = \sum_{i=1}^{n} \left\{X_{ij}^{(t)} X_{ik}^{(t)} + C_{jki}^{(t)}\right\}, \quad j,k = 1,2,\dots,p \tag{3.16}$$

Eş. 3.15 ve Eş. 3.16'daki $X_{ij}^{(t)}$ ve $C_{jki}^{(t)}$ gösterimleri aşağıdaki gibi ifade edilir:

$$X_{ij}^{(t)} = \begin{cases} X_{ij}, & X_{ij}\ gözlenmiş\ ise \\ E\left[X_{ij}\middle|X_{(gi)}, \theta^{(t)}\right], & X_{ij}\ kayıp\ ise \end{cases} \quad (3.17)$$

$$C_{jki}^{(t)} = \begin{cases} 0, & X_{ij}\ veya\ X_{ik}\ gözlenmiş\ ise \\ cov\left[X_{ij}, X_{ik}\middle|X_{(gi)}, \theta^{(t)}\right], & X_{ij}\ ve\ X_{ik}\ kayıp\ ise \end{cases} \quad (3.18)$$

$E\{X_{ij}|X_{(gi)}, \theta^{(t)}\}$ yüklenmiş değerleri ve $C_{jki}^{(t)}$ düzeltme değeri $\Sigma^{(t)}$'ye sweep operatör uygulanarak çözümlenebilmektedir (Beale ve Little, 1975).

R adımı için Eş. 3.19 ve Eş. 3.20 kullanılır:

$$\mu_j^{(t+1)} = \sum_{i=1}^{n} w_i X_{ij}^{(t)} \Big/ \sum_{i=1}^{n} w_i \quad (3.19)$$

$$\sigma_{jk}^{(t+1)} = \frac{\sum_{i=1}^{n} w_i^2 \left(X_{ij}^{(t)} - \mu_j^{(t+1)}\right)\left(X_{ik}^{(t)} - \mu_k^{(t+1)}\right) + C_{jki}^{(t)}}{\sum_{i=1}^{n} w_i^2 - 1} \quad (3.20)$$

w_i, Mahalanobis uzaklıklarına bağlı bir ölçüttür ve aşağıdaki gibidir:

$$w_i = \frac{w(d_i)}{d_i} \quad (3.21)$$

Burada d_i, t. iterasyonda değişkenlerin karesel Mahalanobis uzaklığından elde edilir:

$$d_i^2 = \left[X_{(gi)}^{(t)} - \mu_{(gi)}^{(t)}\right]^T \left[\Sigma_{(ggi)}^{(t)}\right]^{-1} \left[X_{(gi)}^{(t)} - \mu_{(gi)}^{(t)}\right] \quad (3.22)$$

$w()$ ise Hampel (1974)'in iki parametreli sınırlı etki fonksiyonudur ve aşağıdaki gibi tanımlanır:

$$w(d_i) = \begin{cases} d_i & d_i \le d_{gi} \\ d_{gi} \exp\{-\dfrac{\left(d_i - d_{gi}\right)^2}{2b_2^2}\} & d_i > d_{gi} \end{cases} \quad (3.23)$$

Burada, $d_{gi} = \sqrt{g_i} + b_1/2$'dir. b_1 kesim noktasını, b_2 ağırlıkların azalış hızını betimlemektedir. Bu sabitler araştırmacıya bağlıdır. Hampel (1973), b_1=2 ve b_2=1,25 değerlerini önermiştir.

Cheng ve Victoria-Feser (2002), ER algoritmasının kayıp veriyi iyi şekilde doldurduğunu ancak yüksek bozulma noktasına sahip olmamasının önemli bir götürü olduğunu açıklamışlardır. Copt ve Victoria-Feser (2003), ER algoritmasının bozulma noktasının yaklaşık olarak $1/(p+1)$ olduğunu belirtmişlerdir. Bu durumda bozulma oranı artınca ER algoritması sağlamlığı bozulmaktadır.

3.4.2. EM Algoritmasında Ağırlıklandırma ile Sağlam Kestirimler

Little (1988a), çalışmasında tamamlanmamış veri kümesinin ML kestirimlerini elde etmek için EM algoritmasının E adımında koşullu beklenen değerleri ağırlıklandırmalarla oluşturmuştur.

y_i, i. gözlem için $Y_1, Y_2, \cdots, Y_p$ sürekli değişkenlerinin değerler vektörü olsun. Kayıp değerler olduğundan veri kümesindeki $y_{g,i}$ i. gözlem içinde gözlenmişleri, $y_{k,i}$ kayıpları tanımlanır ve $y_g = \{ y_{g,i} : i = 1,2, \dots, n\}$, $y_k = \{ y_{k,i} : i = 1,2, \dots, n\}$ olarak yazılabilir. EM algoritmasında yapılan değişiklik için varsayımlar aşağıdaki gibidir:

- $q = (q_1, q_2, \dots, q_n)$ gözlenmemiş değerleri skaler olarak göstersin. Bu durumda $(y_1, y_2, \cdots, y_n), q = (q_1, q_2, \dots, q_n)$ değerleriyle koşullu olarak μ ortalama ve Ψ/q_i kovaryans matrisi ile p boyutlu normal dağılımdan rasgele çekilmiş n büyüklüğünde bir örneklemdir.
- $q = (q_1, q_2, \dots, q_n)$, $q_i > 0$ olmak üzere $h(q_i)$ olasılık yoğunluk fonksiyonundan rasgele dağılan örneklemdir.
- Kayıp veri RK'dir.

Son koşulun özelliği sayesinde olabilirlik çıkarımı, kayıp veri düzeneği düşünülmeden y_g'nin marjinal dağılımına dayanmaktadır. y_k ve q değerleri kayıp veri olarak düşünüldüğünde, μ ve Ψ'nin ML kestirimleri EM algoritması uygulanarak bulunur. Eğer y_k ve q gözlenmiş durumda ise veri, tamamlanmış veri yapısının yeterli istatistikleri ile üstel ailesine aittir:

$$s_0 = \sum_{i=1}^{n} q_i,\ s_y = \sum_{i=1}^{n} q_i\, y_i \text{ ve } s_{yy} = \sum_{i=1}^{n} q_i\, y_i y_i^T \qquad (3.24)$$

μ ve Ψ'nin ML kestirimleri olan ağırlıklandırılmış ortalama ve kovaryans aşağıdaki gibidir:

$$\hat{\mu} = \frac{\sum_{i=1}^{n} q_i \, y_i}{\sum_{i=1}^{n} q_i} = \frac{s_y}{s_0} \tag{3.25}$$

$$\widehat{\Psi} = \frac{\sum_{i=1}^{n} q_i \, (y_i - \hat{\mu})(y_i - \hat{\mu})^T}{n} = \frac{(s_{yy} - \frac{s_y {s_y}^T}{s_0})}{n} \tag{3.26}$$

y_k ve q kayıp ise üstel aileler için EM algoritmasının teorisi uygulanarak ML kestirimleri oluşturulur (Dempster v. d., 1977). EM algoritmasının $(t+1)$. iterasyonu için E adımında, $\mu^{(t)}$, $\Psi^{(t)}$ kestirimleri ve y_g yardımıyla koşullu beklentiler (s_0, s_y ve s_{yy}) kestirilir. M adımında ise E adımında bulunan s_0, s_y ve s_{yy} değerleri ile $\mu^{(t+1)}$ ve $\Psi^{(t+1)}$ kestirimleri elde edilir. Sağlam bir kestirim ile E adımı daha ayrıntılı açıklanabilir. $\mu^{(t)}$ ve $\Psi^{(t)}$ kestirimlerine bağlı olarak,

$$E(s_0|y_g) = E\left(\sum_{i=1}^{n} q_i \, |y_g\right) = \sum_{i=1}^{n} w_i^{(t)} \tag{3.27}$$

şeklinde yazılmaktadır. Burada $w_i^{(t)} = E(q_i|y_{g,i})$'dir ve $w_i^{(t)}$, q_i'nin modeline bağlıdır. $E(s_y|y_g)$'nin j. bileşeni, $\hat{y}_{ij}^{(t)} = E(y_{ij}|y_{g,i})$ olmak üzere ve $y_{g,i}$ ve q_i bilindiği durumda y_{ij}'nin koşullu ortalaması q_i ile ilişkili olmaksızın aşağıdaki gibidir:

$$E\left(\sum_{i=1}^{n} q_i y_{ij} \,\middle|\, y_g\right) = \sum_{i=1}^{n} E\left\{q_i E(y_{ij}|y_{g,i}, q_i)|y_{g,i}\right\} = \sum_{i=1}^{n} w_i^{(t)} \hat{y}_{ij}^{(t)} \tag{3.28}$$

$E(s_{yy}|y_g)$'un (j,k). elemanı aşağıdaki gibidir:

$$\begin{aligned} E\left(\sum_{i=1}^{n} q_i y_{ij} y_{jk} \,\middle|\, y_g\right) &= \sum_{i=1}^{n} E\left\{q_i E(y_{ij} y_{jk}|y_{g,i}, q_i)|y_{g,i}\right\} \\ &= \sum_{i=1}^{n} E\left\{q_i [\hat{y}_{ij}^{(t)} \hat{y}_{jk}^{(t)} + cov(y_{ij} y_{jk}|y_{g,i}, q_i)]|y_{g,i}\right\} \\ &= \sum_{i=1}^{n} \left(w_i^{(t)} \hat{y}_{ij}^{(t)} \hat{y}_{jk}^{(t)} + \Psi_{jk,g,i}^{(t)}\right) \end{aligned} \tag{3.29}$$

Eğer y_{ij} veya y_{ik} gözlenmişse $\Psi_{jk.g,i}^{(t)}$ düzeltme değeri sıfırdır. Eğer y_{ij} ve y_{ik} değerlerinin ikisi de kayıpsa y_{ij} ve y_{ik}'nın kovaryans artıkları q_i'nin (j,k). elemanı ve $y_{g,i}$ kullanılarak elde edilir.

$\Psi^{(t)}$, i durumu için gözlenmiş ve kayıp değişkenlerin bölünmüş haliyle aşağıdaki gibi yazılabilir:

$$\begin{pmatrix} \Psi_{k,i}^{(t)} & \Psi_{cov,i}^{(t)} \\ \Psi_{cov,i}^{(t)} & \Psi_{g,i}^{(t)} \end{pmatrix} \tag{3.30}$$

Eş. 3.30'a göre hem y_{ij} hem de y_{jk} kayıp olduğu durumda $\Psi_{jk.g,i}^{(t)}$ değeri, $\Psi_{k,i}^{(t)} - \Psi_{cov,i}^{(t)\ T}\Psi_{g,i}^{(t)-1}\Psi_{cov,i}^{(t)}$ matrisinin (j,k). elemanı olur. Ψ'nin iterasyondaki kestirimi $\Psi^{(t)}$ üzerinden sweep operatörüyle $\hat{y}_{ij}^{(t)}$ ve $\Psi_{jk.g,i}^{(t)}$'nin değerleri bulunabilir. Kovaryans matrisine uygulanan bu işlemler normal EM algoritmasına uygulananlardan farksızdır (ER algoritmasında da farksızdır). Tek değişiklik son algoritma adımında toplamların ve kareler toplamlarının w_i ile ağırlıklandırması ve vektörel çarpımın M adımına geçmesidir. EM algoritmasında yapılan bu değişiklik ile aykırı değerlerin olduğu gözlemlerin ağırlıkları küçültülerek sorunlara çözüm getirmek amaçlanmıştır.

$h(q_i)$'nin farklı seçimleri için ağırlıkların değişik alınması önemlidir. İlk olarak, $q_i v$, v serbestlik derecesinde ki-kare dağıldığı varsayıldığında q_i incelensin. Bu durumda marjinal dağılım,

$$y_i \sim_{bağımsız} t_p(\mu, \Psi, \nu) \tag{3.31}$$

şeklindedir. μ ortalama, Ψ varyans matrisi ve ν serbestlik derecesi ile Student t dağılır. EM algoritmasında yapılan bu değişik sayesinde Eş. 3.31'den μ ve Ψ'nin ML kestirimleri elde edilir. $\nu > 2$ için y_i'nin kovaryans matrisi $v\Psi/(\nu - 2)$ ve ML kestirimi $v\widehat{\Psi}/v - 2$'dir.

$y_{g,i}$, $\mu^{(t)}$ ve $\Psi^{(t)}$'si verilmiş q_i'nin dağılımı $v + g_i$ serbestlik derecesi ile $q_i(v + d_i^{(t)2})$ şeklinde ki-kare dağılmaktadır. Burada g_i, $i.$ durum için gözlenmiş olan değişkenlerin sayısı ve $\mu_{g,i}$ ve $\Psi_{g,i}$ değerleri parametrenin o andaki kestirimlerinden

hesaplanmış ortalama vektörü ve ölçek matrisi kestirimleri olmak üzere, gözlenmiş değişkenlerin ortalamadan uzaklıklarının karesi aşağıdaki gibidir:

$$d_i^{(t)2} = \left(y_{g,i} - \mu_{g,i}^{(t)}\right)^{\mathrm{T}} \Psi_{g,i}^{(t)^{-1}} \left(y_{g,i} - \mu_{g,i}^{(t)}\right) \tag{3.32}$$

Tam veri için gerekli olan ağırlıklar, Maronna (1976) ve Rubin (1983)'de verilmiştir. Verilen bu ağırlıklardan hareketle $w_i^{(t)}$ ağırlığı Eş. 3.33 ile bulunur:

$$w_i^{(t)} = E\left(q_i | y_{g,i}, \mu^{(t)}, \Psi^{(t)}\right) = (v + g_i)/(v + d_i^{(t)2}) \tag{3.33}$$

$0 < \delta < 1,\ \lambda > 0$ olup bu iki değerinde bilindiği varsayımı altında, q_i için ikinci model aşağıdaki gibidir:

$$h(q_i) = \begin{cases} 1 - \delta & eğer\ q_i = 1\ ise \\ \delta & eğer\ q_i = \lambda\ ise \\ 0 & diğ.durumlar \end{cases} \tag{3.34}$$

Bu durumda y_i'nin marjinal dağılımı $N(\mu, \Psi)$ ve $N(\mu, \Psi/\lambda)$ dağılımlarının karışımından oluşmaktadır. λ değerinin 1'den çok küçük $(\lambda \ll 1)$ olduğu düşünülürse bozulmuş çok değişkenli normal model elde edilmektedir. $h(q_i)$'nin dağılımı için $w_i^{(t)}$ ağırlığı Eş. 3.35 gibidir ve burada bulunan $d_i^{(t)2}$'nin değeri Eş. 3.32'de verildiği gibidir.

$$w_i^{(t)} = \frac{[1 - \delta + \delta\lambda^{1+p_i/2} \exp\left\{\frac{(1-\lambda)d_i^{(t)2}}{2}\right\}]}{[1 - \delta + \delta\lambda^{p_i/2} exp\left\{\frac{(1-\lambda)d_i^{(t)2}}{2}\right\}]} \tag{3.35}$$

Eş. 3.35'teki ağırlık kullanıldığında y_i'nin koşulsuz kovaryans matrisi ve ML kestirimi ise sırasıyla $\Psi(1 - \delta + \delta/\lambda)$ ve $\hat{\Psi}(1 - \delta + \delta/\lambda)$ şeklindedir. Model örnekleme ait olmayan aykırı değerleri incelemek amacıyla kullanılırsa, bozulmanın olmadığı durumda $\hat{\Psi}$, y_i'nin kovaryans matrisinin sağlam kestirimidir.

Little (1988a) çalışması sonucunda hem çok değişkenli t dağılımı hem de bozulmuş çok değişkenli normal dağılım için büyük karesel uzaklıklardaki durumların ağırlıkları azaltılmıştır.

3.4.3. Kayıp Veri Durumunda Sağlam Doğrusal Regresyon

Atkinson ve Cheng (2000), EM algoritması ve çoklu veri yükleme yöntemleri arasındaki ilişkileri ve farklılıkları incelemişlerdir. Kayıp veri için yapılacak veri yükleme yöntemlerinde veri kümesinde aykırı değerlerin olmasının yükleme yöntemlerini etkilediğini görmüşlerdir. Kayıp veri kümelerinde aykırı değerlerin etkisinden arınacak bir çözümleme için EM algoritması ya da çoklu veri yükleme yöntemi ile yüksek bozulma noktasına sahip kestiricilerin ileriye doğru araştırma algoritmasını (FSA- forward search algorithm) birleştirmişlerdir.

Regresyon modeli,

$$Y = \mathfrak{X}\beta + \varepsilon = \beta_0 + X\gamma + \varepsilon \tag{3.36}$$

biçimindedir. Burada, Y, bağımlı değişken vektörü, $\mathfrak{X}$, $p-1$ açıklayıcı değişkenli, $n \times p$ boyutlu matristir. $\beta = (\beta_0 \quad \beta_1 \quad \dots \quad \beta_{p-1})$, regresyon katsayıları vektörü ve ε ise $n \times 1$ boyutlu ve $N(0, \sigma^2 I)$ dağılımlı hatalar vektörüdür. Tamamlanmış veri için EKK kestiricisi $\hat{\beta} = (\mathfrak{X}^T\mathfrak{X})^{-1}(\mathfrak{X}^T Y)$'dir ve sağlam bir kestirici olmaması nedeniyle bir veya birkaç aykırı değerden etkilenir.

Tamamlanmamış veri yapısında ise $n \times p$ veri kümesinde bazı değerler, $Z = (Y, X) = (z_{ij})$, RK'dır. Burada amaç $z = (y, x^T)$ birleşik fonksiyonu andıran dağılımının olabilirliğini maksimum etmektir. z, genellikle çok değişkenli normal dağılımdır:

$$z_i = \begin{pmatrix} y_i \\ x_i \end{pmatrix} \sim \left[\begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_y^2 & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{pmatrix} \right] \tag{3.37}$$

$\theta = (\mu, \Sigma)$'nın ML hesaplamaları için iteratif yöntemler EM algoritması kullanılarak gösterilebilir. θ için başlangıç değeri θ_0 olsun ve R_i, i. durum içinde gözlenen değişkenlerin vektörü ve σ_{jk}, Σ'nın (j,k). bileşenini göstersin. E adımında veri matrisi doldurulur ve gözlenen değişkenler yardımıyla gözlenmeyen değişkenlerin koşullu kovaryansları kestirilir. Her $i = 1,2,\dots,n$ ve her $j = 1,2,\dots,p$ için yüklenen değerler genellikle koşullu beklentilerdir. Bu değerler önceki çalışmalarda açıklanmıştır.

Regresyon modeli Eş. 3.36'daki ikinci eşitlik şeklinde ifade edildiğinde regresyon parametreleri β ve σ^2'nin ML kestirimleri aşağıdaki gibi elde edilir:

$$\hat{\gamma} = \hat{\Sigma}_{xx}{}^{-1}\hat{\Sigma}_{xy}$$

$$\hat{\beta}_0 = \hat{\mu} - \hat{\beta}^T\hat{\mu}_x \tag{3.38}$$

$$\hat{\sigma}^2 = \hat{\sigma}_y^2 - \hat{\Sigma}_{yx}\hat{\Sigma}_{xx}{}^{-1}\hat{\Sigma}_{xy}$$

Eş. 3.38'den γ'nın ML kestirimlerinin yaklaşık biçimi aşağıdaki gibi yazılabilir:

$$\hat{\gamma} = \left(\hat{X}^T\hat{X} + \hat{C}\right)^{-1}\hat{X}^T Y \tag{3.39}$$

Burada $\hat{C} = \sum_{i=1}^{n}\hat{C}_i$'dir (Shih ve Weisberg, 1986).

Little ve Rubin (1987) yukarıdaki kestirimlerin sonuçlarına ulaşabilmek için sweep operatör uygulamışlardır. Ancak, kestirilen regresyon katsayılarının kovaryans matrisini elde etmekte sweep operatör problem çıkarmaktadır. Little (1979), kovaryans matrisinin kestirimi için bir yöntem önermiştir. Yöntemin yardımıyla tamamlanmamış verinin kovaryans matrisinin $Var(\beta)$ kestirimi Eş. 3.40'taki gibidir (Beale ve Little, 1975):

$$A_w = \hat{\sigma}^2 S_w^{-1} = \hat{\sigma}^2\left(\hat{\mathfrak{X}}^T\widehat{W}\hat{\mathfrak{X}}\right)^{-1} \tag{3.40}$$

Eş. 3.40'ta verilen $\widehat{W}$ köşegen matrisi Eş. 3.41 ile açıklanır:

$$w_i = \begin{cases} 1 & tam\ gözlemler\ için \\ \hat{\sigma}_y^2 \Big/ \hat{\sigma}_{yi}^2 & diğer\ durumlar\ için \end{cases} \tag{3.41}$$

Burada $\hat{\sigma}_{yi}^2$, i. durumda x'in gözlenen kısmı için y'nin kestirilmiş artık varyansıdır. $\hat{\sigma}_y^2$ ise verilen bütün bağımsız değişkenler için y'nin kestirilmiş artık varyansıdır.

Atkinson ve Cheng (2000), çoklu veri yükleme yönteminin özelliklerini ve eşitliklerini vermişlerdir. Veri kümelerinin kayıp gözelerine çeşitli yöntemlerle yükleme yapılabilir. Yükleme yapılan değerler yöntemlere göre değişiklikler içermektedir. Çoklu veri yükleme yönteminde kayıp veri için birden fazla yükleme yapılır ve yapılan yükleme sayısı kadar farklı tam veri kümesi oluşturulur. Her veri kümesi ayrı ayrı tam veri kümeleri için tasarlanmış tekniklerle çözümlenir ve sonuçlar birleştirilir. Sayısal olarak açıklanırsa, Y_g, gözlenen değerler kümesi ve Y_k, kayıp değerler kümesidir. Örneklem ortalaması Q'nun sonsal yoğunluk fonksiyonu,

$f(.)$, kayıp değerlerin sonsal yoğunluk fonksiyonu ve $g(.)$ tamamlanmış verinin sonsal yoğunluk fonksiyonu olmak üzere aşağıdaki gibi yazılır:

$$h(Q|Y_g) = \int g(Q|Y_g, Y_k)\, f(Y_k|Y_g) dY_k \tag{3.42}$$

$\hat{Q}$ ve U tam veri istatistikleri, s tamamlanmış veri kümeleri $\hat{Q}_{*1}, \hat{Q}_{*2}, \dots, \hat{Q}_{*s}$ ve $U_{*1}, U_{*2}, \dots, U_{*s}$ üzerinden hesaplanır. Tekrarlı yüklemenin kestirimi Eş. 3.43'deki gibidir:

$$\bar{Q}_s = \sum_{l=1}^{s} Q_{*l}/s \tag{3.43}$$

$\bar{Q}_s$'nin kovaryans matrisi Eş. 3.44'te, iç veri yükleme değişkenliği Eş. 3.45'te ve veri yüklemeler arası değişkenlik Eş. 3.46'da gösterilmiştir:

$$T_s = \bar{U}_s + \frac{s+1}{s} B_s \tag{3.44}$$

$$\bar{U}_s = \sum_{l=1}^{s} U_{*l}/s \tag{3.45}$$

$$B_s = \frac{\sum_{l=1}^{s} (Q_{*l} - \bar{Q}_s)\,(Q_{*l} - \bar{Q}_s)^T}{s-1} \tag{3.46}$$

s tamamlanmış veri kümesinin çok olması, $(Q - \bar{Q}_s)$'yi T_s kovaryans matrisi ile normal dağılıma götürür. $s = \infty$ durumunda, $T_\infty = \bar{U}_s + B_s$ şeklinde olmak üzere $(Q - \bar{Q}_\infty) \sim N(0, T_\infty)$'dir.

Atkinson ve Cheng (2000), EM algoritması ve çoklu veri yükleme arasında yapılan benzetim çalışmalarında çoklu veri yükleme yönteminin EM algoritmasından daha iyi sonuçlar verdiğini açıklamışlardır. Çalışmanın devamında FSA ile veriden rasgele $m = p + 1$ gözlem alt örneklem ve başlangıç gösterimi $\left(\hat{\mu}_m^{(0)}, \hat{\Sigma}_m^{(0)}\right)$ olarak alınmıştır. EM algoritmasının kullanılmasıyla kayıp değerler yüklenir; m durumlarının ortalama ve kovaryans matrisleri, $(\hat{\mu}_m, \hat{\Sigma}_m)$, elde edilir; $\hat{X}_m$ tamamlanmış veri olarak hazırlanır. Regresyon katsayıları ise $\hat{X}_m$ tasarım matrisini ve Y_m, m gözleme ait bağımlı değişken olmak üzere Eş. 3.47'deki gibidir:

$$\hat{\beta}(m) = \left(\hat{X}_m^T \hat{X}_m + \hat{C}_m\right)^{-1} \hat{X}_m^T Y_m \tag{3.47}$$

$\hat{C}_m = \sum_{i\epsilon m} \hat{C}_i$'dir. Verilen m için bu kestirimdeki artıklar, $e_{i,m} = y_i - \hat{x}_i^T \hat{\beta}(m)$'dir $(i = 1,2, \dots, n)$. Artıklar, en küçükten başlayarak sıralandığında, $e^2_{(1),m} \leq e^2_{(2),m} \leq \cdots \leq e^2_{(n),m}$, LTS ölçütüne göre varyans kestirimi,

$$\hat{\sigma}_q^2(m) = \frac{\sum_{i=1}^{q} e^2_{(i),m}}{q - p} \tag{3.48}$$

biçiminde verilir. Atkinson ve Cheng (1999), Eş. 3.48 için q'nun seçimini tartışmışlardır. Her FS adımı $m = n$ oluncaya kadar başarılı bir artış içerir ve $\hat{\sigma}_q^2(m)$ serisi elde edilir. $\hat{\sigma}_{q,j}^2$'nin en küçük değeri, j. araştırmanın performansını tanımlamaktadır. σ^2'nin genel kestirimi araştırmalara göre $\hat{\sigma}_q^2(m) = min_j \hat{\sigma}_{q,j}^2$ durumudur. İleriye doğru araştırmalar, altkümede içerilmeyen gözlemler için ölçeklendirilmiş artıkları da kullanır. İçerilen m gözlemde EKK artığı için $e_{i,m}^2$ kullanılır. Ancak, $n - m$ gözlem içerilmemesine rağmen, EKK artıkları varyans kestirimi ile ölçeklendirilir. $\mathcal{M}$, m durumlarının alt kümesi kabul edildiğinde n tane artığın karesi r_i^2 sıralaması aşağıdaki gibidir:

$$r_i^2 = \begin{cases} e_{i,m}^2, & i \epsilon \mathcal{M} \\ \dfrac{e_{i,m}^2}{1 + d_i}, & i \notin \mathcal{M} \end{cases} \tag{3.49}$$

Burada $d_i = \hat{x}_i^T \left(\hat{X}_m^T \hat{X}_m + \hat{C}_m\right)^{-1} \hat{x}_i$'dir. Eş. 3.49'daki artıklar sıralaması kuralına dayalı olarak altküme büyüklüğü $m + 1$'e çıkartılır. Genel olarak altkümeye bir gözlem eklenmektedir ama bazen iki veya daha fazla gözlemler eklenebilmektedir. $m + 1$ durumu temelinde yeni iteratif EM algoritması başlangıç noktası $\left(\hat{\mu}_{m+1}^{(0)}, \hat{\Sigma}_{m+1}^{(0)}\right)$ verilerek, $(\hat{\mu}_{m+1}, \hat{\Sigma}_{m+1})$ değerleri, yüklenen değerler ve gerekli kestirimleri oluşturmak için yeniden başlatılır.

FSA'nın önemli bir getirisi, aykırı gözlemleri ilerleyen süreçlerde çıkarabilir ve bu durumda yüklenmiş değerleri iyi veri üzerinden elde edilmiş olur. FS, artıkları düzeltmek için bağımsız değişken X ile aynı modelde yapay veriyi kullanırken Y için standart normal dağılımla işlemler sürdürülür. Hesaplamalarda zaman kazanabilmek adına X'in kayıp değerleri için ortanca değerleri yüklenebilir. Bu adımda amaç, σ_q^2 kestiriminde yanı azaltmak ve artıkları kolaylıkla belirleyebilecek şekilde ölçeklendirmektir.

Aykırı değerleri belirlemek için kullanılan artıklar, standartlaştırılmış student tipi artıklardır:

$$t_i = \begin{cases} \dfrac{e_{i,m}\bar{\sigma}_q(m)}{\tilde{\sigma}_q(m)\sqrt{(1-h_i)}}, & i \in \mathcal{M} \\ \dfrac{e_{i,m}\bar{\sigma}_q(m)}{\hat{\sigma}_q(m)\sqrt{(1+d_i)}}, & i \notin \mathcal{M} \end{cases} \tag{3.50}$$

Burada h_i, d_i ile aynı formda ve şapka (hat) matris köşegen değeridir. Ancak $i \epsilon \mathcal{M}$ durumu için hesaplanır. $\bar{\sigma}_q(m)$ ise 100 FS benzetimi ile m için elde edilen $\hat{\sigma}_q(m)$'in ortalamasıdır. Eğer çoklu veri yükleme algoritmadaki EM adımıyla yer değiştirirse, aradaki fark regresyon katsayısının kestiriminde Eş. 3.39'daki $\bar{b}$'nin yerine Eş. 3.51'deki eşitlikleri kullanmaktır:

$$b_l = (\hat{\mathfrak{X}}_l^T \hat{\mathfrak{X}}_l)^{-1} \hat{\mathfrak{X}} Y, \quad l = 1,2,\dots,s$$

$$\bar{b} = \sum_{l=1}^{s} b_l / s$$

$$U = \frac{\sum_{l=1}^{s} \hat{\sigma}_{yl}^2 \left(\hat{\mathfrak{X}}_l^T \hat{\mathfrak{X}}_l\right)^{-1}}{s} \tag{3.51}$$

$$B = \sum_{l=1}^{s} (b_l - \bar{b})\left(b_l - \bar{b}\right)^T / (s-1)$$

$$Var\left(\bar{b}\right) = U + \frac{s+1}{s} B$$

Eş. 3.50 için sapka matrisi ve artıklar ise aşağıdaki şekilde elde edilir:

$$e_{il} = y_i - \hat{x}_{i,l}^T b_l \qquad i = 1,2,\dots,n \; ve \; l = 1,2,\dots,s$$

$$\bar{e}_i = \sum_{i=1}^{s} \frac{e_{il}}{s} \tag{3.52}$$

$$d_{il} = \hat{x}_{i,l}^T \left(\hat{X}_{m.l}^T \hat{X}_{m,l}\right)^{-1} \hat{x}_i$$

$$\bar{d}_i = \sum_{i=1}^{s} \frac{d_{il}}{s} \tag{3.53}$$

3.4.4. Kayıp Veri ile Yüksek Bozulma Noktasına Sahip Kestiriciler

Cheng ve Victoria-Feser (2002), tamamlanmamış çok değişkenli verideki aykırı değer problemini ele almışlardır. ER algoritmasının bazı yüksek boyutlarda sağlam olmaktan uzak olduğunu gözlemleyerek iki alternatif yol önermişlerdir: İteratif yöntemlerin başlangıç noktasını yüksek bozulma noktasına sahip kestiricilerle birleştirmek ve yüksek bozulma noktalı kestiricileri ER algoritmasının kestirim adımlarıyla kullanmak. Her iki alternatifte de önerilen yüksek bozulma noktalı kestiriciler MCD ve translated-Bisquare S (TBS)'dir.

Cheng ve Victoria-Feser (2002), çalışmalarında kayıp verinin en az RK veri düzeneğinde olma durumunu göz önünde bulundurmuşlardır. Kayıp veri düzeneği ihmal edilebilir varsayıldığında gözlenen değerler üzerinden ML kestirimi ya da sağlam kestirimler yapılabilir. Ancak, bu ihmal iki önemli soruna sebep olabilir: Birkaç kayıp değerden kurtulmak için örneklem büyüklüğünde önemli bir azalma ile karşılaşılabilir; örneklem büyüklüğündeki kayıp değerler yüzünden parametreler kestirilemeyebilir.

Klasik kestirimlerde kayıp verinin ML kestirimleri ya da beklenen değerleri yardımıyla yüklenmelerinin yerine EM algoritmasıyla yüklenmeleri iyi sonuçlar vermektedir. Ancak veri kümesinde aykırı değer bulunması durumunda EM algoritmasının kestirim sonuçları bozulur. Little ve Smith (1987) ve Little (1988a) bu sorunu ER algoritması ile aşmaya çalışmıştır. Ancak ER algoritmasının bozulma noktası $1/(1+p)$'dir ve bozulma oranı yüksek olan veri kümelerinde sağlam değildir. Bu yüzden yüksek bozulma noktalı kestiriciler kullanılabilir.

Atkinson (1994) tarafından açıklanan FSA'nın MCD kestiricisi için adımları, altküme Q_k ve artış miktarı s olmak üzere aşağıdaki gibidir:

- Kümeden $\bar{y}(Q_k)$ ve $S(Q_k)$ olan örneklem ortalaması ve kovaryans matrisi hesaplanır.
- $\bar{y}(Q_k)$ ve $S(Q_k)$ kullanılarak Mahalanobis uzaklıkları bütün n gözlemleri için sıralanır.
- İlk h gözlem seçilerek bu gözlemlerin $\bar{y}_h(Q_k)$, $S_h(Q_k)$ ve D_k ile gösterilen örneklem ortalaması, kovaryans matrisi ve determinantı hesaplanır.

- Eğer $D_k < D_{k-1}$ ise ilk kestiriciler olan μ_{MCD} , Σ_{MCD} ve D, elde edilen $\bar{y}_h(Q_k)$, $S_h(Q_k)$ ve D_k ile değiştirilir.
- Eğer q_k altküme gözlem sayısı n'ye eşitse işlem durur. Değilse, 2. adımda elde edilen sıralı gözlemlerin ilk $q_{k+1} = q_k + s \leq n$'si seçilir ve bu küme Q_{k+1}olarak tanımlanarak 1. adıma dönülür ve Q_k ile yer değiştirilir.

Altküme $q_1 = p + 1$ gözlem ile başlar ve FS adımı en küçük determinant D'yi bulana kadar devam eder. FS yöntemi birkaç rasgele seçilen başlangıç altkümeleri ile tekrar etmektedir.

Veri kümesinde kayıp veri bulunduğunda MCD kestiricilerinin örneklem ortalaması ve kovaryansı olarak tanımlanması, kovaryans matrisinin determinantı en küçük olacak şekilde h gözlemli altkümelerinin EM algoritması aracılığıyla sağlanır. Burada FSA'nın kayıp veriye uyumu gerekmektedir. Tamamlanmış verideki olağan h seçimi, $h = \lfloor (n + p + 1)/2 \rfloor$, olmasına rağmen tamamlanmamış veride bilgi kaybından dolayı h değerini daha küçük bulma olasılığı vardır. Önceki deneyimlerden α bozulma oranı olmak üzere h değeri, $\lfloor n(1 - \alpha) \rfloor$ gibi yüksek bir değer almalıdır. FSA uygulanırken Q_k altkümesindeki kayıp veri durumu için MCD kestiricisine basit bir uyarlama yapılmaktadır. Tamamlanmış veri kümesinden tek farkı, gözlemlerin sıralandığı durumlardaki Mahalanobis uzaklıklarıdır. i. gözlemde kayıp veri varsa bu durumda Mahalanobis uzaklığı aşağıdaki gibi hesaplanır:

$$d^2_{[gi]}(Q_k) = \left(y_{[gi]} - \bar{y}(Q_k)_{[gi]}\right)^T S(Q_k)^{-1}_{[ggi]} \left(y_{[gi]} - \bar{y}(Q_k)_{[gi]}\right) \tag{3.54}$$

Eş. 3.54'te Q_k, FSA için seçilen alt kümeyi, $\bar{y}(Q_k)_{[gi]}$ gözlenen değerler üzerinden altkümenin ortalamasını, $S(Q_k)_{[ggi]}$ gözlenen değerler üzerinden kovaryans matrisini göstermektedir. $d^2_{[gi]}(Q_k)$ ise i. durum için gözlenen değerler üzerinden hesaplanan Mahalanobis karesel uzaklığıdır. Her gözlemdeki kayıp sayısının eşit olmadığı düşünülerek ağırlıkları sıralayabilmek için standartlaştırma uygulanır. Standartlaştırma için Little ve Smith (1987) tarafından önerilen ve Eş. 3.55'teki gibi gösterilen ki-kare dağılımının Wilson-Hilferty dönüşümüdür. i gözleminde bozulma yoksa, veri normalse ve kayıp veri RK ise Mahalanobis uzaklığı asimptotik olarak $\chi^2_{g_i}$biçiminde dağılmaktadır (Kendall ve Stuart, 1969).

$$Z_i = \frac{((d_i)^2/g_i)^{1/3} - 1 + 2/(9g_i)}{\sqrt{2/(9g_i)}} \tag{3.55}$$

Kayıp verinin yerine yüklenen değerlerin ortalamaya yakın olması, yüklenen değerlerin gözlemleri için düşük Mahalanobis uzaklıklarına sahip olmasına sebep olur. Bu durumda kayıpları yüklenmiş gözlemlerin ilk h gözlemin içine girme şansı artmaktadır. Standartlaştırma yapılmasının amacı kayıp verisi bulunan gözlemlerin şansını ortadan kaldırmaktır. Ancak Eş. 3.54'te tanımlanan Mahalanobis uzaklığı gözlenen kısımlar üzerinden hesaplamalar yaptığı için bu sıkıntı olmayacaktır. Ancak Cheng ve Victoria-Feser (2002), yine de MCD için FSA'nın 2. adımında dönüşüm kullanmışlardır.

Diğer taraftan TBS ise yüksek bozulma noktasına sahip diğer bir kestiricidir ve iteratif adımlar şeklinde çözülebilir. Rocke (1996)'un önerisi bu iteratif yöntemlerden biridir. Cheng ve Victori-Feser (2002) ise yeni bir öneri getirilmişlerdir. Bu öneride çözümlemeler ER algoritmasındaki adımlardan yapılmaktadır. E adımında $\hat{y}_i$, C_i ve $d_i = d(y_{[gi]})$ değerleri sırasıyla aşağıdaki eşitliklerden hesaplanır:

$$\hat{y}_{ij} = \begin{cases} y_{ij} & y_{ij}\ gözlenmiş \\ \mu_{[ki]} + \Sigma_{[kgi]}\Sigma_{[ggi]}^{-1}(y_{[gi]} - \mu_{[kgi]}) & y_{ij}\ kayıp \end{cases} \tag{3.56}$$

$$C_{ijk} = \begin{cases} 0 & y_{ij}\ veya\ y_{ik}\ gözlenmiş \\ \Sigma_{[kki]} - \Sigma_{[kgi]}\Sigma_{[ggi]}^{-1}\Sigma_{[gki]} & y_{ij}\ ve\ y_{ik}\ gözlenmemiş \end{cases} \tag{3.57}$$

$$d_{[gi]}^2 = \left(y_{[gi]} - \mu_{[gi]}\right)^{\mathrm{T}}\Sigma_{[ggi]}^{-1}\left(y_{[gi]} - \mu_{[gi]}\right) \tag{3.58}$$

R adımında ise E adımında hesaplanan $\hat{y}_i$, C_i ve d_i değerlerinin yardımı ile ağırlıklar ve k değeri aşağıdaki eşitliklerle hesaplanır:

$$k = \frac{d_{(q)}}{\chi_p^2(q/(n+1)} \tag{3.59}$$

$$w_i^{\mu} = v_1(d_i) = \frac{k\psi\left(\frac{d_i}{k}\right)}{d_i} \tag{3.60}$$

$$w_i^{\eta} = v_2(d_i) = p\ v_1(d_i/k) \tag{3.61}$$

$$w_i^{\delta} = v_3(d_i) = \frac{\psi\left(\frac{d_i}{k}\right) d_i}{k} \tag{3.62}$$

Yukarıdaki eşitlikler kullanılarak μ ve Σ Eş. 3.63 yardımıyla güncellenir:

$$\frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} w_i^{\mu}(\mu_i - \hat{y}_i) \\ w_i^{\delta} vech(\Sigma) - w_i^{\eta}(vech((\mu_i - \hat{y}_i)(\mu_i - \hat{y}_i)^T - vech(C_i)) \end{bmatrix} = 0 \tag{3.63}$$

Eş. 3.63'teki $vech()$ fonksiyonu $(p \times p)$ boyutlu simetrik matrisin elemanlarını tekrar etmeden $p(p+1)/2$ elemanlı bir sütun vektörü olarak tanımlanmıştır.

Cheng ve Victoria-Feser (2002), ER algoritmasını açıklamışlar ve ER algoritması kullanılarak oluşturulan MCD ve TBS kestiricileri ile benzetim çalışması yapmışlardır. Kayıp veri bulunması durumunda ER algoritmasının iteratif adımlarını kullanarak yüksek bozulma noktasına sahip kestiricilerin kayıp veri sorunlarını çözümlemişlerdir.

3.4.5. Mikroçip (Microarray) Veride Sağlam Kayıp Veri Yükleme Yöntemi

Mikroçiplerde kayıp veri sorunu yıllardır çözülmeye çalışılmaktadır. Kayıp veri için şimdiye kadar ortalama ile veri yükleme, satır ortalamasını yükleme, en yakın komşu yöntemi, ortalama kümeleri, organize şemalar, EKK, tekil değer ayrışımı ile veri yükleme gibi yöntemler kullanılmıştır. Ayrıca kısmi EKK (PLS) yöntemi de yine kullanılmış çözümlemelerden biridir. Ancak bu yöntemlerin çoğu aykırı değerlere karşı duyarlıdır ve veri kümesinden aykırı değer bulunması durumunda verinin geneli hakkında bilgi vermeyebilirler.

Cao ve Poh (2006), en küçük mutlak sapma (LAD) yöntemini kullanarak veri yüklemesi yapabilmek için bir algoritma önermişlerdir. LADimpute şeklinde adlandırdıkları yöntemin biri Öklid uzaklıklarıyla ilişkili iken (LADimpute / L2) diğeri Pearson ilişki katsayısı ile ilişkilidir(LADimpute / PC).

LADimpute için önerilen algoritma adımları aşağıdaki gibidir:

- Kayıp gözlemler satır ortalamaları ile yüklenir.
- Hedef gözlem olarak ilk kayıp gözlem içeren satır alınır (g_t).
- g_t hedef gözleminin Öklid uzaklığına ya da pearson ilişki katsayısına bakarak en yakın k gözlem seçilir ($g_{s_i}, s = 1,2, \dots, k$).

- k en yakın ilişkili gözlemlerle, g_t hedef gözleminin $\hat{y}_1$, $\hat{y}_2$, … , $\hat{y}_k$ kayıp değer(ler)i g_{s_1}, g_{s_2}, … , g_{s_k} gözlemleri üzerinden LAD temelinde regresyon değerleri yardımıyla hesaplanır.
- g_t'nin bütün kayıp değerleri $\hat{y}_1$, $\hat{y}_2$, … , $\hat{y}_k$ değerlerinin ağırlıklı ortalamalarından hesaplanır.
- Veri kümesinde başka kayıp veri varsa 2. adıma geri dönülür.

4. adımda $\hat{y}_1$, $\hat{y}_2$, … , $\hat{y}_k$'ler g_t'deki kayıp veri durumuna göre vektör olabilirler. 5. adımdaki ağırlıklar en yakın gözlemlere büyük ağırlıkları vermek üzerine kurulmuştur. Bu durumda r_{ts_i}, g_t ve g_{s_i}'ler arasındaki ilişkileri göstermek üzere, g_{s_i} için Pearson ilişki katsayısında kullanılan ağırlık aşağıdaki gibidir:

$$w_i = \frac{r_{ts_i}^2}{1 - r_{ts_i}^2 + \varepsilon} \tag{3.64}$$

Eş. 3.64'teki ε değeri paydanın sıfır olmasını engellemek için ufak bir değerdir. g_t ve g_{s_i}'ler arasındaki Öklid uzaklığı $\|g_t - g_{s_i}\|_2$ gibi ifade edilirse diğer ağırlık aşağıdaki gibi dir:

$$w_i = \frac{1}{\|g_t - g_{s_i}\|_2} \tag{3.65}$$

Ağırlıklar standartlaştırıldığı zaman ise toplamları 1 olacaktır. Yapılan gerçek veri çalışmasında LADimpute yönteminin her iki ağırlıkla kullanımında da en yakın komşu yöntemi ve EKK yöntemlerinden daha iyi sonuçlar verdiği görülmüştür. Ayrıca Cao ve Poh (2006), LADimpute yönteminin %25'lik kayıp veri durumunda bile sağlamlığını kaybetmediği açıklamışlardır.

3.4.6. Kısmi En Küçük Kareler (PLS) ve Kısmi Sağlam M-Regresyonu (PRM)

PLS yöntemi sınıflandırmada, veri madenciliğinde ve regresyon amaçlı kullanılabilme özelliği nedeniyle son zamanlarda popüler bir yöntemdir. Başarısı, kısa bir sürede yüksek boyutlu veriyi iki ilişkili blok olarak ayırabilmesidir. Aynı zamanda çoklu bağlantı sorununu da çözümleyebilmektedir. Ancak aykırı değerlere karşı duyarlı olması nedeniyle sağlam bir yöntem değildir. Bu durumda PRM modelleri kullanışlıdır. Ancak kayıp veri problemi durumunda ikisinin de iteratif çözümlemelere ihtiyaçları olacaktır. İki yöntem de EM algoritmasının iteratif

yöntemlerini kullanmaktadır. Kayıp veri bulunması durumunda EM-PLS yönteminin adımları aşağıdaki gibidir:

- Satır ve sütunların ortalamasından kayıp veri yüklenir.
- Ortalama ile X ve y'ler merkezileştirilir.
- Faktör sayısı 1 için;
    - PLS skorları, yüklemeleri ve ağırlıkları hesaplanır.
    - Kestirilen X ve y'ler ile gerçek X ve y'ler değiştirilir.
- Çapraz geçerliliğin kök hata kareler ortalaması üzerinden optimal faktör sayısını seçebilmek için PLS skor ve yüklerini kullanarak X ve y kestirilir.
- X matrisindeki kayıp veri, geçici modeldeki kestirimleriyle yüklenir ve yakınsaklık koşulu sağlanmadıysa 2.adıma geri dönülür.

EM-PRM'nin adımları ise EM-PLS'ye benzerdir ve aşağıdaki gibidir:

- Satır ve sütunların ortancasını kullanarak kayıp veri yüklenir.
- Ortanca ile X ve y'ler merkezileştirilir.
- Faktör sayısı 1 için;
    - PRM sağlam skorları, sağlam yüklemeleri ve ağırlıkları modeldeki ağırlıklar ve PRM skorları uzayındaki Mahalonobis uzaklığı ile hesaplanır.
    - Kestirilen X ve y'ler ile gerçek X ve y'ler değiştirilir.
- PRS skor ve yüklerini kullanarak X kestirilir.
- X matrisindeki kayıp veri, sağlam modeldeki kestirimleriyle yüklenir ve yakınsaklık koşulu sağlanmadıysa 2. adıma geri dönülür.

Görülebileceği gibi farklılıklar başlangıç değerlerinin atanmasında ve uzaklık ifadesindedir. Yapılan benzetim çalışmasında Stanimirova v.d. (2007), kayıp veri oranı arttıkça hata kareler ortalama karekökünün arttığını belirtmişlerdir.

Aykırı değerler ve bozulmalar olmadığı durumlarda EM-PLS biraz daha iyi sonuçlar verirken aykırı değerler arttırıldığında ise EM-PRM daha iyi sonuçlar vermekte ve aykırı değer oranı arttıkça EM-PLS'den daha etkin hale gelmektedir. Ayrıca y yönünde güçlü aykırı değerlerin ve kötü uç gözlemlerin olduğu durumlarda EM-PRM, EM-PLS'den daha iyi sonuçlar vermektedir.

3.4.7. Kayıp Veri Durumunda Sağlam Kestirimler ile İlgili Diğer Çalışmalar

Kayıp veri durumunda yapılan öneriler ve sağlam çalışmalar genel olarak yukarıdaki gibidir. Son yıllarda bilgisayarın ve yazılımların gelişmesiyle aktarılan yöntemlere alternatif öneriler getirilmektedir.

Devlin v. d. (1981), aykırı değerlerin bulunduğu veri kümelerinde Monte Carlo yöntemini kullanarak sağlam kestiricileri temel bileşenler analizi (TBA)'nde etkinliklerini karşılaştırmıştır. Bu çözümlemeyi yaparken karşılaştırdığı sağlam kestiriciler için veri kümesinde %10 kayıp veri oluşturarak kestiricilerin etkinliklerindeki değişiklikleri incelemiştir. Ancak çalışmada kayıp veri durumu için liste bazında veri silme yöntemi kullanmıştır. Veri silme yönteminin etkin olmadığını ve birçok götürüsünün olduğu önceden açıklanmıştı.

Çok değişkenli verinin çözümlenebilmesi için TBA önemli bir yöntemdir. Ancak kayıp veri ve aykırı değer bulunması durumunda çözümlemeleri etkili değildir. Smoliñski v. d. (2002), kayıp verinin elde edilmesi için McLaachlan ve Krishnan (1997) tarafından açıklanan TBA'da EM algoritmasının iteratif adımlarını ve aykırı değerlerin etkisini önlemek için Croux ve Ruiz-Gazen (1996) tarafından açıklanan sağlam TBA (RobPCA) yöntemini birleştirmişlerdir. Ayrıca kayıp verinin çözümlenmesi için gereken başlangıç değerlerinin aykırı değerlerden etkilenmesini engellemek ve daha sağlam bir sonuç elde edebilmek için Walczak (1995a,1995b) tarafından açıklanan sağlam kısmi EKK (RobPLS-partial least squares) yöntemini kullanılmışlardır. Bu sayede hem başlangıç değerlerinin aykırı değerlerden etkilenmesini önlemişlerdir hem de model tabanlı RobPLS yönteminden elde edilen değerleri başlangıç değeri olarak atayıp sağlam TBA'yı EM algoritmasının adımlarıyla açıklamışlardır.

Yoon v.d. (2007), mikroçiplerle ilgili çalışmalarda kullanılan yöntemleri karşılaştırmışlardır. En yakın komşu yöntemi için seçilecek en yakın gözlemlerin ilişki katsayıları ve seçilen k kümenin sayısından etkilenmesi sorununa değinmişlerdir. Ayrıca yerel EKK (LLSimpute) çözümlemesi yapılırken yine k kümenin seçilmesini, bir kayıp için bile k kadar küme üzerinden işlem yapılmasını, etkin kestirimlerin k küme seçimi ile bağlantılı olmasını sorun olarak açıklamışlardır. Sınırlamaları engellemek amacıyla TB'leri kullanarak sağlam EKK (RLSP) yöntemini önermişlerdir. LLSimpute yöntemindeki gibi, ele alınan

gözlemleri kullanıp elde edilen TB'lerle kantil (Quantile) regresyon çözümlemesi yapmışlar ve bu çözümlemeyle kayıp verinin yüklenebileceğini önermişlerdir. Yaptıkları benzetim çalışmasıyla yakın komşu yöntemi ve LLSimpute yöntemlerinden daha etkin sonuçlar elde ettiklerini görmüşlerdir.

Serneels ve Verdonck (2008), hem aykırı değer hem de kayıp değerin bulunduğu veri kümelerinde TBA uygulamasını esas almışlardır. Çalışmalarında iki farklı yaklaşım kullanmışlardır. Birinci yaklaşımda amaç, elde edilen veri üzerinden kovaryans matrisinde ayrışımı kullanmaktır. Bu çözümleme, değişken sayısı gözlem sayısını aştığında sonuç vermemektedir. İkinci yaklaşımda ise kayıp gözlem içeren veri kümelerine ER algoritması uyarlanarak sağlam TBA kullanılması önerilmiştir. TBA'nin sağlam yöntemlerle çözümlenebilmesi için kovaryans matrisini sağlamlaştırma, Huber (1985) tarafından önerilen projeksiyon izleme (projection pursuit), kovaryans matrisinin sağlamlaştırma ve projeksiyon izlemenin birleşimi, sağlam alt uzay kestirimi ve küresel ve eliptik TBA kullanımlarını açıklamışlardır. Çalışmanın sonunda yapılan geniş bir benzetim çalışmasıyla ER yaklaşımıyla oluşturulan sağlam TBA'nın her türlü veriye uygun olduğunu da göstermişlerdir.

Serneels ve Verdonck (2009), Temel Bileşenler Regresyonu (TBR) ile ER algoritmasının adımlarını oluşturmuşlardır. TBR için yapmış oldukları yöntem daha önce TBA için ER algoritmasının iteratif adımlarına uyarlama çalışması ile benzerdir. İteratif ER algoritma mantığını TBR için oluşturmuşlardır. Bu algoritma ile oluşturulan çok değişkenli regresyon yöntemi kayıp veri sorununu aşabilen ve aykırı değere karşı duyarsız olan ilk yöntemdir.

Hron v.d. (2010), birleştirilmiş veride (compositional data), Aitchison uzaklığından yararlanarak en yakın komşu yöntemi ile yükleme yapmışlardır. Bu yöntemde birleşen parçaların toplam büyüklüğü için kayıp veri kestirimi yapılmalıdır. Ayrıca çalışmada iteratif olarak bir regresyon yöntemi uygulamışlardır. Dönüştürülmüş uzay üzerinde iyi performans sağlayan ve veri yapısına göre klasik ya da sağlam regresyon yöntemleri kullanarak sırayla bir değişkeni bağımlı diğerlerini bağımsız yapacak şekilde modeller kurmuşlardır. Çalışmanın sonunda yapmış oldukları benzetim çalışmasında önerdikleri yöntemin standart veri yükleme yöntemlerinden daha iyi sonuçlar verdiğini, veri kümesinde aykırı değerler bulunduğunda iteratif

adımlardaki regresyon modelinin sağlam olmasının sonuçları daha etkin hale getirdiğini açıklamışlardır.

Son dönemlerde yapılan çalışmalara ve konularına bakıldığında aykırı değerlerle birlikte kayıp veri sorunun çözümlenmeye çalışılması ilgi alanı olmuştur. Özellikle EM ve ER algoritmalarının iteratif adımları yöntemlerdeki aykırı değer ve kayıp veri sorunlarını çözümlemede ilham kaynağı olmaktadır.

DÖRDÜNCÜ BÖLÜM

4. UYGULAMA

3. bölümde verilen kayıp veri yöntemlerini ve sağlam kestirimlerin bu yöntemlerde kullanımını incelenmek üzere 2 uygulama yapılmıştır: Birinci uygulamada, X veri matrisi oluşturulmuş ve bu veri kümesi %10 ve %20 bozulmuştur. Daha sonra bozulmamış, %10 bozulmuş ve %20 bozulmuş veri matrislerinden %5 ve %10 RK olacak şekilde değerler silinmiştir. İkinci uygulamada ise aykırı değerler içeren regresyon veri kümesinde %10 RK oluşturulmuş ve klasik veri yükleme, sağlam veri yükleme yöntemleri karşılaştırılmıştır. Bu bölümde çalışmanın amacına uygun olacak şekilde "R" paket programı kullanılmıştır.

4.1. Kayıp Veri Durumunda Sağlam Ortalama Vektörü ve Kovaryans Matrisi Kestirimleri

Uygulama 1 için veri kümesi çok değişkenli standart normal dağılımdan ($X \sim \boldsymbol{MN}(\mathbf{0}, \boldsymbol{I})$) elde edilmiştir. Bu verinin %10 ve %20'lik kısmı $X \sim \boldsymbol{MN}(\mathbf{5}, \boldsymbol{I})$ çok değişkenli normal dağılımından türetilerek veri matrisi bozulmuştur. Çözümlemelerde incelemeler yapılırken bozulmamış veri kümesinde, %10 ve %20 bozulmuş veri kümelerinde sırasıyla %5 ve %10 RK oluşturulmuştur. Ayrıca %30 bozulmuş veri kümesi ve %20 RK durumları da incelenmiş, farklı bir yorum getirmediği için çalışmada yer verilmemiştir.

Elde edilen veri matrislerinde veri silme yöntemleri, ortalama ve ortanca ile veri yükleme yöntemleri, EM ve ER Algoritmaları kullanılarak ortalama vektörü ve kovaryans matrisi elde edilmiştir. Ayrıca aynı çözümlemede sağlam veri yükleme yöntemi kullanılarak elde edilen tam veri matrisinin ortalama vektörü ve kovaryans matrisi MCD, OGK, SD ve BS kestiricileri yardımıyla ayrı ayrı elde edilerek yorumlanmıştır. Sonuçları verildiği çizelgelerde "LBVS", liste bazında veri silme; "ÇBVC", çiftler bazında veri silme; "Ort. Yük.", ortalama ile veri yükleme, "Ortanca Yük.", ortanca ile veri yükleme, "EM Alg.", EM algoritmasını, "ER Alg.", ER algoritmasını göstermektedir.

Çizelge 4.1, %5 oranında RK düzeneğine uyacak şekilde verisi silinen X matrisine ilişkin elde edilen ortalama vektörü ve kovaryans matrisi kestirimlerini göstermektedir. Çizelge 4.1 ortalama vektörü sonuçları incelendiğinde çiftler bazında veri silme yöntemi, ortalama ve ortanca ile veri yükleme yöntemleri

sonuçları benzerdir. Bozulum bulunmadığından EM ve ER algoritmalarının sonuçları aynıdır. Ortalama ile veri yükleme yöntemi sonucunda elde edilen ortalama vektörü kestirimi, çiftler bazında veri silme yönteminin ortalama vektörü kestirimi ile aynıdır. Ancak, ortalama ile veri yükleme yönteminin daha düşük kovaryans kestirimi verdiği görülmektedir. Çiftler bazında veri silme yöntemi kayıp veriye herhangi bir değer yüklemeden işlem yapmaktadır. Ortalama ile veri yükleme yönteminde ise kayıp veri değerleri ortalama ile yüklendiğinden varyansta azalma meydana gelmektedir. Azalma, gözlenmemiş değerlerde merkezi değer yüklenmesinden kaynaklanmaktadır.

Çizelge 4.1. Bozulmamış, %5'i Kayıp Veride Klasik ve Sağlam Veri Yükleme Yöntemleri ile Elde Edilen Ortalama Vektörü ve Kovaryans Matrisi

	$\bar{X}^T$	$Kov(X)$					$\bar{X}^T$	$Kov(X)$			
LBVS	-0,030	**1,018**	-0,062	-0,077	0,069	**ER Alg.**	-0,046	**1,088**	-0,065	-0,065	0,025
	0,033	-0,062	**1,108**	-0,026	-0,162		0,096	-0,065	**1,011**	-0,001	-0,183
	-0,086	-0,077	-0,026	**1,040**	-0,168		-0,063	-0,065	-0,001	**0,996**	-0,148
	-0,038	0,069	-0,162	-0,168	**1,063**		-0,063	0,025	-0,183	-0,148	**1,058**
ÇBVS	-0,045	**1,090**	-0,067	-0,059	0,025	**MCD**	-0,090	**1,190**	-0,195	0,011	0,206
	0,096	-0,067	**1,020**	-0,008	-0,193		0,160	-0,195	**1,127**	0,065	0,019
	-0,065	-0,059	-0,008	**0,999**	-0,142		-0,070	0,011	0,065	**1,188**	-0,329
	-0,062	0,025	-0,193	-0,142	**1,061**		-0,119	0,206	0,019	-0,329	**0,922**
Ort. Yük.	-0,045	**1,077**	-0,062	-0,057	0,023	**OGK**	-0,133	**0,884**	-0,206	0,068	0,105
	0,096	-0,062	**0,980**	-0,003	-0,176		0,126	-0,206	**0,847**	0,032	-0,060
	-0,065	-0,057	-0,003	**0,905**	-0,127		-0,052	0,068	0,032	**0,941**	-0,226
	-0,062	0,023	-0,176	-0,127	**1,036**		-0,095	0,105	-0,060	-0,226	**0,902**
Ortanca Yük.	-0,045	**1,077**	-0,064	-0,057	0,025	**SD**	-0,147	**1,014**	-0,243	0,164	0,060
	0,096	-0,064	**0,981**	-0,002	-0,177		0,097	-0,243	**0,976**	0,140	-0,040
	-0,065	-0,057	-0,002	**0,905**	-0,126		-0,042	0,164	0,140	**1,180**	-0,365
	-0,062	0,025	-0,177	-0,126	**1,036**		-0,098	0,060	-0,040	-0,365	**1,152**
EM Alg.	-0,046	**1,088**	-0,065	-0,065	0,025	**BS**	-0,127	**1,080**	-0,200	0,092	0,103
	0,096	-0,065	**1,011**	-0,001	-0,183		0,129	-0,200	**0,964**	0,082	-0,042
	-0,063	-0,065	-0,001	**0,996**	-0,148		-0,052	0,092	0,082	**1,107**	-0,297
	-0,063	0,025	-0,183	-0,148	**1,058**		-0,118	0,103	-0,042	-0,297	**1,029**

Sağlam veri yükleme sonucunda yapılan kestirimlerde, BS kestiricisi gerçek değerlere yakın sonucu verse de hepsi ortalama vektörü açısında diğer yöntemlere göre kötü sonuç vermiştir. OGK kestiricisi ise ortalama vektörünü kötü kestirmesine ek olarak kovaryans matrisi değerini de beklenenden daha küçük elde etmiştir. OGK kestiricisinin kovaryans matrisini küçük elde edilmesinin

sebebi, kovaryans matrisini kesilmiş gözlem sayısı üzerinden hesaplanan ikili değişkenlerden elde etmesidir. Bozulma olmadığından en büyük sapmaya sahip değerlerin çıkartılması varyansın küçük kestirilmesine neden olmaktadır.

Çizelge 4.2, %10 oranında RK düzeneğine uyacak şekilde verisi silinen X matrisine ilişkin elde edilen ortalama vektörü ve kovaryans matrisi kestirimlerini göstermektedir. Çizelge 4.1 ve Çizelge 4.2 karşılaştırıldığında %5 RK olduğu durumda liste bazında veri silme yönteminin ortalama vektörü kestirimi ile %10 RK olduğu durumunda liste bazında veri silme yönteminin ortalama vektörü kestirimi sonuçları arasında diğer yöntemlerden daha fazla farklılık oluştuğu görülmüştür.

Çizelge 4.2. Bozulmamış, %10'u Kayıp Veride Klasik ve Sağlam Veri Yükleme Yöntemleri ile Elde Edilen Ortalama Vektörü ve Kovaryans Matrisi

	$\bar{X}^T$	$Kov(X)$					$\bar{X}^T$	$Kov(X)$			
LBVS	-0,170	**1,131**	-0,098	-0,174	0,077	**ER Alg.**	-0,057	**1,119**	-0,231	-0,111	0,005
	0,139	-0,098	**0,769**	-0,002	0,029		0,081	-0,231	**0,851**	-0,007	-0,024
	-0,078	-0,174	-0,002	**1,146**	-0,051		-0,017	-0,111	-0,007	**1,105**	-0,129
	-0,072	0,077	0,029	-0,051	**0,887**		-0,080	0,005	-0,024	-0,129	**0,954**
ÇBVS	-0,057	**1,124**	-0,245	-0,104	0,001	**MCD**	-0,171	**1,018**	-0,507	0,402	-0,093
	0,095	-0,245	**0,859**	-0,021	-0,041		0,111	-0,507	**1,177**	0,004	-0,014
	-0,016	-0,104	-0,021	**1,104**	-0,117		0,060	0,402	0,004	**1,303**	-0,312
	-0,078	0,001	-0,041	-0,117	**0,960**		-0,189	-0,093	-0,014	-0,312	**0,955**
Ort. Yük.	-0,057	**1,019**	-0,185	-0,088	-0,007	**OGK**	-0,136	**0,816**	-0,243	0,103	0,086
	0,095	-0,185	**0,755**	-0,011	-0,029		0,083	-0,243	**0,663**	0,017	0,025
	-0,016	-0,088	-0,011	**1,026**	-0,098		0,040	0,103	0,017	**0,849**	-0,251
	-0,078	-0,007	-0,029	-0,098	**0,858**		-0,129	0,086	0,025	-0,251	**0,688**
Ortanca Yük.	-0,060	**1,019**	-0,181	-0,088	-0,011	**SD**	-0,199	**1,015**	-0,423	0,428	-0,009
	0,105	-0,181	**0,756**	-0,007	-0,031		0,130	-0,423	**0,848**	-0,017	0,003
	-0,017	-0,088	-0,007	**1,026**	-0,100		0,054	0,428	-0,017	**1,242**	-0,153
	-0,086	-0,011	-0,031	-0,100	**0,859**		-0,132	-0,009	0,003	-0,153	**0,706**
EM Alg.	-0,057	**1,119**	-0,231	-0,111	0,005	**BS**	-0,172	**0,905**	-0,327	0,265	0,055
	0,081	-0,231	**0,851**	-0,007	-0,024		0,134	-0,327	**0,806**	-0,003	-0,013
	-0,017	-0,111	-0,007	**1,105**	-0,129		0,031	0,265	-0,003	**1,084**	-0,242
	-0,080	0,005	-0,024	-0,129	**0,954**		-0,103	0,055	-0,013	-0,242	**0,822**

Liste bazında veri silme yöntemi kayıp veri oranı arttıkça dikdörtgen biçimi sağlamak için kayıp verisi bulunan tüm gözlemleri sildiğinden veri kümesi küçülmüştür. Bu durum sonuçların diğer yöntemlere göre daha farklı çıkmasına sebep olmuştur. Veride bozulma olmadığında en iyi sonuçlar çiftler bazında veri silme yöntemi ile EM ve ER algoritmalarından elde edilmektedir. Çizelge 4.2'de

ortalama ile veri yükleme yöntemi, çiftler bazında veri silme yöntemiyle aynı ortalama vektörü kestirimini vermiştir. Ancak ortalama ile veri yükleme yönteminin kovaryans matrisi kestirimi, merkeze yakın değerler yüklenmesinden etkilenerek çiftler bazında veri silme yönteminin kovaryans matrisi kestiriminden daha düşük elde edilmiştir. Aynı durum ortanca ile veri yükleme yönteminde de görülebilir. Üstelik kovaryans matrisi kestirimlerindeki bu düşüşler, Çizelge 4.1'deki düşüşlerden daha fazla olmuştur. Bu duruma, kayıp veri oranının artmasından (%5'ten %10'a çıkmasından) dolayı kayıp değerlerin merkeze yakın olan (ortalama ve ortanca) değerlerle daha çok yüklenmesi sebep olmaktadır.

Veri kümesi bozulduğunda yöntemleri karşılaştırmak için yukarıda elde edilen X matrisi %10 oranında bozulmuştur. Veri matrisi %10 oranında bozulduktan sonra %5 oranında RK düzeneğine uyacak şekilde verisi silinen X matrisi kestirimleri Çizelge 4.3'te verilmiştir. Veri silme yöntemleri, ortalama ve ortanca ile veri yükleme yöntemleri, EM ve ER algoritmaları ile elde edilen kestirimler verideki bozulumdan etkilenmiştir. Ortanca ile veri yükleme yöntemi, ortalama ile veri yükleme yöntemine göre kestirimler bakımından bozulumdan daha az etkilenmiştir. Ortalama ile veri yükleme yöntemi ve çiftler bazında veri silme yönteminin ortalama vektörleri için elde ettiği kestirimler benzer şekildedir.

Çizelge 4.1 ve Çizelge 4.2'de bozulma yokken ortalama ve ortanca ile veri yükleme yöntemleri arasında fark yoktur. Bozulma olduğunda ise (Çizelge 4.3) ortalama ile veri yükleme yöntemi daha kötü sonuçlar vermiştir. Aynı şekilde bozulmanın olmadığı durumda EM ve ER algoritmalarından elde edilen sonuçlar benzerken bozulma olduğunda ER algoritması kestirimleri EM algoritması kestirimlerine göre daha az etkilenmiştir. Klasik yöntemler veri kümesindeki bozulmalarda iyi sonuç vermemiştir. ER algoritmasının bozulma noktasının düşük olması sonuçlarının yüksek bozulma noktasına sahip kestiricilerden daha tutarlı olmasını engellemiştir. Yüksek bozulma noktasına sahip sağlam kestiriciler klasik yöntemlere göre oldukça iyi sonuçlar vermiştir.

Sağlam yöntemlerden en iyi sonuçlar MCD ve BS kestiricilerinden elde edilmiştir. SD kestiricisi de OGK kestiricisinden daha iyi sonuç vermiştir. OGK kestiricisi bozulma olduğu durumda da varyans kestirimini beklenenden daha küçük vermektedir.

EM ve ER algoritmalarındaki benzerlik aykırı değer olmadığından kayıp veri oranı arttırılmasına rağmen devam etmektedir. Sağlam veri yükleme sonrasındaki kestirimler, veride herhangi bir bozulma olmadığından klasik yöntemlere göre iyi sonuçlar vermemiştir.

Çizelge 4.3. %10 bozulmuş, %5'i Kayıp Veride Klasik ve Sağlam Veri Yükleme Yöntemleri ile Elde Edilen Ortalama Vektörü ve Kovaryans Matrisi

	$\bar{X}^T$	$Kov(X)$					$\bar{X}^T$	$Kov(X)$			
LBVS	0,530	**3,175**	2,169	2,175	2,121	**ER Alg.**	0,335	**1,601**	0,755	0,756	0,758
	0,534	2,169	**3,408**	2,331	2,372		0,295	0,755	**1,875**	0,885	0,843
	0,593	2,175	2,331	**3,056**	2,116		0,375	0,756	0,885	**1,595**	0,658
	0,495	2,121	2,372	2,116	**2,911**		0,325	0,758	0,843	0,658	**1,442**
ÇBVS	0,503	**3,214**	2,342	2,303	2,272	**MCD**	0,017	**1,005**	0,072	0,005	-0,109
	0,520	2,342	**3,854**	2,748	2,448		0,008	0,072	**1,433**	0,150	0,144
	0,579	2,303	2,748	**3,350**	2,199		0,070	0,005	0,150	**1,192**	-0,103
	0,517	2,272	2,448	2,199	**3,011**		-0,031	-0,109	0,144	-0,103	**0,825**
Ort. Yük.	0,503	**2,841**	1,929	1,909	1,849	**OGK**	0,080	**0,737**	0,041	0,079	0,077
	0,520	1,929	**3,511**	2,402	2,054		-0,062	0,041	**0,994**	0,092	-0,003
	0,579	1,909	2,402	**3,015**	1,821		0,110	0,079	0,092	**0,882**	-0,020
	0,517	1,849	2,054	1,821	**2,633**		-0,031	0,077	-0,003	-0,020	**0,639**
Ortanca Yük.	0,490	**2,845**	1,934	1,910	1,851	**SD**	0,061	**0,881**	0,057	-0,006	-0,012
	0,506	1,934	**3,517**	2,400	2,033		-0,069	0,057	**1,294**	0,154	-0,007
	0,575	1,910	2,400	**3,016**	1,810		0,061	-0,006	0,154	**1,149**	-0,196
	0,499	1,851	2,033	1,810	**2,639**		-0,059	-0,012	-0,007	-0,196	**0,835**
EM Alg.	0,523	**3,055**	2,183	2,103	2,021	**BS**	0,046	**0,973**	0,014	0,024	0,021
	0,526	2,183	**3,518**	2,385	2,316		-0,083	0,014	**1,194**	0,089	-0,009
	0,581	2,103	2,385	**2,998**	2,003		0,040	0,024	0,089	**1,174**	-0,118
	0,546	2,021	2,316	2,003	**2,849**		-0,032	0,021	-0,009	-0,118	**0,839**

Çizelge 4.4, %10 oranında bozulduktan sonra %10 oranında RK düzeneğine uyacak şekilde verisi silinen X matrisine ilişkin elde edilen ortalama vektörü ve kovaryans matrisi kestirimlerini göstermektedir. Çizelgeye göre yapılan incelemede sağlam veri yükleme yöntemlerinin klasik yöntemlere göre büyük bir üstünlüğü söz konusudur. Özellikle yüksek bozulma noktasına sahip MCD, OGK, SD ve BS kestiricileri kayıp oranının artmasından biraz etkilenseler de diğer kestiricilere göre iyi sonuçlar verdiği görülmüştür.

Veri silme yöntemleri, ortalama ve ortanca veri yükleme yöntemleri ve EM algoritması uygulandıktan sonra elde edilen kestirimlerin veri matrisindeki bozulmadan etkilendikleri burada da görülmektedir. Verideki bozulmanın %5'ten

%10'a çıkması ve kayıp oranının artması ortanca ile veri yükleme yönteminin ortalama ile veri yükleme yönteminden daha iyi sonuçlar vermesini sağlamaktadır.

Çizelge 4.4. %10 bozulmuş, %10'u Kayıp Veride Klasik ve Sağlam Veri Yükleme Yöntemleri ile Elde Edilen Ortalama Vektörü ve Kovaryans Matrisi

	$\bar{X}^T$	$Kov(X)$					$\bar{X}^T$	$Kov(X)$			
LBVS	0,691	**2,856**	2,606	2,347	2,270	**ER Alg.**	0,436	**1,600**	0,907	0,999	0,905
	0,470	2,606	**4,121**	2,881	2,714		0,283	0,907	**2,074**	0,962	0,978
	0,648	2,347	2,881	**3,163**	2,297		0,465	0,999	0,962	**1,715**	0,739
	0,614	2,270	2,714	2,297	**3,346**		0,331	0,905	0,978	0,739	**1,569**
ÇBVS	0,588	**3,353**	2,591	2,889	2,246	**MCD**	0,031	**1,113**	0,166	0,156	0,156
	0,514	2,591	**4,102**	3,230	2,837		0,069	0,166	**0,669**	0,143	0,051
	0,668	2,889	3,230	**3,672**	2,503		0,027	0,156	0,143	**1,424**	-0,258
	0,547	2,246	2,837	2,503	**3,120**		-0,096	0,156	0,051	-0,258	**1,097**
Ort. Yük.	0,588	**2,704**	1,788	1,952	1,560	**OGK**	0,139	**0,706**	-0,010	0,183	0,148
	0,514	1,788	**3,295**	2,130	1,948		-0,112	-0,010	**0,849**	0,032	0,023
	0,668	1,952	2,130	**2,867**	1,704		0,184	0,183	0,032	**0,858**	-0,011
	0,547	1,560	1,948	1,704	**2,564**		-0,067	0,148	0,023	-0,011	**0,666**
Ortanca Yük.	0,550	**2,715**	1,797	1,942	1,558	**SD**	0,118	**0,952**	0,090	0,129	0,170
	0,442	1,797	**3,325**	2,146	1,960		-0,016	0,090	**0,757**	0,061	-0,014
	0,613	1,942	2,146	**2,890**	1,721		0,096	0,129	0,061	**1,360**	0,013
	0,506	1,558	1,960	1,721	**2,580**		-0,053	0,170	-0,014	0,013	**0,953**
EM Alg.	0,602	**2,998**	2,292	2,228	1,984	**BS**	0,117	**0,897**	0,013	0,126	0,102
	0,493	2,292	**3,716**	2,408	2,325		-0,069	0,013	**1,011**	0,016	0,015
	0,654	2,228	2,408	**3,012**	1,950		0,080	0,126	0,016	**1,127**	-0,020
	0,544	1,984	2,325	1,950	**2,821**		-0,017	0,102	0,015	-0,020	**0,924**

Verideki bozulma oranının %10 olduğu ve kayıp değer oranının %5'ten %10'a çıkartıldığında ortalama ve ortanca ile veri yükleme yöntemlerinin elde etmiş olduğu kovaryans matrisi kestirimlerindeki değişim Çizelge 4.3 ve Çizelge 4.4.'den karşılaştırılabilir. Bu karşılaştırmaya göre, merkezi değerler yüklemelerinden dolayı çiftler bazıda veri silme yöntemine göre daha küçük varyansları olduğu görülmektedir. ER algoritması, diğer sağlam kestiricilere nazaran verideki bozulmadan daha çok etkilenmiştir. Ancak, EM algoritmasından daha iyi sonuç vermiştir. Sağlam yöntemlerden en iyi sonucu MCD kestiricisi vermiştir. BS ve SD kestiricileri, OGK kestiricisinden daha iyi sonuçlar vermiştir.

Veri matrisi %20 oranında bozulduktan sonra %5 oranında RK düzeneğine uyacak şekilde verisi silinen X matrisine ilişkin elde edilen ortalama vektörü ve kovaryans matrisi kestirimleri Çizelge 4.5'te verilmiştir.

Çizelge 4.5. %20 bozulmuş, %5'i Kayıp Veride Klasik ve Sağlam Veri Yükleme Yöntemleri ile Elde Edilen Ortalama Vektörü ve Kovaryans Matrisi

	$\bar{X}^T$	$Kov(X)$					$\bar{X}^T$	$Kov(X)$			
LBVS	1,048	**5,228**	4,464	4,793	4,634	**ER Alg.**	0,784	**5,192**	4,316	4,489	4,401
	1,283	4,464	**5,316**	4,961	4,532		1,031	4,316	**5,103**	4,564	4,223
	1,232	4,793	4,961	**6,446**	5,170		0,928	4,489	4,564	**6,041**	4,744
	1,309	4,634	4,532	5,170	**5,613**		1,064	4,401	4,223	4,744	**5,380**
ÇBVS	0,869	**5,692**	4,894	5,705	5,537	**MCD**	-0,165	**1,310**	0,362	-0,053	0,187
	1,034	4,894	**5,761**	5,570	5,453		0,102	0,362	**1,568**	-0,004	-0,055
	1,036	5,705	5,570	**6,968**	6,525		-0,133	-0,053	-0,004	**1,224**	-0,030
	1,168	5,537	5,453	6,525	**6,615**		0,051	0,187	-0,055	-0,030	**1,452**
Ort. Yük.	0,869	**4,688**	3,660	4,020	3,885	**OGK**	-0,122	**0,857**	0,212	-0,037	0,024
	1,034	3,660	**4,550**	3,968	3,650		0,099	0,212	**1,050**	-0,010	-0,050
	1,036	4,020	3,968	**5,538**	4,227		-0,142	-0,037	-0,010	**0,887**	-0,002
	1,168	3,885	3,650	4,227	**4,773**		0,181	0,024	-0,050	-0,002	**0,732**
Ortanca Yük.	0,829	**4,713**	3,677	4,177	3,954	**SD**	-0,189	**1,099**	0,257	0,084	0,153
	1,010	3,677	**4,564**	4,027	3,698		0,151	0,257	**1,552**	0,151	-0,065
	0,978	4,177	4,027	**5,583**	4,372		-0,152	0,084	0,151	**1,331**	0,097
	1,102	3,954	3,698	4,372	**4,813**		0,211	0,153	-0,065	0,097	**0,987**
EM Alg.	0,803	**4,814**	3,976	4,206	4,044	**BS**	-0,164	**1,145**	0,262	-0,023	0,061
	1,040	3,976	**4,771**	4,289	3,914		0,113	0,262	**1,449**	0,000	-0,112
	0,949	4,206	4,289	**5,687**	4,417		-0,112	-0,023	0,000	**1,075**	-0,016
	1,076	4,044	3,914	4,417	**4,948**		0,131	0,061	-0,112	-0,016	**1,136**

Klasik kayıp veri yöntemleri sonucundaki kestirimler bozulma oranından etkilenmiştir. Aynı şekilde bozulma noktası düşük olan ER algoritması da %20 bozulmaya karşı dirençli olmadığından kestirimleri etkilenmiştir.

Sağlam kestiricilerden MCD ve BS kestiricileri diğerlerine göre daha iyi sonuçlar vermiştir ancak %10'u bozuk veriye göre ortalama vektörü sıfırdan biraz daha uzaklaşmıştır. Sağlam yöntemler çözümleme bakımından karşılaştırıldığında klasik yöntemlerden oldukça üstündür.

EM algoritması genellikle çok iyi veri yüklemesine rağmen ML kestiricileri üzerinden işlem yapıyor olması aykırı değerlerden fazlaca etkilendiğini göstermektedir.

Çizelge 4.6 verinin %20 oranında bozulumundan sonra %10 oranında RK düzeneğine uyacak şekilde verisi silinen X matrisine ilişkin elde edilen ortalama vektörü ve kovaryans matrisi kestirimlerini göstermektedir.

Çizelge 4.6. %20 bozulmuş, %10'u Kayıp Veride Klasik ve Sağlam Veri Yükleme Yöntemleri ile Elde Edilen Ortalama Vektörü ve Kovaryans Matrisi

	$\bar{X}^T$	$Kov(X)$					$\bar{X}^T$	$Kov(X)$			
LBVS	0,802	**4,831**	3,999	4,148	3,904	**ER Alg.**	0,688	**4,942**	4,179	4,004	4,112
	1,157	3,999	**4,714**	4,199	3,793		1,025	4,179	**5,065**	4,173	4,110
	1,027	4,148	4,199	**5,696**	4,435		0,894	4,004	4,173	**5,348**	4,328
	1,089	3,904	3,793	4,435	**4,932**		1,079	4,112	4,110	4,328	**5,270**
ÇBVS	0,787	**5,591**	5,492	4,334	4,989	**MCD**	-0,163	**1,393**	0,358	-0,205	0,019
	1,082	5,492	**6,247**	4,718	5,243		0,060	0,358	**1,685**	-0,030	-0,263
	0,812	4,334	4,718	**5,501**	4,631		-0,083	-0,205	-0,030	**1,162**	-0,080
	1,163	4,989	5,243	4,631	**6,229**		0,196	0,019	-0,263	-0,080	**1,473**
Ort. Yük.	0,787	**4,571**	3,881	3,166	3,149	**OGK**	-0,294	**1,088**	0,340	-0,139	0,152
	1,082	3,881	**4,769**	3,344	3,213		0,141	0,340	**1,008**	-0,004	-0,126
	0,812	3,166	3,344	**4,500**	3,108		-0,126	-0,139	-0,004	**0,831**	-0,042
	1,163	3,149	3,213	3,108	**4,335**		0,207	0,152	-0,126	-0,042	**0,826**
Ortanca Yük.	0,733	**4,600**	3,951	3,108	3,213	**SD**	-0,295	**1,201**	0,631	-0,059	0,262
	1,037	3,951	**4,796**	3,296	3,235		0,109	0,631	**1,511**	-0,005	-0,032
	0,760	3,108	3,296	**4,531**	3,061		-0,056	-0,059	-0,005	**1,155**	-0,093
	1,084	3,213	3,235	3,061	**4,381**		0,208	0,262	-0,032	-0,093	**1,124**
EM Alg.	0,740	**4,786**	4,061	3,981	3,973	**BS**	-0,261	**1,215**	0,409	-0,092	0,124
	1,064	4,061	**4,913**	4,123	3,961		0,087	0,409	**1,337**	-0,041	-0,158
	0,946	3,981	4,123	**5,264**	4,215		-0,042	-0,092	-0,041	**0,980**	-0,065
	1,123	3,973	3,961	4,215	**5,000**		0,148	0,124	-0,158	-0,065	**1,019**

Klasik kayıp veri yöntemleri kullanılarak yapılan çözümlemelerin bozulma oranındaki artıştan etkilendikleri önceki çizelgelerde görüldü. %20 bozulma oranı değiştirilmeden kayıp oranı %5'ten %10'a çıkartıldığında ise kestirimler Çizelge 4.5 ve Çizelge 4.6'dan karşılaştırılabilir. Yapılan inceleme sonucunda kestirimlerin kayıp oranının arttırılmasından etkilendikleri görülmektedir. Ancak bu etkilenme Çizelge 4.3 ve Çizelge 4.4 arasındaki kayıp veri oranındaki artıştan daha önemsizdir. Buradan da yöntemlerin kayıp oranının arttırılmasından çok bozulma oranından etkilendikleri görülmektedir. Kayıp oranının arttırılması kestirimleri karşılaştırma da farklılık yaratmamıştır. Sağlam veri yükleme ile elde edilen tam verinin sağlam kestirimleri klasik veri yükleme yöntemi ile elde edilen kestirimlerden daha iyidir.

Klasik veri yükleme yöntemlerinin yanında ER algoritmasının sonuçları da iyi değildir. MCD ve BS kestiricileri en iyi sonuçları vermiştir. Bozulma oranının %10'dan %20'ye çıkmış olması kovaryans matrisinin sağlam kestirimleri üzerinde

etkili olmamıştır. Ancak klasik kestirimler bozulma oranın arttırılmasından çok fazla etkilenmiştir.

Sonuç olarak, bozulmanın olmadığı durumlarda klasik veri yükleme yöntemleri, sağlam yöntemlere göre daha üstündür. Bozulmanın olmadığı durumlarda kayıp oranının arttırılması kestirimlerde değişikliklere neden olmaktadır. Bozulma oranının artması yüksek bozulma noktasına sahip kestiricilerde anlamlı bir etki yaratmasa da klasik veri yükleme yöntemlerinin ve yüksek bozulma noktasına sahip olmayan ER algoritmasının üzerinde etkilidir. Belli bir bozulma oranında kayıp veri oranının arttırılması yine kestirimleri etkilemektedir.

Bozulma oranının arttırılması özellikle klasik veri yükleme yöntemleri sonucunda elde edilen kovaryans kestirimlerinde büyük değişikliklere sebep olmaktadır. Ortalama ve ortanca ile veri yükleme yöntemleri, aynı oranda kayıp veriye uygulanan çiftler bazında veri silme yöntemine göre kovaryans matrisini daha küçük kestirmektedir. Kayıp oranı arttıkça kestirimler arasındaki fark artmaktadır. Ancak kovaryans kestirimlerindeki bu düşüklükler, bahsedilen klasik veri yükleme yöntemlerinin iyi kestirimlerinden değil merkeze yakın değerlerin boş gözelere yüklenmesinden kaynaklanmaktadır.

Bozulma durumlarında BS ve MCD yöntemi diğer yüksek bozulma noktasına sahip sağlam kestirimlerden ve klasik veri yükleme yöntemleriyle elde edilen kestirimlerden daha başarılı sonuçlar elde etmiştir. Bozulma durumlarında kayıp oranının artması sağlam kestirimleri karşılaştırmada bir farklılık yaratmamıştır.

4.2. Kayıp Veri Durumunda Klasik ve Sağlam Veri Yükleme Yöntemlerinin Regresyon Verisi için İncelenmesi

Tez çalışmasının 2. uygulaması olarak Hawkins, Bradu ve Kass (1984)'ın veri kümesi kullanılmıştır. Bu veri kümesi 14 aykırı değer içermektedir ve Çizelge 4.7'de verilmiştir.Veri kümesindeki değerler %10 oranında RK olacak şekilde silinmiştir. Bu uygulamada kayıp veri yükleme yöntemlerinden ortalama ile veri yükleme, ardışık veri yükleme, ER algoritması, sağlam veri yükleme yöntemi kullanılmıştır. Yüklenen veriyle gerçek değerleri arasındaki farklar ortalama veri yükleme hatası bakımından karşılaştırılmıştır. Veri kümesindeki RK olarak silinen değerler koyu ve altı çizilmiş şekilde verilmiştir.

Çizelge 4.7. Hawkins, Bradu ve Kass(1984)'ın Veri Kümesi

	x_1	x_2	x_3	y		x_1	x_2	x_3	y		x_1	x_2	x_3	y
1	10,10	19,60	28,30	9,70	**26**	0,90	3,30	2,50	-0,80	**51**	2,30	**1,50**	0,40	0,70
2	9,50	20,50	28,90	10,10	**27**	3,30	2,50	2,90	-0,70	**52**	3,30	**0,60**	1,20	-0,50
3	10,70	20,20	31,00	10,30	**28**	1,80	0,80	2,00	0,30	**53**	0,30	0,40	3,30	0,70
4	9,90	21,50	31,70	9,50	**29**	1,20	0,90	0,80	0,30	**54**	**1,10**	3,00	**0,30**	0,70
5	10,30	21,10	31,10	10,00	**30**	1,20	0,70	3,40	-0,30	**55**	0,50	2,40	0,90	0,00
6	10,80	20,40	29,20	10,00	**31**	3,10	1,40	1,00	0,00	**56**	1,80	3,20	0,90	0,10
7	10,50	20,90	29,10	10,80	**32**	0,50	2,40	**0,30**	-0,40	**57**	1,80	0,70	0,70	0,70
8	9,90	19,60	28,80	10,30	**33**	1,50	3,10	1,50	-0,60	**58**	**2,40**	3,40	1,50	-0,10
9	9,70	20,70	**31,00**	9,60	**34**	0,40	0,00	0,70	-0,70	**59**	**1,60**	2,10	3,00	-0,30
10	9,30	19,70	30,30	9,90	**35**	3,10	2,40	3,00	0,30	**60**	0,30	1,50	3,30	-0,90
11	11,00	24,00	35,00	-0,20	**36**	0,10	2,20	2,70	-1,00	**61**	0,40	3,40	3,00	-0,30
12	12,00	23,00	37,00	-0,40	**37**	0,10	3,00	2,60	-0,60	**62**	0,90	0,10	0,30	0,60
13	12,00	26,00	34,00	0,70	**38**	1,50	1,20	0,20	0,90	**63**	1,10	2,70	0,20	-0,30
14	11,00	34,00	34,00	0,10	**39**	2,10	0,00	1,20	-0,70	**64**	**2,80**	3,00	2,90	-0,50
15	**3,40**	2,90	2,10	-0,40	**40**	0,50	2,00	1,20	-0,50	**65**	2,00	0,70	2,70	0,60
16	3,10	2,20	0,30	0,60	**41**	**3,40**	**1,60**	2,90	-0,10	**66**	0,20	1,80	**0,80**	-0,90
17	0,00	1,60	0,20	-0,20	**42**	0,30	1,00	2,70	-0,70	**67**	1,60	2,00	1,20	-0,70
18	2,30	**1,60**	2,00	0,00	**43**	0,10	3,30	0,90	0,60	**68**	0,10	0,00	1,10	0,60
19	0,80	2,90	1,60	0,10	**44**	1,80	0,50	3,20	-0,70	**69**	2,00	0,60	0,30	0,20
20	3,10	3,40	2,20	0,40	**45**	1,90	0,10	0,60	-0,50	**70**	1,00	2,20	2,90	0,70
21	2,60	2,20	1,90	0,90	**46**	1,80	**0,50**	3,00	**-0,40**	**71**	2,20	2,50	2,30	0,20
22	**0,40**	**3,20**	3,83	**0,30**	**47**	3,00	0,10	0,80	-0,90	**72**	0,60	2,00	1,50	-0,20
23	2,00	2,30	0,80	-0,80	**48**	3,10	1,60	3,00	0,10	**73**	0,30	1,70	2,20	0,40
24	1,30	2,30	0,50	0,70	**49**	3,10	2,50	1,90	**0,90**	**74**	0,00	**2,20**	1,60	-0,90
25	1,00	0,00	0,40	-0,30	**50**	2,10	2,80	**2,90**	-0,40	**75**	0,30	0,40	2,60	0,20

RK olacak şekilde silinmiş olan değerlerin veri yükleme yöntemleri uygulandıktan sonraki değerleri Çizelge 4.8'de verilmiştir.

RK oluşturulduktan sonra aykırı değer olan 8. ve 9. gözlemlerdeki kayıp değerlere Çizelge 4.8'deki gibi yükleme yapıldığında ardışık veri yükleme yöntemiyle yüklenen değerler, ER algoritması ile yüklenen değerler ve sağlam veri ile yüklenen değerler gerçek veriye yakındır. Ancak ortalama ile veri yükleme yöntemi aykırı gözlemlerin kayıp değerlerini yüklemede başarısızdır.

Yapılan yüklemeleri karşılaştırmak amacıyla "ortalama veri yükleme hataları (OVYH)" kullanılmıştır.

Çizelge 4.8. RK Silinen Değerlerin Gerçek ve Yüklenen Değerleri

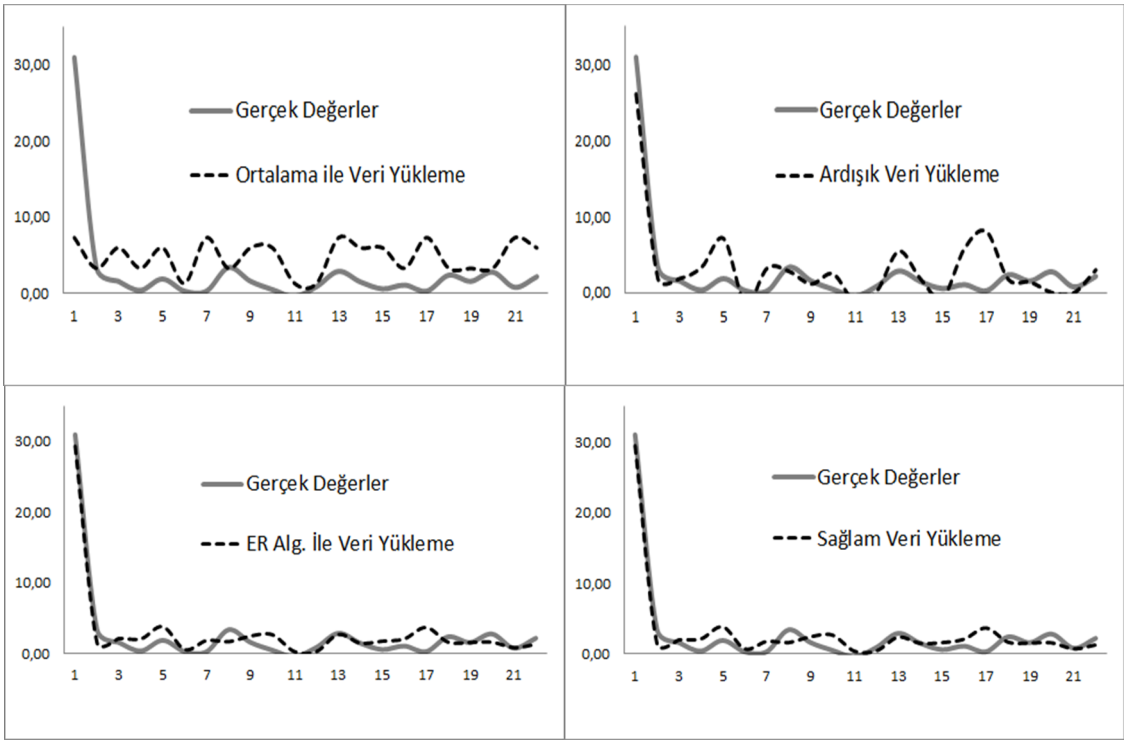
Bulunduğu Değişken	Gözlem Sırası	Gerçek Değeri	Ortalama İle Veri Yükleme	Ardışık Veri Yükleme	ER Alg. ile Yükleme	Sağlam Veri Yükleme
x_3	**9**	31,00	7,32	26,15	29,37	29,43
x_1	**15**	3,40	3,30	1,90	1,57	1,42
x_2	**18**	1,60	5,99	1,91	2,16	1,98
x_1	**22**	0,40	3,30	3,39	2,10	2,11
x_2	**22**	3,20	5,99	7,16	3,84	3,83
y	**22**	0,30	1,32	-1,38	0,59	0,70
x_3	**32**	0,30	7,32	3,30	1,88	1,75
x_1	**41**	3,40	3,30	2,84	1,70	1,60
x_2	**41**	1,60	5,99	1,19	2,51	2,41
x_2	**46**	0,50	5,99	2,53	2,68	2,65
y	**46**	-0,40	1,32	-1,17	0,30	0,38
y	**49**	0,90	1,32	0,26	0,31	0,42
x_3	**50**	2,90	7,32	5,51	2,73	2,36
x_2	**51**	1,50	5,99	1,84	1,50	1,50
x_2	**52**	0,60	5,99	-0,64	1,78	1,59
x_1	**54**	1,10	3,30	5,85	2,10	2,07
x_3	**54**	0,30	7,32	8,09	3,73	3,66
x_1	**58**	2,40	3,30	1,81	1,64	1,66
x_1	**59**	1,60	3,30	1,45	1,60	1,54
x_1	**64**	2,80	3,30	0,05	1,64	1,56
x_3	**66**	0,80	7,32	0,03	0,86	0,71
x_2	**74**	2,20	5,99	3,03	1,35	1,27

OVYH için elde edilen değerler Çizelge 4.9'da verilmiştir. Sağlam veri yükleme yöntemi ve ER algoritması ile veri yükleme yöntemi en az ortalama hataya sahiptir. Ortalama ile veri yükleme yöntemi ise beklenildiği gibi, kayıp değerler için yapılan yüklemelerde en çok ortalama hataya sahiptir.

Çizelge 4.9. Veri Yükleme Yöntemlerinin Ortalama Veri Yükleme Hatası

Veri Yükleme Yöntemi	OVYH
Ortalama ile Veri Yükleme	40,13
Ardışık Veri Yükleme Yöntemi	7,68
ER Algoritması ile Veri Yükleme	1,74
Sağlam Veri Yükleme	1,72

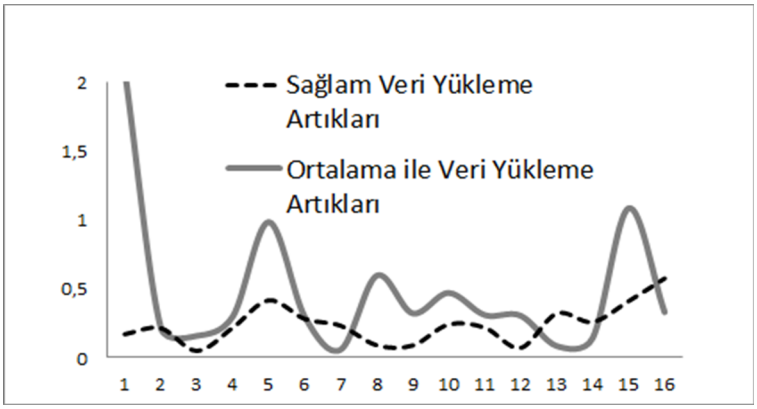
Gerçek değerlerle, yüklenen değerler arasındaki farklar, Şekil 4.1'den görülebilir. ER algoritması ile sağlam veri yükleme yönteminin sonuçları gerçek değere çok yakındır. Ancak ortalama ile veri yükleme yöntemi, özellikle aykırı gözlemin kayıp değerinde başarısız bir şekilde veri yüklemesi yapmıştır.



Şekil 4.1. Gerçek Değerler ile Yüklenen Değerler Arasındaki Farklılıklar

Veri kümesindeki kayıpların yanlış şekilde yüklenmesi oluşturulacak regresyon kestirimini ve uyum kestirimlerini etkileyecektir. Bu durumu incelemek amacıyla tam veri kümesi ve veri kümesinde oluşturulan kayıp değerlerin yüklendiği veri kümesi için M kestiricisi ile sağlam regresyon kestirimi yapılmıştır.

Veri kümesinde herhangi bir RK oluşturulmadan elde edilen regresyon modelinin bağımlı değişken kestirimleri elde edilmiştir. Daha sonra %10 RK şeklinde oluşturulan veri kümesinin ortalama ile veri yükleme ve sağlam veri yükleme yöntemleri ile elde edilen tam veri kümelerinden de regresyon modeli kestirilmiştir. Kayıp değerleri bulunan gözlemlerin uyum kestirimleri elde edilerek farklılıklar incelenmiştir. Yapılan çözümlemeler sonucunda, elde edilen tam veri kümesinin bağımlı değişken değerleri için uyum kestirimi ile ortalama ve sağlam veri yükleme yöntemleriyle kayıpları yüklenen veri kümelerinin uyum kestirimlerinin farkları Şekil 4.2'de gösterilmiştir.



Şekil 4.2. Veri Yükleme Yöntemlerinin Regresyon Artığı Bakımından Farklılıkları

Ortalama ile veri yükleme yöntemi sayesinde elde edilen kayıp veri değerleri regresyon modelinden elde edilen uyum kestirimlerini de etkilemiştir. Sağlam veri yükleme yöntemiyle elde edilen uyum kestirimleri ile gerçek veri kümesinin regresyon modelinden elde edilen uyum kestirimleri arasındaki değerlerin küçük olduğu Şekil 4.2'den görülmektedir.

Uygulama 2'de görüldüğü gibi verinin uygun değerlerle yüklenmesi, yapılacak parametre ve regresyon kestiricilerini de etkilemektedir. Bu nedenle çalışmalarda seçilecek kayıp veri yükleme yöntemi önemlidir. Uygulama 2'deki veri kümesiyle de aykırı değerler veya bozulumun olduğu durum incelenmiş ve sağlam veri yükleme yöntemlerinin kullanılması gerektiği görülmüştür.

BEŞİNCİ BÖLÜM

5. SONUÇ VE TARIŞMA

Bu çalışmada kayıp verinin nasıl ilgi alanı olduğu, geçmişten günümüze hangi aşamalardan geçtiği aktarılmıştır. Kayıp veri için veri silme ve veri yükleme yöntemleri, belirli sınıflamalar altında bir araya getirilmiştir. Kayıp veri yöntemlerinin kullanılabilecekleri kayıp veri mekanizmaları tanıtılmış, yanlış kullanımlarda karşılaşılabilecek sorunlar tartışılmıştır. Veri kümelerinde hem aykırı değer hem de kayıp veri bulunması durumunda ise sağlam kestirimlerin kullanılması gerektiği örneklendirilerek açıklanmış ve bu alanda yapılan öncü çalışmalar verilmiştir. Çalışmanın uygulama kısmında aykırı değere sahip olan veri kümelerinde RK oluşturulduğunda elde edilen ortalama vektörü ve kovaryans matrisi kestirimleri ile kayıp verinin yerine yüklenen değerler sağlam ve klasik veri yükleme yöntemlerinin kullanımına göre incelenmiştir.

Önceki çalışmalardan elde edilen çıkarımlara göre, veri kümelerinde aykırı değer varken kayıp değerlerin sağlam yöntemlerle yüklenmesi çok değişkenli yöntemlerdeki bazı çözümlemeler için büyük fayda sağlamıştır. EM algoritmasının iteratif adımları ile elde edilen sağlam veri yükleme yöntemleri sayesinde kayıp verili kümelerde sağlam TBA, sağlam PLS gibi çok değişkenli yöntemlerin kayıp veri sorununu çözümlediği ve daha etkili çözümler verdiği bazı çalışmalar üzerinden açıklanmıştır.

Uygulamalar sonucunda ise liste bazında veri silme yönteminin, kayıp veriden kurtulmak isterken elde edilmiş birçok değeri veri kümesinden çıkarttığı ve kestirim değerlerinin kayıp oranı arttıkça değişkenlik gösterdiği görülmüştür. Tek bir değer yüklemesi yapan model tabanlı olmayan yöntemlerin (ortalama, ortanca değeri ile veri yükleme) veri kümesinin merkezinde atamalar yaptığı için kovaryans kestirimlerine etki ettikleri uygulamada gösterilmiştir. Model tabanlı yöntemlerin daha iyi sonuçlar verdiği ve aykırı değer bulunduğunda iteratif adımlarının sağlam yaklaşım ile desteklenmesi doğru kestirimler vermiştir. Örneğin Little ve Smith (1987)'in EM algoritmasının adımında yaptığı basit bir sağlam ağırlıklandırma değişikliği (ER algoritması) ile yüklenen veri, aykırı değerlerden etkilenmeden elde edilebilmektedir. Sağlam veri yükleme yöntemi kullanılarak tam veri haline getirilen veri kümelerinde uygulanan sağlam kestiricilerden BS ve MCD'nin daha etkin

sonuçlar verdiği görülmüştür. OGK kestiricisi ikili değişkenler üzerinden hesaplanırken yüksek kestirime sahip gözlemlerin çıkartılmasının düşük kovaryans kestirimlerine neden olduğu görülmüştür. Klasik veri yükleme yöntemleri sonunda elde edilen kestirimlerin bozulma olmadığı durumlarda iyi sonuçlar verdiği görülmüştür. Bozulma durumunda sağlam kestiricilerin klasik yöntemlerden daha üstün oldukları görülmüştür. Yüksek bozulma noktasına sahip kestiricilerin kullanıldığı yöntemlerin ER algoritmasından daha iyi sonuçlar verdiği görülmüştür.

Diğer taraftan Branden ve Verboven (2009) tarafından ardışık veri yükleme yönteminde yapılan sağlam adım değişikliği ile elde edilen sağlam veri yükleme yöntemi gösterilmiştir. Bu yöntemin veri yüklemede ER algoritmasından OVYH küçük olacak şekilde ER algoritmasına benzer yüklemeleri yapabildiği görülmüştür. Ardışık veri yükleme ve ortalama ile veri yükleme yöntemleri aykırı değerlerden dolayı sağlam kayıp veri yükleme yönteminden daha başarısız olduğu görülmüştür.

Hawkins, Bradu ve Kass (1984)'ın veri kümesi üzerinde yapılan çalışmalar sonucunda RK olan veri için yüklenen değerlerin, gerçek değerlere yakınlığı kullanılan yöntemler bakımından incelenmiştir. Sağlam veri yükleme yöntemi ile kullanılan diğer yöntemlerin yüklenen değerleri karşılaştırılmış ve sağlam veri yükleme yönteminin gerçek değere daha yakın değerler yüklediği örnek üzerinden gösterilmiştir. Aynı zamanda kayıp verinin iyi kestirilemediği durumda regresyon modelinden elde edilecek bağımlı değişken uyum değerlerinin de etkilendiği, ortalama ile veri yükleme ve sağlam veri yükleme yöntemleri üzerinden gösterilmiştir.

KAYNAKLAR

Afifi, A. A., Elashoff, R. M., 1966, Missing observations in multivariate statistics I. Review of the literature, Journal of the American Statistical Association, 61, 595-605.

Afifi, A. A., Elashoff, R. M., 1967, Missing Observations in multivariate statistics II. Point estimation in simple linear regression, Journal of the American Statistical Association, 62, 10-29.

Afifi, A. A., Elashoff, R. M., 1969a, Missing Observations in Multivariate Statistics III. Large sample analysis of simple linear regression, Journal of the American Statistical Association, 64, 337-358.

Afifi, A. A., Elashoff, R. M., 1969b, Missing Observations in Multivariate Statistics III. Large sample analysis of simple linear regression, Journal of the American Statistical Association, 64, 259-365.

Allan, F. E., Wishart, J., 1930, A method of estimating the yield of a missing plot in field experimental work, J. Agric. Sci., 20, 399-406.

Allison, P. D., 2000, Multiple imputation for missing data: A cautionary tale, Sociological Methods and Research, 28, 301-309.

Anderson, T. W., 1957, Maximum Likelihood Estimates for a Multivariate Normal Distribution when some Observations are Missing, Journal of the American Statistical Association, 278, 200-203.

Andrews, D. F., Bickel, P.J., Hampel, F.R., Huber, P. J., Rogers, W.H., Tukey, J.W., 1972, Robust Estiamtes of Location, Press, Princeton, 523 p.

Atalay, U., 2003, Yanıtsızlık Durumunda Kayıp Değer Tahmini: İmputasyon, DİE Uzmanlık Tezi, Ankara.

Atkinson, A. C., 1994, Fast very robust methods for the detection of multiple outliers, Journal of the American Statistical Association, 89, 1329-1339.

Atkinson, A. C., Cheng, T. C., 1999, The forward search for the minimum covariance determinant estimator, Technical report, London School of Economics, London WC2A2AE, submitted.

Atkinson, A. C., Cheng, T. C., 2000, On robust linear regression with incomplete data, Computational Statistics and Data Analysis, 33, 361-380.

Baygül, A., 2007, Kayıp Veri Analizinde Sıklıkla Kullanılan Etkin Yöntemlerin Değerlendirilmesi, Yüksek Lisans Tezi, İ. Ü. Sağlık Bilimleri Enstitüsü, İstanbul.

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.

Box, G. E. P., 1953, Non-normality and tests on variances, Biometrika, 40, 318-335.

Branden , K. V., Verboven S., 2009, Robust data imputation, Computational Biology and Chemistry, 33(1), 7-13.

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, 22, 302–306.

Butler, R. W., Davies, P. L., Jhun, M., 1993, Asymptotics for the minimum covariance determinant, The Annals of Statistics, 21(3), 1385-1400.

Candan, M., 1995, Doğrusal Regresyon Çözümlemesinde Sağlam Kestiriciler, Yüksek Lisans Tezi, H.Ü. Fen Bilimleri Enstitüsü, Ankara.

Cao, Y., Poh, K. L., 2006, An accurate and robust missing value estimation for Microarray data: least absolute deviation imputation, Proceedings of the 5th International Conference on Machine Learning and Applications (ICMLA'06).

Cheng, T. S., Victoria-Feser, M. P., 2002, High-breakdown estimation of multivariate mean and covariance with missing observations, British J. Math. Statist. Psych., 55, 317–335.

Cook, R. J., Zeng, L., Yi, G. Y., 2004, Marginal analysis of incomplete longitudinal binary data: A cautionary note on LOCF imputation, Biometrics, 60, 820-828.

Copt, S.,Victoria-Feser, M.P., 2003, Fast algorithms for computing high breakdown covariance matrices with missing data, Cahiers du département d'économétrie août 2003, Faculté des sciences économiques et sociales, Université de Genève, Geneva, Switzerland.

Croux, C., Ruiz-Gazen, A., 1996, A fast algorithm for Robust principal components based on projection pursuit, In: COMPSTAT: Proceedings in Computational Statistics, Physica-Verlag, Heidelberg, 211–217.

David, M. H., Little, R. J. A., Samuhel, M. E., Triest, R.K., 1983, Imputation methods based on the propensity to respond, Proc. Bus. Econ. Sec., Am. Stat. Assoc., 168-173.

Davies, P.L., 1987, Asymptotic behavior of S-estimators of multivariate location parameters and dispersion matrices, The Annals of Statistics, 15, 1269–1292.

Dear, R. E., 1959, A principal-component missing-data method for multiple regression models, SP-86, System Development Corporation, Santa Monica, California.

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, 39, 1–38.

Dempster, A. P., Rubin, D. B., 1983, Introduction of Incomplete Data in Sample Surveys (Volume 2) Theory and Bibliography (W. G. Madow, I. Olkin, D.B. Rubin eds.), p 3-10, New York.

Devlin, S. J., Gnanadesikan, R., Kettenring, J. R., 1981, Robust estimation of dispersion matrices and principal components, J. Amer. Statist. Ass., 76, 354-362.

Dixon, W. J., 1988, BMDP Statistical Software, Los Angeles: University of California Press, 1500 p.

Donders, A.R.T., Heijden, G.J.M.G., Stijnen, T., Moons, K.G.M., 2006, Review: A gentle introduction to imputation of missing values, Journal of Clinical Epidemiology, 59, 1087-1091.

Donoho, D.L., 1982, Breakdown properties of multivariate location estimators, Ph.D, Qualifying Paper, Harvard University.

Edgett, G. L., 1956, Multiple Regression with Missing Observations among the Independent Variables, Journal of the American Statistical Association, 273, 122-131.

Enders, C. K., 2010, Applied Missing Data Analysis, The Guilford Press, New York, 401 p.

Glasser, M., 1964, Linear regression analysis with missing observations among the independent variables, Journal of the American Statistical Association, 59, 834-844.

Gnanadesikan, R., Kettenring, J.R., 1972, Robust estimates, residuals, and outlier detection with multiresponse data, Biometrics, 28, 81–124.

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, Multivar. Behav. Res, 31, 197-218.

Haitovsky, Y., 1968, Missing Data in Regression Analysis, Journal of the Royal Statistical Society. Series B (Methodological), 30(1), 67-82.

Hampel, F.R., 1971, A general definition of qualitative robustness, The Annals of Mathematical Statistics, 42, 1887–1896.

Hampel, F. R., 1973, Robust estimation: A condensed partial survey, Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 27, 87-104.

Hampel, F.R., 1974, The influence curve and its role in robust estimation, The Annals of Statistics, 69, 383–393.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A., 1986, Robust Statistics: The Approach Based on Influence Functions. New York: John Wiley & Sons, Inc., 387 p.

Hartley, H.O., 1956, Programming analysis of variance for general-purpose computers, Biometrics, 12, 110-122.

Hawkins, D. M., Bradu, D., Kass, G., V.,1984, Location of several outliers in multiple regression data using elemental sets, Technometrics, 26, 197-208.

Healy, M.J.R:, Wesmacott, M., 1956, Missing values in experiments analyzed on automatic computers, Appl. Statist., 5, 203-206

Hron, K., Templ, M., Filzmoser, P., 2010, Imputation of missing values for compositional data using classical and robust methods, Computational Statistics Data Analysis, 54, 2095-3107.

Huber, P. J., 1964, Robust estimation of a location parameter, The Annals of Mathematical Statistics, 35, 73–101.

Huber, P. J., 1981, Robust Statistics, New York: John Wiley & Sons, Inc, 301 p.

Huber, P., 1985, Projection pursuit, Ann. Statist., 13, 435–475.

Jarrett, R.G., 1978, The analysis of designed experiments with missing observations, Appl. Statist., 27, 38-46.

Kendall, M. G., Stuart, A., 1969, The Advanced Theory of Statistics, Vol. 1, New York: Hafner Press, 604 p.

Little, R.J.A., 1979, Maximum likelihood inference for multiple regression with missing values: a simulation study, J. Roy. Statist. Soc. Ser. B, 44, 226-233.

Little, R. J. A., 1988a, Robust estimation of the mean and covariance matrix from data with missing values, Applied Statistics, 37, 23-38.

Little, R. J. A., 1988b, A test of missing completely at random for multivariate data with missing values, Journal of the American Statistical Association, 83, 1198–1202.

Little, R. J. A., 1992, Regression with missing X's: A review, Journal of the American Statistical Association, 87,1227–1237.

Little, R. J. A., Rubin, D. B., 1987, Statistical Analysis with Missing Data, Hoboken, N. Jersey, Wiley, 371 p.

Little, R. J. A., Rubin, D. B., 2002, Statistical Analysis with Missing Data (2nd ed.), Hoboken, N. Jersey, Wiley, 381 p.

Little, R. J. A., Smith P. J., 1987, Editing and imputing for quantitative survey data, Journal of the American Statistical Association 82, 58-68.

Liu, G., Gould, A. L., 2002, Comparison of alternative strategies for analysis of longitudinal trials, Journal of Biopharmaceutical Statistics, 12, 207–226.

Longford, N.T., 2006, Missing Data and Small-area Estimation, Springer, 376 p.

Lopuhaa H. P., 1989, On the relation between S-estimators and M-estimators of multivariate location and covariance, The Annals of Statistics, 17, 1662-1683.

Lord, F. M., 1955, Estimation of parameters from incomplete data, Journal of the American Statistical Association, 50, 870-76.

Mallinckrodt, C. H., Clark, W. S., David, S. R., 2001, Accounting for dropout bias using mixed effects models, Journal of Biopharmaceutical Statistics, 11, 9–21.

Marker, D.A., Judkins, D.R., Winglee, M., 2002, Large scale imputation for complex surveys, Survey Nonresponse, Ed. Groves, R.M., Dillman, D.A., Eltinge, J.l., Little, R.J.A., John Wiley: New York, 329-341.

Maronna, R.A., 1976, Robust M-estimators of multivariate location and scatter, The Annals of Statistics, 4, 51–67.

Maronna, R. A., Martin, R. D., Yohai, V. J., 2006, Robust Statistics: Theory and Methods, John Wiley and Sons, England, 426 p.

Maronna, R.A., Yohai, V.J., 1995, The behavior of the Stahel-Donoho robust multivariate estimator, Journal of the American Statistical Association, 90, 330–341.

Maronna, R.A., Zamar, R.H., 2002, Robust estimation of location and dispersion for high-dimensional data sets, Technometrics, 44, 307–317.

McLaachlan, G. J., Krishnan, T., 1997, The EM Algorithm and Extensions, New York: Wiley, 347 p.

Molenberghs, G., Kenward, M. G., 2007, Missing data in clinical studies, West Sussex, UK: Wiley.

Nicholson, G. E. Jr., 1957, Estimation of Parameters From Incomplete Multivariate Samples, Journal of the American Statistical Association, 280, 523-526.

Orchard, T., Woodbury, M.A., 1972, A missing information principle: theory and applications, Proc. Sixth Berkeley Symp. on Math. Statist. and Prob., 1, 697-715.

Pearce, S. C., 1965, Biological Statistics: an Introduction, Chapter 7, New York: McGraw-Hill, 76p.

Pison, G., Aelst, S. V., Willems, G., 2002, Small sample correlations for LTS and MCD, Metrika, 55, 11-123.

Preece, D. A., 1971, Iterative procedures for missing values in experiments, Technometrics,13, 743-754.

Prescher D., 2003, A tutorial on the expectation-maximization algorithm including maximum-likelihood estimation and EM training of probabilistic, Context-Free Grammars, Presented at the 15th European Summer School in Logic, Language and Information (ESSLLI 2003).

Rocke, D. M., 1996, Robustness Properties of S-Estimators of Multivariate Location and Shape in High Dimension, The Annals of Statistics, 24, 1327-1345.

Rousseeuw, P.J., Leroy, A.M., 1987, Robust Regression and Outlier Detection, New York: John Wiley & Sons, Inc., 360 p.

Rubin, D. B., 1976, Inference and missing data, Biometrika, 63, 581–592.

Rubin, D. B., 1983, Iteratively reweighted least squares, Entry in Encyclopedia of the Statistical Sciences, Vol. 4, Kotz, S., Johnson, N. L. and Read, C. B., eds, New York: Wiley.

Rubin, D. B., 1987a, Bayesian inference for causal effects: The role of randomization, Annals of Statistics, 6, 34–58.

Rubin, D. B., 1987b, Multiple imputations in sample surveys—A phenomenological Bayesian approach to Nonresponse, Proceedings of the Survey Research Methods Section of the American Statistical Association, 30–34.

Satıcı, E., 2009, Kayıp Gözlem Olması Durumunda Kitle Ortalaması Tahmini, Doktora Tezi, H.Ü. Fen Bilimleri Enstitüsü, Ankara.

Schafer, J. L., 1997, Analysis of incomplete multivariate data, Boca Raton, FL: Chapman & Hall, 430 p.

Schafer, J. L., Graham, J. W., 2002, Missing data: Our view of the state of the art, Psychological Methods, 7, 147–177.

Serneels, S., Verdonck, T., 2008, Principal component analysis for data containing outliers and missing elements, Comp. Statistics Data Anal., 52, 1712 – 1727.

Serneels, S., Verdonck, T., 2009, Principal component regression for data containing outliers and missing elements, Computational Statistics and Data Analysis, 53, 3855-3863.

Shih, W.J., Weisberg, S., 1986, Assessing inuence in multiple linear regression with incomplete data, Technometrics, 28, 231-239.

Smolínski, A., Walczak, B., Einax, J.W., 2002, Exploratory analysis of data sets with missing elements and outliers, Chemosphere, 49, 233-245.

Stahel, W.A., 1981, Breakdown of covariance estimators, Research Report 31, Fachgruppe Fùr Statistik, ETH, Zurich.

Stanimirova, I., Serneels, S., Espen, P. J. V., Walczak B., 2007, How to construct a multiple regression model for data with missing elements and outlying objects, Analytica Chimica Acta, 581, 324–332.

Tanner, M. A., Wong, W. H., 1987, The calculation of posterior distributions by data augmentation, Journal of the American Statistical Association, 82, 528–540.

Todorov, V., Filzmoser, P., 2009, An Object-Oriented Framework for Robust Multivariate Analysis, Journal of Statistical Software, 32(3), 1-47.

Verboven, S., Branden K.V., Goos, P., 2007, Sequential imputation for missing values, Computational Biology and Chemistry, 31, 320-327.

Walczak, B., 1995a, Outlier detection in multivariate calibration, Chemometrics and Intelligent Laboratory Systems, 28, 259–272.

Walczak, B., 1995b, Outlier detection in bilinear calibration. Chemometrics and Intelligent Laboratory Systems, 29, 63-73.

Walsh, J. E., 1959, Computer-feasible general method for fitting and using regression functions when data incomplete, SP-71, System Development Corporation, Santa Monica, California.

Wilkinson, G. N., 1957, The analysis of covariance with incomplete data, Biometrics, 13, 363-372.

Wilkinson, G. N.,1958, The Analysis of Variance and Derivation of Standard Errors for Incomplete Data, Biometrics, 14(3), 360-384.

Wilkinson, L., 1999, Statistical methods in psychology journals: Guidelines and explanations, American Psychologist, Task Force on Statistical Inference, 54, 594–604.

Wilks, S. S., 1932, Moments and distributions of estimates of population parameters from fragmentary samples, The Annals of Mathematical Statistics, 3, 163–195.

Wynn, H. P., 1970, The sequential generation of D-optimum experimental designs, Ann. Math. Stat., 41, 1655-1664.

Yates, F., 1933, The analysis o replicated experiments when field results are incomplete, Emp. J. Exp. Agric., 1, 129-142.

Yates, F., 1936, Incomplete randomized blocks, Ann. Eugen., Lond., 7, 121-140.

Yazıcı, F., 2005, EM Algoritması ve Uzantıları, Yüksek Lisans Tezi, H.Ü. Fen Bilimleri Enstitüsü, Ankara.

Yoon, D., Lee, E. K., Park, T., 2007, Robust imputation method for missing values in microarray data, BMC Bioinformatics, 8(2), 1-7.

ÖZGEÇMİŞ

Adı Soyadı : ONUR TOKA

Doğum Yeri : Ankara

Doğum Yılı : 1986

Medeni Hali : Evli

Eğitim ve Akademik Durumu:

Lise 2000-2004 Sincan (Yabancı Dil Ağırlıklı) Lisesi

Lisans 2004-2009 Hacettepe Üniversitesi, Fen Fakültesi, İstatistik Bölümü

Lisans 2006-2011 Anadolu Üniversitesi, İşletme Fakültesi, İşletme Bölümü

Yabancı Dil: İngilizce

İş Tecrübesi:

2009 - : Araştırma Görevlisi - Hacettepe Üniversitesi, Fen Fakültesi, İstatistik Bölümü