ANKARA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

DOKTORA TEZİ

HÜCRESEL AYKIRI GÖZLEM OLMASI DURUMUNDA SAĞLAM TAHMİN YÖNTEMLERİ İLE İSTATİSTİKSEL VERİ ANALİZİ

Elif ŞEN

İSTATİSTİK ANABİLİM DALI

ANKARA
2023

ÖZET

Doktora Tezi

HÜCRESEL AYKIRI GÖZLEM OLMASI DURUMUNDA SAĞLAM TAHMİN YÖNTEMLERİ İLE İSTATİSTİKSEL VERİ ANALİZİ

Elif ŞEN

Ankara Üniversitesi
Fen Bilimleri Enstitüsü
İstatistik Anabilim Dalı

Danışman: Prof. Dr. Olçay ARSLAN

Bu çalışmanın genel amacı, çok değişkenli veri setinde hücresel aykırı değer olması durumda bu gözlemlerden daha az etkilenecek sağlam istatistiksel yöntemler kullanarak yerine değer atama (imputasyon) yöntemlerinin incelemesidir. Literatürde karşılaşılan aykırı değer ve kayıp veri kavramlarından bahsedilmiştir. Tek ve çok değişkenli durumunda aykırı değer problemlerinde kullanılan sağlam yöntemler incelenmiştir. Çok değişkenli veri setinde, hücresel ve satırsal aykırı değer problemleri genelde bir arada ya da sadece satırsal olduğu durumlarda yapılan çok sayıda çalışma mevcut iken, sadece hücresel aykırı değer durumunda kullanılan sağlam yöntemler için az sayıda araştırmayla karşılaşılmıştır. Çalışmada çok değişkenli veri setinde sadece hücresel aykırı değer varlığında karşılaşılan problem ele alınmıştır. Kayıp veriyi ve aykırı değeri eş zamanlı aynı veri setinde değerlendirebilmek için, çok değişkenli veri setinde kayıp veri durumunda karşılaşılan problem, hücresel aykırı değer gibi değerlendirilmiştir. Uygulamada IWGPS[1] kuruluşları tarafından kayıp veri durumunda TÜFE[2] hesaplamalarında kullanılan imputasyon yöntemleri ele alınmıştır. Teknolojik cihazların yaygınlaşmasıyla, istatistik ofislerinin veri derleme araçlarına uyarlanabilecek şekilde, anında alandan veri derleme ve istatistik üretme talebine uygun yöntemler önerilmiştir. Önerilen yöntemlerin imputasyon sonuçları, IWGPS kuruluşlarınca kullanılan yöntem sonuçlarıyla ve istatistiksel programlama dilindeki hücresel aykırı değer ve kayıp veri imputasyon sonuçlarıyla karşılaştırılmıştır. Önerilen yöntemler arasında $i\_müd19$ sonuçları, sağlam aykırı değer imputasyon sonuçlarına benzerlik göstermiştir. Tüm dünyada TÜFE hesaplamasında kullanılan ve istatistiklerden üretilen imputasyon araçlarına yardımcı olması için önerilen $i\_müd19$'un, tüm kullanıcılara kolaylık sağlaması amaçlanmıştır ve başka panel veri türü çalışmalarda veri yapısına göre uyarlanabilirdir. Çok değişkenli veri setinde hem hücresel aykırı değer hem de kayıp veri durumu için ortak bir ağırlıklı imputasyon yöntemi olarak da önerilmektedir.

Ağustos 2023, 131 sayfa

Anahtar Kelimeler: Hücresel Aykırı Değer, Değer Atama, Sağlam Tahmin Yöntemleri, TÜFE Değer Atama Yöntemleri.

[1] Fiyat İstatistikleri Kuruluşları Sekreterlikler Arası Çalışma Grubu (IWGPS); Avrupa Birliği İstatistik Ofisi (Eurostat), Uluslararası Çalışma Örgütü (ILO), Uluslararası Para Fonu (IMF), Ekonomik Kalkınma ve İşbirliği Örgütü (OECD), Birleşmiş Milletler Avrupa Ekonomik Komisyonu (UNECE) ve Dünya Bankası (WB)' dan oluşmaktadır.

[2] TÜFE; Tüketici Fiyat Endeksi

ABSTRACT

Ph. D. Thesis

STATISTICAL DATA ANALYSIS WITH ROBUST ESTIMATION METHODS IN CELLWISE OUTLIER OBSERVATION

Elif ŞEN

Ankara University
Graduate School of Natural and Applied Sciences
Department of Statistics

Supervisor: Prof. Dr. Olçay ARSLAN

The general aim of this study is to examine the methods of assigning values using robust statistical methods that will be less affected by these observations in the case of cellwise outliers in the multivariate dataset. The concepts of outlier and missing data encountered in the literature are mentioned. Robust methods used in outlier problems in the case of univariate and multivariate are examined. In the multivariate dataset, there are many studies conducted in cases where cellwise and casewise outlier problems are generally together or only casewise, while there are few studies have been encountered for a robust method used for imputation only in the case of a cellwise outlier. In the study, the problem encountered only in the presence of cellwise outliers in the multivariate dataset is discussed. In order to be able to evaluate missing data and outlier simultaneously in the same data set, the problem encountered in the case of missing data in the multivariate data set is evaluated as a cellwise outlier. In practice, imputation methods used by IWGPS[3] organizations in CPI[4] calculations in case of missing data are discussed. With the widespread use of technological devices, methods suitable for the demand of collecting data and producing statistics from the field immediately, in a way that can be adapted to the data collection tools of statistics offices have been proposed. The imputation results of the proposed methods are compared with the method results used by the IWGPS organizations and imputation results of cellwise outlier and missing data in the statistical programming language. Among the proposed methods, $i\_müd19$ results were similar to the results of robust outlier imputation. The $i\_müd19$ proposed to assist the imputation tools used in CPI calculation all over the world and produced from statistics is intended to provide convenience to all users and can be adapted according to the data structure in other panel data type studies. It is also suggested as a common weighted imputation method for both cellwise outlier and missing data case in multivariate dataset.

August 2023, 131 pages

Key Words: Cellwise Outlier, Imputation, Robust Estimation Methods, CPI Imputation Methods.

[3] Intersecretariat Working Group on Price Statistics (IWGPS) Organizations consist of The Statistical Office of the European Union (Eurostat), International Labour Organization (ILO), International Monetary Fund (IMF), The Organisation for Economic Co-operation and Development (OECD), The United Nations Economic Commission for Europe (UNECE) and The World Bank (WB).

[4] CPI; Consumer Price Index

TEŞEKKÜR

‘Hücresel Aykırı Gözlem Olması Durumunda Sağlam Tahmin Yöntemleri ile İstatistiksel Veri Analizi’ ile ilgili çalışmamda bana çalışma olanağı sağlayan ve her zaman bilgilendiren, hiçbir konuda yardımını esirgemeyen danışman hocam Sayın Prof. Dr. Olçay ARSLAN’ a en içten teşekkürlerimi sunarım. İstatistik bilimini öğreten Ankara Üniversitesi İstatistik Bölümü’ ndeki emektar hocalarıma teşekkür ederim. Eğitimim boyunca hep yanımda olan eşim, annem, babam ve kardeşlerime verdikleri destekten dolayı çok teşekkür ederim.

Elif ŞEN
Ankara, Ağustos 2023

İÇİNDEKİLER

SİMGELER DİZİNİ

$b(i)$	silinen i.satır ile tahmin edilen katsayılar
β	tahmin edilecek parametrelerin vektörü, $\beta \in \mathbb{R}^{p\times 1}$
$\hat{\beta}$	$y = X\beta + e$ şeklinde iken β 'nın en küçük kareler tahmini
$\hat{\beta}_{(-i)}$	i. gözlem veriden çıkartıldıktan sonra hesaplanan parametre vektörü,
χ_p^2	χ_p^2 dağılımı, χ_p^2
C_R	sağlam (robust) konum tahmin edicileri
C_M	sağlam (robust) bir kestirici olan medyan vektörleri
D_i	Cook Uzaklığı (Cook Distance)
e	hata (artık) vektörü, $e \in \mathbb{R}^{n\times 1}$
ϵ	rastgele konumlarda bozulmuş hücrelerin bir oranı
$\bar{\epsilon}$	bozulmuş satırların beklenen oranı, $\bar{\epsilon} = 1 - (1-\epsilon)^p$
F	F dağılımı, $F_{n,n-p}$
H	$n \times n$ boyutlu şapka matrisi, $H = X(X'X)^{-1}X'$
IC	etki eğrisi (influence curve, IC)
IF	etki fonksiyonu (influence function, IF)
L_1	en küçük mutlak, (Least Absolute), $LA = L_1 = \underset{\hat{\beta}}{\text{Minimize}} \sum_{i=1}^{n} \lvert r_i \rvert$
L_2	en küçük kareler, (Least Squares), $LS = L_2 = \underset{\hat{\beta}}{\text{Minimize}} \sum_{i=1}^{n} r_i^2$
λ^*	yerel-kayma-duyarlılığı (local-shift-sensitivity)
μ	aritmetik ortalama
$\tilde{\mu}$	μ tahmini değeri için özel tanımlı tahmin edici
M	kayıp veri gösterge matrisi (missing data indicator matrix), $M = (m_{ij})$
MD_i	Mahalanobis Uzaklığı (Mahalanobis Distance)
n	örneklem hacmi, gözlem sayısı
p	tahmin edilecek parametre sayısı
p^*	red noktası (rejection point)
RD_i	Sağlam Uzaklık (Robust Distances)
S	hesaplanan örneklem varyans-kovaryans matrisi
S^*	şekil tahmini
S_R	sağlam (robust) ölçek tahmin edicileri
S_M	sağlam bir kestirici olan medyan vektörleri C_M kullanılarak hesaplanan kovaryans
SC_n	duyarlılık eğrisi (sensitivity curve)
$s(i)$	tüm regresyon i 'inci gözlem olmadan tekrar çalıştırıldığında σ 'nın tahmini
σ	standart sapma
Σ	Kovaryans Matrisi
Σ^{-1}	Kovaryans Matrisinin Tersi

t_i	standart artıklar
$t(i)$	student artık
T_n	tahmin edici dizisi, $\{T_n; n \geq 1\}$
v_i	0 veya 1 değeri alan ağırlık fonksiyonu
y	bağımlı değişkenlerin vektörü, $y \in \mathbb{R}^{n\times 1}$
$\hat{y}$	y nin tahmin vektörü, $\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y$
Y	n gözlemden alınan p değişkenin değerlerini gösteren veri matrisi
y_{ij}	j. değişkenin i. gözlemde aldığı değer, $i = 1,2,\dots,n$ ve $j = 1,2,\dots,p$
$\underline{Y}$	tam veri vektörü, $\underline{Y} = \left[\underline{Y}_{obs}, \underline{Y}_{miss}\right]^t$
$\underline{Y}_{obs}$	gözlenebilen veri vektörü, tam olmayan veri (incomplete data)
$\underline{Y}_{miss}$	gözlenemeyen veri vektörü
γ^*	büyük hata duyarlılığı (gross-error-sensitivity)
Z	Z skorları, $Z = (x - \mu)/\sigma$
x	gözlenen değer
X	bağımsız değişkenler matrisi, $X \in \mathbb{R}^{n\times p}$
X_c	Tam matris yapısı (completed)
$\bar{X}^*$	konum tahmini
$\bar{x}$	veriden hesaplanan örneklem ortalama vektörü

Kısaltmalar

$CLIME$	Ters Matris Tahmini için Kısıtlı ℓ_1 minimizasyonu (Constrained ℓ_1 minimization for Inverse Matrix Estimation)
$COVRATIO$	Belirlenmiş kovaryans matrislerinin oranı
$DFBETA$	Regresyon katsayılarındaki değişikliğin etki ölçüsü
$DFFIT$	Regresyon tahmin edilen değerlerdeki değişikliğin etki ölçüsü
EM	Beklenti Maksimizasyon algoritması (Expectation Maximization algorithm)
$FVARATIO$	Belirlenmiş varyansdaki değişikliğin etki ölçüsü
GGQ	(Glasso Gauss Q_n, Gaussian rank correlation, Q_n-estimator as a scale)
$GLASSO$	Grafiksel Lasso (Graphical Lasso)
GSE	Genelleştirilmiş S-tahmincisi (Generalized S-Estimator)
I	İhmal Edilebilir (Ignorable)
ICM	Bağımsız Bozulma Modeli (Independent Contamination Model)
IQR	Çeyreklikler Arası Dağılım Aralığı (Inter Quantile Range)
$kNNI$	k En Yakın Komşuluk İmputasyonu (k Nearest Neighbor Imputation)
$Lasso$	En Küçük Mutlak Çekme ve Seçim Operatörü (Least Absolute Shrinkage and Selection Operatör)
LMS	En Küçük Medyan Kareler Yöntemi (Least Median of Squares)
$LOCF$	Son Gözlemi İleri Taşıma (Last Observation Carried Forward)

LS	En Küçük Kareler Yöntemi (Least Squares)
LTS	En Küçük Budanmış Kareler (Least Trimmed Squares)
MAR	Rasgele Kayıp (Missing at Random)
$MCAR$	Tamamıyla Rasgele Kayıp (Missing Completely at Random)
MCD	Minimum Kovaryans Determinant (Minimum Covariance Determinant)
$MCMC$	Markov Zincirli Monte Carlo yöntem (Markov Chain Monte Carlo)
MI	Çoklu İmputasyon (Multiple Imputation)
ML	En Çok Olabilirlik tahmin edicisi (Maximum Likelihood estimation)
$MMRWAL$	MM-Sağlam Ağırlıklı Adaptif Lasso (MM-Robust Weighted Adaptive Lasso)
MSD	Mahalanobis Kare Uzaklık (Mahalanobis Squared Distance)
MVE	Minimum Hacim Elipsoidi (Minimum Volume Ellipsoid)
NI	İhmal Edilemez (Nonignorable)
NNI	En Yakın Komşuluk İmputasyonu (Nearest Neighbor Imputation)
NPD	Belirli bir simetrik matrise en yakın pozitif tanımlanmış matris (Nearest Positive Definite)
$OCD(A)L$	Aykırı Düzeltilmiş Veri (Ayarlanabilir) Lasso (Outlier Corrected Data (Adaptive) Lasso , $(OCD-(A)\ Lasso)$)
p-value	p değeri, bir istatistiksel modele bağlı olarak gözlemlenen örneklem sonuçlarının (bir test istatistiği) ne kadar aşırı olduğunu ölçmek için kullanılan bir fonksiyondur.
Q_1	1.Çeyreklik (1.Quantile)
Q_3	3.Çeyreklik (3.Quantile)
$QUIC$	Kuadratik Ters Covaryans (Quadratic Inverse Covariance)
PO	Tahmin Edici Aykırılığı (Predictor Outlyingness)
$ROBI$	$ROBi$mpute sağlam atama, konum ve dağılım (scatter) için sağlam bir tahmin edici ekleyerek daha sağlam bir atama yöntemi
SDO	Stahel Donoho Aykırılığı (Stahel Donoho Outlyingness)
$SEQI$	$SEQi$mpute kovaryansının determinantını en aza indirerek bir veri setindeki kayıp değerleri sırayla tamamlayan ardışık imputasyon (sequential imputation)
$SKNN$	Ardışık K En Yakın Komşuluk İmputasyonu (Sequential K Nearest Neighbour)
$THCM$	Tukey–Huber Bozulma Modeli (Tukey–Huber Contamination Model)
$TSGS$	İki aşamalı genelleştirilmiş S-tahmin edicisi (two-step generalized S-estimator)

ŞEKİLLER DİZİNİ

ÇİZELGELER DİZİNİ

1. GİRİŞ

İstatistik ve istatistik uygulamaları için tam veriye ulaşmak büyük önem taşımaktadır. Verilerin toplanması, özetlenmesi, çözümlenmesi, analizi ve sonuç çıkarımı istatistik biliminin konusu; doğru, yansız ve güvenilir parametreler elde etmek amacıdır. Tam veri, tam olmayan veri ve kayıp veri türleri ele alındığında; araştırmacıların uygulamada karşılaştığı kayıp veri sorunu ile ilgili çeşitli yöntem ve öneriler mevcuttur. Benzer şekilde araştırmacıların uygulama sürecinde elde ettiği veri setinde aykırı değer sorunuyla da karşılaşılmaktadır. Veri setindeki kayıp veri ve aykırı değer yapıları birbirinden tamamen farklı durumlardır. Ayrıca, hem aykırı değeri belirleme problemi hem de bu problemle başa çıkma yöntemlerini belirleme ayrı birer çalışma konusu olduğu açıktır. Veri setine göre tek ve çok değişkenli yapılar göz önünde bulundurularak, son zamanlarda çok değişkenli bir veri setinde aykırı değerleri belirlemek için hücresel ve satırsal aykırı değer bir arada değerlendirilen çalışmalar karşımıza çıkmaktadır.

Pek çok istatistiksel araştırmada, derlenen veri setlerinde eksiklikler, kayıp veri olarak adlandırılır ve bu problem araştırmacıların karşılaştığı bilinen bir sorundur. Başka bir değişle, araştırmadaki anlamlı bir değeri gizleyen, analiz aşamasında anketi anlamlı kılacak gözlemlenmemiş değerlere kayıp veri denir (Little ve Rubin 1987). Bu problem, yapılacak istatistiki araştırmalarda kitle hakkında anlamlı kestirimler yapmamızı etkileyecektir. Anket ve veri toplama maliyetinin yüksek olması veri setinin kayıp değerler içermesinde etkili rol oynayabilir. Anket sürecindeki cevaplayıcıdan kaynaklı veri kayıpları da göz ardı edilmemelidir. Örneğin hanehalkı ile ilgili bir ankette hanehalkı üyesi, gelir bilgisi ile ilgili bilgi vermeyi ret edebilir, anketteki diğer tüm soruları cevaplayıp gelir bilgisinin atlanmasını talep edebilir. Bu kayıp, istatistiksel analiz ve sonuç çıkarım süreçlerini olumsuz etkileyecektir. Gerçek hayattaki bu olumsuz durum, teorik olarak istatistik biliminde örneklem çapının azalması sorunu olarak karşılaştığımız durumlara bir örnektir. Kayıp verinin, n örneklem hacmindeki azalmaya sebebiyet verdiği açıktır. Kitlenin homojenliği fazla ise alınması gereken örneklem büyüklüğünün azaldığı bilinmektedir. Ancak kitledeki birimler arasında farklılaşma büyüdükçe doğru istatistiksel analizler için büyük örneklem hacmi almak

gerekmektedir. Büyük örneklem boyutu, standart hatayı azaltır. Standart hata büyürse, onun kitlenin parametresini tahmin etme ya da kitleyi temsil etme becerisi azalmaktadır. Yani daha fazla kayıp hücre olması durumunda n örneklem hacmi azalması; yanlılığa sebep olabilir, standart hatayı artırabilir bu da kitlenin parametresini tahmin etme ya da kitleyi temsil etme becerisini azaltmaktadır. Araştırmacıların kayıp veri problemini, elde ettikleri veriler üzerinde yapacakları analizlerden doğru sonuçlar alabilmek için çözmeleri gerekmektedir. Bunun için araştırmacılar, veriye yeni gözlemlerin eklenmesi, kayıp gözlemlerin veri setinden çıkartılması, kayıp veriyi tahmin edici ile hesaplanması ve elde edilen yaklaşık değerlerin kayıp veriler yerine kullanılması yöntemlerinden birini uygulayarak kayıp veri problemini çözmeye çalışabilirler.

Kayıp veri için araştırmacıların karşılaştığı bu veri derleme sürecindeki sorunlara benzer olarak, veri setinde aykırı değer sorunları ile de karşılaşılmaktadır. Aykırı değer; bir veri kümesinde, bu veri setinin geri kalanıyla tutarsız görünen bir gözlem olarak tanımlanmıştır (Barnett ve Lewis 1978). Bilindiği gibi araştırmalarda veri setlerindeki gözlemlerin aynı yapıya sahip olduğunu bir başka deyişle aynı dağılımdan geldiğini varsayılır. Veri setinde bulunan ve "diğerlerinden farklı" olarak adlandırılan aykırı değerleri belirlemek ve bu gözlemlere ilişkin gerekli çalışmayı yapmak ve verinin yapısını belirlemek araştırmacıların daha doğru sonuçlar elde etmesi için gereklidir. Çünkü verinin yapısı, değişken sayısı ve verideki şüpheli gözlemlerin sayısı hangi yöntemin kullanılacağını belirlemektedir. Veri setindeki aykırı değer tek değişkenli durumunda değişkenler tek tek ele alınarak kolaylıkla belirlenebilir. Grafiksel yöntemler ve istatistiksel testler yardımıyla aykırı değerler hakkında bilgi sahibi olunabilir. Ancak birden fazla değişken ele alındığında çok değişkenli bir yapıda verilerin genel eğilimine uymayan gözlemlerin belirlenmesi oldukça zordur ve aykırı değerleri belirlemek için farklı yöntemler kullanmak gerekmektedir. Veri setindeki aykırı değer belirlendikten sonra; aykırı değeri veri setinden çıkartma, aykırı değerin etkilerini azaltacak istatistiksel yöntemler kullanma veya aykırı değer yerine yeni değer atama olarak uygun yöntem seçimi yapılır. Aykırı değer tespit edilip belirlendikten sonra çözümleme aşamasında, son zamanlarda yapılan incelemeler sonucunda araştırmacılar, aykırı değerlerin veri setinden atılması yerine; imputasyon veya sağlam tahmin yöntemleri ile aykırı değerlerin parametre tahminine olan olumsuz etkilerinin en aza indirilmeye

çalışmışlardır. Aykırı değeri belirleme ve yerine değer atama problemlerinin yanı sıra, bu aykırı değer varlığında yapılacak kestirimler de başka bir çalışma konusu olduğu unutulmamalıdır. Bu problemle başa çıkmak için, yapılacak istatistiki araştırmalarda kitle hakkında anlamlı kestirimler yapmamıza yardımcı olarak sağlam kestirici yöntemler üzerine çalışmalar mevcuttur.

Son zamanlarda çok değişkenli veri setinde satırsal ve hücresel aykırı değer kavramı içeren iki çeşit yapı ile karşılaşmaktayız. Yani, çok değişkenli yapıda veride sadece aykırı değer varlığından söz edemeyiz. Bu aykırı değerin veri setindeki gözlemlenme durumuna göre, satırsal ve hücresel olmak üzere iki farklı aykırı değer tanımı karşımıza çıkmaktadır. Literatürde bu iki tür satırsal ve hücresel aykırı değerler genelde bir arada ya da sadece satırsal olduğu durumlarda yapılan çalışmalar dikkat çekmektedir. Yani çok değişkenli veri setinde sadece hücresel aykırı değer belirleme ve çözümleme ile ilgili çalışmalar nadirdir. Bu çalışmada, çok değişkenli veri setinde sadece hücresel aykırı değer varlığında karşılaşılan problemler ele alınacaktır.

Amaç : Çok değişkenli veri setinde hücresel aykırı değer olduğu durumda yerine değer atama (imputasyon) yöntemleri incelenecek, imputasyonda istatistiksel olarak veri setini en az etkileyecek şekilde kullanılan sağlam (robust) yöntemler üzerinde çalışılacaktır. Kayıp veriyi ve aykırı değeri eş zamanlı, bir veri setinde değerlendirebilmek için; çok değişkenli veri setinde kayıp veri durumunda karşılaşılan problem, hücresel aykırı değer gibi değerlendirilecektir. Örnek bir veri setinde, sağlam aykırı değer ve kayıp veri imputasyon yöntem sonuçları, önerilen yöntem imputasyon sonuçları ve mevcut uygulamada kullanılan imputasyon yöntem sonuçları karşılaştırılacaktır.

Literatür Taraması : Kayıp veri ve aykırı değer kavramları istatistik biliminin başa çıkmak için uğraştıkları birbirinden tamamen farklı iki problemdir. Bu iki konu için yapılan literatür çalışmalarının detayları ilgili bölüm tarihçelerinde tekrar ele alınacaktır. Bu çalışmanın genel amacını tetikleyen başlıca unsur, çok değişkenli veri setinde son zamanlarda yapılan çalışmalarda karşılaştığımız hücresel ve satırsal aykırı değer kavramlarıdır. 1960’ larda çok değişkenli veri setinde aykırı değer kavramı olarak

verideki aykırı sıraları (durum, olay veya satırları) tespit edebilen ve bu yöntemlere alternatif araştıran çalışmalar yapılmıştır. Çok değişkenli veri setinde hücresel aykırı değer, 2002 yılında ilk hücresel bozulma modeli olarak araştırılmış ve daha sonra 2009 yılında Alqallaf vd. (2002, 2009) tarafından detaylı olarak tanımlanmıştır. Bu çalışmalar çok değişkenli veri setinde aykırı değer probleminin, hücresel ve satırsal aykırı değer tanım başlıkları altında incelenmesi gerekliliğini doğurmuştur. Aşağıda inceleyeceğimiz makalelerde; önce çok değişkenli yapıda sadece aykırı değer kavramı kullanılırken, sonra bu ifadenin hücresel ve satırsal aykırı değer olmak üzere iki tanım altında nasıl zamanla geliştiğine güzel bir örnek olacaktır. Son zamanlarda yapılan araştırmalarda bu ifadelerin sağlam teknikler ile buluşması ile birlikte, hücresel ve satırsal bozulma modeli olarak da geliştirildiğini görebiliriz.

Rocke (1996) çalışmasında; çok değişkenli konum ve ölçeğin sağlam tahmini problemi için S-tahmincilerini, genellikle yapıldığı gibi boyutuna bakılmaksızın, sabit bir ρ fonksiyonunun ölçek dönüşümlerini kullanarak tanımlamanın, bozuk bir sonuca yol açabileceğini söylemiştir. Yani, yüksek boyuttaki tahmin edicilerin % 50' ye yaklaşan bir bozulma noktası olabileceğini, ancak yine de ana nokta kütlesinden büyük mesafeler olan aykırı noktalar olarak reddedilemeyeceğinden bahsetmiştir. Bunun önemli pratik sonuçları olan bir tür yanıltıcılığa (nonrobustness) yol açacağı düşüncesiyle S-tahmin edicilerinde iyileşme sağlayan tahmin ediciler tanımlamıştır.

Danilov vd. (2012) tarafından yapılan araştırmada, veri kalitesini ve bir tahmin edicinin performansını etkileyebilecek sorunlardan bahsetmişlerdir. Veri kalitesi ile ilgili iki ana konuyu veri yapısının bozulması (aykırı değerler) ve veri tamamlama (kayıp veriler) olarak ele almışlardır. Temel bileşenler, kanonik korelasyon, dikriminant analiz vb. gibi birçok sağlam çok değişkenli analiz tekniği için önem taşıyan; çok değişkenli konum ve dağılım değerlerini tahmin etmek amacıyla, aykırı değer ve kayıp veri sorunu eş zamanlı ele almışlardır. Kayıp veri ve aykırı değerlerle eş zamanlı olarak başa çıkmak için çok değişkenli konum ve dağılımın S-tahmin edicilerinin tanımını genelleştirmişlerdir.

Agostinelli vd. (2015) tarafından, çok değişkenli konum ve dağılım matrisinin sağlam tahmini üzerinde durulmuş ve verilerin, bağımsız hücresel ve satırsal aykırı değerleri içermesi durumda çalışılmıştır. Çok sayıda değişken ve az sayıda durum içeren düz veri setleri durumunda, geleneksel sağlam prosedürler tarafından gerçekleştirildiği gibi, bir durumun yani satırın tümünün genel olarak azaltılması, zayıf sonuçlara yol açabildiği düşünülmüştür. Hücresel aykırı değerlerle etkin bir şekilde başa çıkabilen ve aynı zamanda satırsal aykırı değerler altında iyi performans gösterebilecek yeni nesil sağlam tahmin ediciler ele alınmıştır.

Rousseeuw ve Bossche (2017) tarafından yapılan araştırmada, çok değişkenli bir veri seti üzerinde çalışıp, p boyutlarındaki n durumdan oluşan ve genellikle $n \times p$ boyutlu veri matrisinde depolanan gerçek verilerin aykırı değerler içerebileceğini vurgulamışlardır. Duruma göre aykırı değerlerin, veri analizini olumsuz etkileyebilecek veya beklenmedik bilgilerin değerli olabileceği istenmeyen hatalar olabileceğine dikkat çekmişlerdir. İstatistik ve veri analizinde, aykırı değer genellikle veri matrisinin bir sırasına karşılık geldiğini ve bu aykırı değerlerin tespit edilmesi için yöntemlerin yalnızca satırların en az yarısı temiz olduğunda çalıştığını söylemişlerdir. Ancak genellikle birçok sıra, her bir değişkene (sütun) ayrı ayrı bakılarak görülemeyen birkaç bozulmuş hücre değerine sahip olduğu vurgulanmıştır. Değişkenler arasındaki korelasyonları dikkate alan çok değişkenli bir örnekte, aykırı veri hücrelerini saptamak için yöntem önermişlerdir.

Croux ve Öllerer (2015) tarafından yapılan yorumlarında, hem hücresel hem de satırsal bozulmaya dayanıklı, çok değişkenli konum ve dağılım için bir tahmin edici önermelerinden bahsedilmiştir. Literatürde belirtildiği gibi, dağılım matrisinin tahmini "köşe taşı" dır ve uygulamalar (ana bileşen analizi, faktör analizi ve çoklu doğrusal regresyon), kovaryans matrisi Σ yerine kovaryans matrisinin tersi $\Theta = \Sigma^{-1}$ matrisini gerektirir. Önerilen iki aşamalı genelleştirilmiş S-tahmin edicisinin ($TSGS$, two-step generalized S-estimator) tersi, kovaryans matrisinin tersinin bir tahminini vereceğini söylemişlerdir. Önerilen iki aşamalı genelleştirilmiş S-tahmin edicisi, hem hücresel hem de satırsal bozulma altında var olan tam ve sağlam bir tahmin edici olduğu tespit

edilmiştir. Yaptıkları simülasyon deneylerinde çeşitli bozulma şemaları denenmiş ve önerilen $TSGS$' nin bunların hepsinde çok iyi performans gösterdiğini belirtmişlerdir.

Tarr vd. (2016) tarafından yapılan araştırmada, temel bileşen analizi ve doğrusal diskriminant analizi gibi birçok standart tekniğin doğal olarak aykırı değerlere duyarlı olduğu sağlam tekniklere büyük ihtiyaç olduğu vurgulanmıştır. Standart sağlam prosedürler, bir veri matrisinin gözlem sıralarının (satırlarının) yarısından daha azında bozulma olduğunu varsayar; ancak bu varsayımın, gözlenen özelliklerin sayısı fazla olduğunda gerçekçi olmayacağı düşünülebilir. Hücresel bozulma altında kovaryans ve ters kovaryans matrisleri tahmin etme problemi üzerinde çalışmışlardır.

Leung (2016) tarafından yapılan çalışmasında; hücresel aykırı değerlerin, büyük boyutlu veri kümelerinde satırsal aykırı değerle birlikte ortaya çıkması vurgulanmıştır. Son çalışmalar, bu tür veri kümelerine uygulandığında geleneksel yüksek kırılma noktası prosedürlerinin başarısız olabileceğini göstermiştir. Amaçlarının çok değişkenli konum ve dağılım matrisini tahmin etmek ve her ikisi de çok değişkenli veri analizinde köşe taşları olan çıkarım için katsayılar ve güven aralıkları regresyonunu tahmin etmek olduğunda bu sorunu göz önünde bulundurduklarını söylemişlerdir. İlk sorunu çözmek için, hücresel ve satırsal aykırı değerleriyle başa çıkmak için iki adımlı bir prosedür önerilmiştir. Daha sonra konum ve dağılım matrisini tahmin etmek için önerilen iki aşamalı prosedür uygulamışlardır. Simülasyon sonuçları ve gerçek veri örnekleri, önerilen yöntemlerin hem hücresel hem de satırsal aykırı değerlerle benzer şekilde baş edebildiği gösterilmiştir.

Leung vd. (2016) tarafından yapılan araştırmada hücresel aykırı değerler, nispeten büyük boyutlu modern veri kümelerinde satırsal aykırı değerler ile birlikte ortaya çıktığını belirtmişlerdir. Son çalışmalar, bu tür veri kümelerine uygulandığında geleneksel sağlam regresyon yöntemlerin başarısız olabileceğini göstermiştir. Satırsal aykırı değerler için sağlam regresyon hakkında geniş bir literatür varken, sadece hücresel aykırı değerler için dar bir literatür olduğunu, ek olarak regresyon bağlamında hücresel ve satırsal aykırı değer tipi için literatürün yetersiz olduğunu vurgulamışlardır.

Hücresel ve satırsal aykırı değerleriyle başa çıkabilen üç adımlı bir regresyon tahmin edicisi önermişlerdir. Ayrıca, tahmin edicilerinin merkezi regresyon modeli dağılımında tutarlı ve asimptotik olarak normal olduğunu kanıtlamışlardır.

Benzer şekilde Leung vd. (2017) araştırmalarında, satırsal aykırı değerleri için sağlam tahmin üzerine geniş bir literatür olduğunu, ancak sadece hücresel aykırı değerler için yetersiz bir literatür varlığından bahsedilmiştir. Çok değişkenli konum ve dağılım matrisi tahmininin, çok değişkenli veri analizinde bir köşe taşı olduğu vurgulanmıştır. Çok değişkenli konum ve dağılım matrisinin hücresel ve satırsal aykırı değerlerin varlığında sağlam bir tahmin gerçekleştirmesi için 2016 da iki aşamalı önerilen yaklaşımlarını tekrar ele almışlardır. Bir simülasyon çalışması ve gerçek bir veri örneğiyle, orijinal iki aşamalı prosedürün aksine, değiştirilmiş iki aşamalı yaklaşımın yüksek boyutta iyi performans gösterdiğini belirtmişlerdir. Ayrıca, değiştirilmiş prosedürün hücresel ve satırsal veri bozulmasında orijinal prosedürden ve diğer son teknoloji ürünü sağlam prosedürlerden daha iyi performans gösterdiğini söylemişlerdir.

Maronnaa ve Yohai' de (2017) yazarların amacı, kontrol edilebilir bir sonlu-örnek etkinliğine (efficiency) ve büyük p için uygulanabilirliğe sahip, oldukça sağlam olmaları anlamında sadece yüksek bir bozulma noktasına değil aynı zamanda nispeten düşük bir kontaminasyon yanlılığına (contamination bias) sahip, çok değişkenli konum ve dağılımın en iyi tahmin edicilerini seçmektir. En sık kullanılan tahmin edicilerin bu bakımdan tatmin edici olmadığını düşünmüşlerdir.

Rousseeuw ve Bossche (2015) tarafından yapılan araştırmada yazarlar, hücresel ve satırsal aykırı değerlerin kombinasyonunu ele almak için özel olarak tasarlanmış bir tahmin edici önermişlerdir. Hücresel bozulma kendi başına yeterince zor olsa da, verilerde hem hücresel hem de satırsal aykırı değerlerin meydana gelebileceği gerçekçi durumun daha da zor olduğunu dile getirmişlerdir. Geliştirilen yöntemlerle, aykırı değerlerin bir türüyle ya da diğeriyle başedebileceğini, ancak her ikisiyle birden başedilemeyeceğini vurgulamışlardır. Önerdikleri yöntemin, şu anda hücresel ve satırsal

aykırı değerlerin kombinasyonu ile başa çıkmak için mevcut en iyi yöntem olduğunu düşünmüşlerdir.

Machkour vd. (2017a) tarafından yapılan araştırmada, sağlam lineer regresyon tahmin edicilerinin büyük çoğunluğunda, satırsal aykırı değere karşı sağlamlığa odaklanıldığından bahsetmektedir. "Yüksek bozulmalı" olarak adlandırılan tahmin ediciler bile, regresyon matrisi satırlarının çoğunun aykırı değerlerden etkilenmediği varsayımına dayanmaktadır. Yakın zamanda, ilk kez hücresel sağlam regresyon tahmin yöntemleri geliştirilmiştir. Çalışmalarında, hücresel ve satırsal aykırı değerlerin bozulmuş olduğu koşulsuz veya az belirlenmiş doğrusal regresyon problemlerine çözüm bulma sorunu araştırılmıştır, sağlam değişken seçimi yapmak için gereken 'sağlam özellikleri' tanımlanmıştır. Model seçimi için, hücresel aykırı değerleri dikkate alan sağlam ağırlıklı ve uyarlanabilir $Lasso$ tipi düzenleyici bir terim önerilmiş ve analiz edilmiştir.

Machkour vd. (2017b) tarafından yapılan araştırmada, açıklayıcı değişken sayısının örnek boyutunu aştığı aralıklı (sparse) modeller irdelenmiştir. Yüksek boyutlu ortamlarda, birçok aykırı değerden sorumlu tek bir bozulmuş tahmin ediciye sahip olabileceği düşünülmüştür. Bu tahmin edici ile ortaya konan aykırı değerler, mevcut herhangi bir sağlam tahmin edicinin kırılma noktasını kolayca aşacağı belirtilmiştir. Ancak, birçok veri seti, yüksek oranda ve bağımsız olarak bozulmuş tahmin ediciler içerir. Bu nedenle, yeni bir sağlam ve seyrek tahmin edici önerilmiştir. Önerilen yönteminin bir avantajı, hesaplama karmaşıklığının çok daha az olmasıdır. Veri kümesinde yüksek derecede bozulmuş tahmin edicinin üstesinden geldiği ve klasik bozulma modeli altında iyi performans gösterdiği belirtilmiştir.

Toka vd. (2021) tarafından yapılan araştırmada, regresyon analizi ile ilgili iki ana husus, aykırı değerlerin varlığında tahmin ve değişken seçimi ele alınmıştır. Popüler sağlam regresyon tahmin yöntemleri aynı anda sağlam tahmin ve değişken seçimine ulaşmak için değişken seçim yöntemleriyle birleştirilmiştir. Yakın tarihli çalışmaların, tahminde kullanılan sağlam tahmin yöntemlerinin ve değişken seçim prosedürlerinin yalnızca veri

setindeki satırsal aykırı değerlere dirençli olduğu vurgulanmıştır. Sağlam tahminin satırsal aykırı değerlerle başa çıkabildiğini ancak, hücresel aykırı değer olması durumunda çok değişkenli istatistiklerde hücresel aykırı değerlerin yapılan sağlam tahminlerle kolayca çözülemediğini gözlemlemişlerdir. Çalışmalarında, verilerde hem hücresel hem de satırsal aykırı değerlerle başa çıkabilmek için sağlam bir tahmin ve değişken seçim yöntemi önermişlerdir. Simülasyon sonuçlarıyla, önerilen tahminin ve değişken seçim prosedürünün, hücresel ve satırsal aykırı değerlerin varlığında iyi performans gösterdiğini ortaya koymuşlardır.

Literatür taramasında görüldüğü gibi, çok değişkenli yapıda hücresel ve satırsal aykırı değer problemi ile başa çıkabilmek için; korelasyon yapılarını dikkate alan yöntemler, kovaryans ve kovaryans matrisinin tersinin tahminine ilişkin yöntemler, konum ve dağılım matrisinin tahminleri ve bunların sağlam tahminleri için genelleştirilmiş S-tahmin ediciler, M-tahmin ediciler, $Lasso$ tipi değişken seçim operatörleri vb. yöntemler önerilmiştir. Ancak burada dikkat çeken ortak düşünce; çok değişkenli veri setinde hücresel ve satırsal aykırı değer genelde bir arada ele alınmış ve satırsal aykırı değer için geniş literatür çalışmalarıyla karşılaşılmasıdır. Ayrıca sadece hücresel aykırı değer için yetersiz literatür çalışması olduğu ve sağlam yöntemlerin satırsal aykırı değer ile başa çıkabilir iken hücresel aykırı değerlerin kolayca çözülemediği gözlemlenmiştir. Bu doğrultuda, çok değişkenli veri setinde sadece hücresel aykırı değer olduğunda, bu problemle başa çıkabilmek için sağlam yöntemler üzerinde çalışılması planlanmaktadır.

2. KAYIP VERİ ve KAYIP VERİ YÖNTEMLERİ İÇİN GENEL BİLGİLER

Pek çok istatistiksel araştırmada, derlenen veri setlerinde eksiklikler bulunur. Veri setlerindeki bu gözlemlenemeyen değerler, kayıp veri olarak adlandırılır ve bu problem araştırmacıların karşılaştığı yaygın bir sorundur. Bu problem, yapılacak istatistiki araştırmalarda kitle hakkında anlamlı kestirimler yapmamızı etkileyecektir. Başka bir değişle, araştırmadaki anlamlı bir değeri gizleyen, analiz aşamasında anketi anlamlı kılacak gözlemlenmemiş değerlere kayıp veri denir (Little ve Rubin 1987). Kayıp veriler, deneysel (empirical) araştırmalarda yaygın bir sorundur ve esasen her çalışmada ortaya çıkar. Özellikle büyük gruplar üzerinde yürütülen çalışmalarda bilgilerin tüm durumlar için tamamlanması neredeyse olanaksızdır (Cool 2000). Anket ve veri toplama maliyetinin yüksek olması veri setinin kayıp değerler içermesinde etkili rol oynayabilir. Anket sürecindeki cevaplayıcıdan kaynaklı veri kayıpları da göz ardı edilmemelidir. Örneğin, sağlık sektöründeki karşılaşılan zamana ve hastalık çeşitlerine bağlı, az sayıda gözlem imkanı olan hastalık ve tedavi için araştırılan verinin ölçüm aşamasındaki zorluklar da unutulmamalıdır. Ya da hanehalkı ile ilgili bir ankette hanehalkı üyesi, gelir bilgisi ile ilgili bilgi vermeyi ret edebilir, anketteki diğer tüm soruları cevaplayıp gelir bilgisinin atlanmasını talep edebilir. Araştırmacıların kayıp veri problemini, elde ettikleri veriler üzerinde yapacakları analizlerden doğru sonuçlar alabilmek için çözmeleri gerekmektedir. Bunun için araştırmacılar, kayıp veri problemi çözümlemeleri için birçok yöntem üzerinde çalışmışlardır.

2.1 Kayıp Verinin Tarihi Kronolojisi

Veri analizi yaparken veri setinde çeşitli nedenlerle ortaya çıkan kayıp verilerin olması problem yaratmaktadır. Sebeplere dayalı ortaya çıkan bilgi kaybı uygun teknikler ile tamamlanmalıdır. Zamanla bu konuya olan ilgi arttıkça yapılan çalışmaların sayısı da artmıştır. Kayıp veri çözümleme çalışmaları 1930 larda başlamış olsa bile büyük atılımlar bilgisayar teknolojisindeki ilerlemelerle birlikte 1970 lerde maksimum olabilirlik tahmini teknikleri ve çoklu imputasyonun ortaya çıkmasıyla başlamıştır.

Rubin' e (1976) göre, çok değişkenli normallik ile ilgili bazı makalelerde (Wilks 1932, Anderson 1957, Afifi ve Elashoff 1966, Hocking ve Smith 1968, Hartley ve Hocking 1971), kayıp verilere neden olan süreçle ilgili varsayım, veri setindeki her bir değerin eşit derecede kayıp olması gibi görünmektedir. Ayrıca varyans analizi ile ilgili diğer makalelerde (Hartley 1956, Healy ve Westmacott 1956, Wilkinson 1958, Rubin 1972, 1976), bağımlı değişkenlerin değerlerinin, gözlemlenecek değerlere bakılmaksızın kayıp veri varsayıldığını söylemiştir.

Rubin' in (1976) kayıp veri düzeneklerini sınıflandırması, kayıp veri araştırmalarına ilgiyi arttırmıştır. Literatüre baktığımızda yerine değer atama ile ilgili çalışmalar Rubin' in 1976 yılında yaptığı çalışma ile başlamıştır. Günümüzde kullanılan kayıp veri ve yerine değer atama terminolojisi ilk kez Little ve Rubin tarafından kullanılmıştır.

2.2 Kayıp Veri Tanımı

Veri Matrisi: Geleneksel olarak, veri matrisinin satırları, durumlar, gözlemler veya olaylar olarak da adlandırılan birimleri temsil eder ve sütunlar, her birim için ölçülen değişkenleri temsil eder. Little ve Rubin (1987), veri setine ait bilginin cebirsel olarak dikdörtgensel yapıda olduğu ve bu varsayım altında istatistiki araştırmaların yapıldığını açıklamıştır. Matristeki bazı girdiler gözlenmediğinde bu dikdörtgensel yapıyı bozacak kayıp hücre, araştırmadaki istatistikleri etkileyecektir. Veri matrisinde hangi değerlerin gözlemlendiğini ve hangi değerlerin kayıp olduğunu gösteren veri matrisi yapısı aşağıdaki gibi tanımlanmıştır (Little ve Rubin 2002).

2.2.1 Tam veri (Complete-Data)

Veri matrisinde y_{ij}, j. değişkenin i. gözlemde aldığı değer olsun. Bu durumda n gözlemden alınan p değişkenin değerlerini gösteren veri matrisine Y dersek, kayıp verisi olmayan dikdörtgen veri seti $Y = (y_{ij})$' lere tam veri (complete data) denir. Burada, $i = 1,2, \dots, n$ ve $j = 1,2, \dots, p$' dir. Tam-veri matrisi aşağıdaki gibi verilmektedir. Veri matrisindeki y_{ij}' ler gerçek sayılardır.

$$Y = \begin{bmatrix} y_{11} & \cdots & y_{1p} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{np} \end{bmatrix}_{nxp}$$

2.2.2 Tam-olmayan veri (Incomplete-Data)

Araştırmalar sonucunda elde edilen veri setlerini, vektör yapısında ele almak mümkündür. Bu durumda, gözlenebilen veriler $\underline{Y}_{obs}$ vektörüyle, gözlenemeyen veriler ise $\underline{Y}_{miss}$ vektörü ile tanımlansın. Belli bir tanımlama sonucunda bir $\underline{Y}$ vektörü $\underline{Y} = \left[\underline{Y}_{obs}, \underline{Y}_{miss}\right]^t$ biçiminde ele alınabilir. Böylece $\underline{Y}$ veri vektörü tam veri, $\underline{Y}_{obs}$ ise tam-olmayan veri vektörü ya da kısaca tam-olmayan veri (incomplete data) olarak isimlendirilmektedir. Burada, $\underline{Y} = \left[\underline{Y}_{obs}, \underline{Y}_{miss}\right]^t$ birlikte alındığında tam veri değerini oluşturur. $\underline{Y}_{miss}$ değeri bizim tarafımızdan bilinmediği için $\underline{Y}_{obs}$ gözlemlenen (observed) veriler tam-olmayan (incomplete) veridir. Aşağıdaki matris formatındaki daireler kayıp değerleri belirtir. Tüm matris, karşılık gelen mevcut değerlerle verinin gözlenen bir $\underline{Y}_{obs}$ bölümüne ve eksik değerlerin bilinmeyen gerçekleşmeleri ile verilerin kayıp bir $\underline{Y}_{miss}$ kısmına bölünebilir.

$$Y = (y_{ij}) = \begin{bmatrix} y_{11} & y_{12} & \cdots & \circ & y_{1p} \\ \vdots & \circ & & \ddots & \circ \vdots \\ y_{n1} & y_{n2} & \circ & \cdots & y_{np} \end{bmatrix}_{nxp} = \left[\underline{Y}_{obs}, \underline{Y}_{miss}\right]^T$$

Eğer $\underline{Y} = \left[\underline{Y}_{obs}, \underline{Y}_{miss}\right]^t$ tam veri değeri, $\underline{Y} = \underline{Y}_{obs}$ ise örneklem verileri tamamen gözlemlenmiştir (completely observed) diyebiliriz (Buuren 2012).

2.2.3 Kayıp veri (Missing Data)

Veri matrisinde y_{ij}, j. değişkenin i. gözlemde aldığı değer olsun. Burada, $i = 1,2,\ldots,n$ ve $j = 1,2,\ldots,p$' dir. $M = (m_{ij})$ olacak şekilde kayıp veri gösterge matrisi (missing data indicator matrix) tanımlayalım. Burada m_{ij}' lerin aldığı değer; y_{ij} değişkeni kayıp

veri ise $m_{ij} = 1$ ve y_{ij} değişkeni mevcut veri ise $m_{ij} = 0$ değerini alsın. Kayıp veri gösterge matrisi M, aşağıdaki gibi verilmektedir.

$$M = \begin{bmatrix} m_{11} & \cdots & m_{1p} \\ \vdots & \ddots & \vdots \\ m_{n1} & \cdots & m_{np} \end{bmatrix}_{nxp} = \begin{bmatrix} 1 & 0 & 1 & \cdots & 1 \\ \vdots & & \ddots & & \vdots \\ 0 & 1 & 1 & \cdots & 1 \end{bmatrix}_{nxp}$$

Bu durumda n gözlemden alınan p değişkenin değerlerini gösteren veri matrisine Y dersek, kayıp verisi olan dikdörtgen veri seti $Y = (y_{ij})$' lere kayıp veri (missing data) denir.

Literatürde kayıp veri ile ilgili karşılaşılan iki terim ile ilgili olarak, Enders (2010) kayıp veri yapıları (patterns) ile kayıp veri mekanizmaları (mechanisms) arasındaki ayrımı iyi yapmanın yararını vurgulamıştır. Bu terimlerin aslında çok farklı anlamlara sahip olmasına rağmen, araştırmacılar bazen bunları birbirlerinin yerine kullandıklarını tespit etmiştir. Kayıp veri yapıları, bir veri kümesinde gözlenen ve kayıp değerlerin yapılandırılmasını ifade ederken; kayıp veri mekanizmaları ölçülen değişkenler ve kayıp veri olasılığı arasındaki olası ilişkileri açıkladığını belirtmiştir. Kayıp veri yapısının sadece verilerdeki "boşlukların" yerini açıkladığını ve verinin neden kayıp olduğunu açıklamadığına dikkat çekmiştir. Kayıp veri mekanizmaları, kayıp veriler için nedensel bir açıklama sunmasa da, veri seti ve kayıp değer arasındaki genel matematiksel ilişkileri sunmaktadır (örneğin, bir anket tasarımında, eğitim düzeyi ile kayıp veri eğilimi arasında sistematik bir ilişki olabilir). Bahsi geçen bu iki terimin tanımlarını daha detaylı göstermeye çalışalım.

2.3 Kayıp Veri Yapıları (Pattern)

Veri matrisi, n gözlemden alınan p değişkenin değerlerini gösteren, kayıp verisi olmayan dikdörtgensel $Y = (y_{ij})$‘ ler

$$Y = \begin{bmatrix} y_{11} & \cdots & y_{1p} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{np} \end{bmatrix}_{nxp}$$

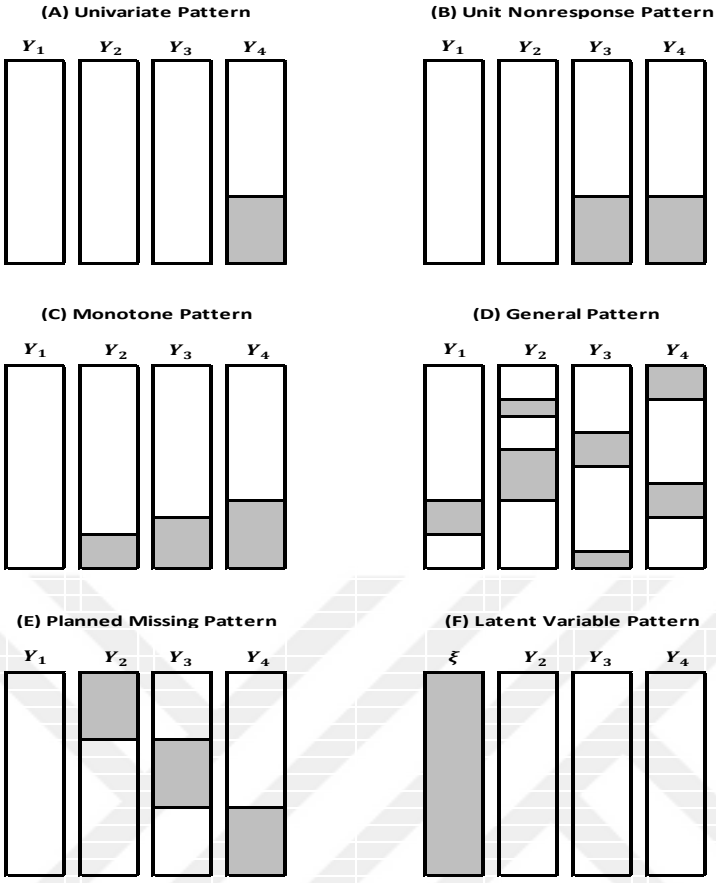
şeklinde olmak üzere, veri kümesindeki kayıp veriyi belirleyebilmek için gösterge matrisi M,

$$M = \begin{bmatrix} m_{11} & \cdots & m_{1p} \\ \vdots & \ddots & \vdots \\ m_{n1} & \cdots & m_{np} \end{bmatrix}_{nxp} = \begin{bmatrix} 1 & 0 & 1 & \cdots & 1 \\ \vdots & & \ddots & & \vdots \\ 0 & 1 & 1 & \cdots & 1 \end{bmatrix}_{nxp}$$

şekildeki gibi elde edildiğinde; kayıp veri yapısı kolaylıkla görülür. Kayıp veri literatüründe altı çeşit veri yapısı ile karşılaşılmaktadır. Veri setinde kayıp verinin oluşturduğu kısımlar taranarak gösterildiğinde kayıp veri yapıları **Şekil 2.1** deki gibi sınıflandırılmaktadır, burada gölgeli alanlar veri kümesindeki kayıp değerlerin yerini temsil etmektedir.

2.3.1 Tek değişkenli yapı (Univariate Pattern)

Şekil 2.1 de bulunan Panel A' daki tek değişkenli yapı (univariate pattern), veri kümesindeki değişkenlerden sadece tek bir değişkende kaybın oluştuğu yapıdır. Bazı çalışmalarda tek değişkenli bir yapı ile ender karşılaşılır, ancak genellikle deneysel çalışmalarda ortaya çıkabilir. Örneğin, dört değişkenli veri seti Y_1, Y_2 ve Y_3' ün manipüle edilmiş değişkenler (örneğin, bir $ANOVA$ tasarımındaki denekler arası faktörler) ve Y_4' ün tamamlanmamış sonuç değişkeni olduğunda, tek değişkenli yapıyı göstermektedir (Enders 2010).



Şekil 2.1 Kayıp Veri Yapısı. Altı tipik kayıp veri yapısı. Gölgeli alanlar, dört değişkenli veri kümesindeki kayıp değerlerin konumunu temsil eder (Enders 2010)

2.3.2 Birim cevapsızlık yapısı (Unit Nonresponse Pattern)

Panel B, birim cevapsızlık yapısı (unit nonresponse pattern) olarak bilinen kayıp değerlerin bir yapılandırmasını gösterir. Burada, aynı gözlemlerin bazı değişkenlerdeki değerleri elde edilemez. Bu yapı genellikle Y_1 ve Y_2' nin örnekleme çerçevesinin her üyesi için kullanılabilen özellikler olduğu ve Y_3 ve Y_4' ün aynı belli birimlerinde kayıpların olduğu bir anket araştırmasında ortaya çıkar. Örneğin, sayım yolu ile yapılan bir araştırmada Y_1 ve Y_2' nin verilerinin kayıp olmayan verilerine karşılık, Y_3 ve Y_4' ün bazı katılımcıların cevap vermeyi reddettiği sorular olduğu ve bu retlerin oluşturduğu kayıpların aynı gözlemlerden oluştuğu anket araştırmasında ortaya çıkar (Enders 2010).

2.3.3 Tekdüze yapı (Monotone Pattern)

Genellikle tekrarlanan ölçümlerde veya uzun süreli veri kümelerinde bir denek belirli bir zaman diliminde çalışmadan ayrılırsa, verileri doğal olarak sonraki tüm zaman dilimlerinde kayıp olacaktır. Panel C' deki tekdüze kayıp veri yapısı (monotone pattern) tipik olarak katılımcıların ayrıldığı ve asla geri dönmediği uzun süreli kesitsiz bir çalışma ile ilişkilidir (literatürde bazen aşınma olarak ifade edilir). Örneğin, yeni bir ilaç için, katılımcıların ilaca olumsuz reaksiyonları olduğu için çalışmayı bıraktıkları klinik bir çalışma düşünüldüğünde, belirli bir değerlendirmede kayıp veri bulunan olaylar, her zaman sonraki ölçümlerden kayıp olur (Enders 2010). Maksimum olabilirlik ve çoklu imputasyonun matematiksel karmaşıklığını büyük ölçüde azalttığı ve yinelemeli tahmin algoritmalarına olan ihtiyacı ortadan kaldırabildiği için monoton kayıp veri yapıları dikkat çekicidir (Schafer 1997).

2.3.4 Genel yapı (General Pattern)

Genel kayıp veri yapısı, kayıp değerlerin en yaygın yapılandırmasıdır. Panel D' de görüldüğü gibi, genel yapı (general pattern), veri matrisi boyunca gelişigüzel bir tarzda dağılmış kayıp değerlere sahiptir. Görünüşte rastgele yapı yanıltıcıdır, çünkü değerler hala sistematik olarak kayıp olabilir. Örneğin, Y_1 değerleri ile Y_2' deki kayıp verilerin eğilimi arasında bir ilişki olabilir. Kayıp veri yapısının; kayıplık nedenlerini değil, kayıp değerlerin yerini tarif ettiğini hatırlamak önemlidir (Enders 2010).

2.3.5 Dosya eşleşme yapısı (File Matching Pattern)

Kasıtlı olarak kayıp veriler üreten bir dizi tasarımını ele alalım. Burada değişkenlerin herhangi biri tamamen gözlenmiş, diğer değişkenler ise birbirleriyle gözlem sayısı kadar gözlenememiş olsun. Dosya eşleşme yapısı (file matching pattern), değişkenler arasındaki ilişkiler tam veri olmadan elde edilebilecek şekilde form yapısına göre özel olarak belirlenir. E Panelinde planlanan kayıp veri yapısı (planned pattern), Graham vd. (1996) tarafından özetlenen üç formlu anket tasarımına karşılık gelmektedir. Üç formlu

tasarımın arkasındaki temel fikir, anketleri farklı formlara dağıtmak ve her katılımcıya formların bir alt kümesini uygulamaktır. Örneğin, Panel E' deki tasarım dört anketi üç forma dağıtır, böylece her form Y_1 içerir, ancak Y_2, Y_3 veya Y_4 eksiktir. Planlanan kayıp veri yapıları, aynı anda katılımcı yükünü azaltırken çok sayıda anket öğesinin toplanmasında yararlıdır (Enders 2010).

2.3.6 Gizli değişken yapısı (Latent Variable Pattern)

Değişkenlerden biri örneklemin tamamında gözlenmediği, Panel F' deki gizli değişken yapısı (latent variable pattern), yapısal eşitlik modelleri (structural equation models) gibi gizli değişken analizlerine özgüdür. Bu yapıda gizli (latent) değişkenlerin değerleri tüm örnek için kayıptır. Örneğin, bir açıklayıcı faktör analizi modeli (confirmatory factor analysis), bir dizi belirgin gösterge değişkeni (örneğin, Y_1 ile Y_3) arasındaki ilişkileri açıklamak için gizli bir faktör kullanır, ancak skorlar faktörü tamamen kayıptır. Gizli değişken yapıları, kayıp veri problemi olarak görmek gerekli olmasa da, araştırmacılar bu modelleri tahmin etmek için kayıp veri algoritmalarını uyarlamışlardır (Enders 2010).

2.4 Kayıp Veri Mekanizması (Mechanism)

Kayıp verileri ortadan kaldırmak için kullanılan yöntemler kayıp verilerin oluşum mekanizmasına bağlıdır. Kayıp veri mekanizmasının iyi bilinmesi kayıp veri probleminin çözümünde doğru analiz yöntemi olur.

Allison (2001) kayıp verilerle ilgili sınıflamayı dört başlık altında aşağıdaki şekilde yapmıştır.

2.4.1 Tamamıyla rasgele kayıp (Missing Completely at Random, $MCAR$)

Belirli bir Y değişkeninde kayıp veri olduğunu varsayalım. Y' deki verilerin kayıp olma olasılığı, Y' nin kendisinin değeri veya veri kümesindeki diğer değişkenlerin değerleri

ile ilgisiz ise, Y üzerindeki verilerin tamamen rastgele kayıp ($MCAR$) olduğu söylenir. Bu varsayım tüm değişkenler için karşılandığında, tam veriye sahip bireyler kümesi, orijinal gözlem kümesinden basit bir rastgele alt örnek olarak kabul edilebilir. $MCAR$' ın Y' deki kayıbın başka bir X değişkenindeki kayıp ile ilişkili olmasına izin verdiği unutmamalıdır. Kayıp olma durumunun belirlenen değişkenlerle ilgili değildir. Örneğin, yaşlarını rapor etmeyi reddeden insanlar gelirlerini rapor etmeyi reddetseler bile, verilerin rastgele tamamen kayıp olması yine de mümkündür. $MCAR$ koşullarının sağlanıp sağlanmadığı genellikle cevap verenler ile vermeyenler arasında gözlemlerin dağılımlarının karşılaştırılmasıyla test edilebilir.

2.4.2 Rasgele kayıp (Missing at Random, MAR)

Analizdeki diğer değişkenleri kontrol ettikten sonra, Y' deki verilerin kayıp olma olasılığı Y' nin değeri ile ilgisiz ise, Y üzerindeki verilerin rastgele kayıp (MAR) olduğu söylenir ve oldukça zayıf bir varsayımdır. X' in her zaman gözlemlendiği ve Y' nin bazen kayıp olduğu yalnızca iki X ve Y değişkeni olduğunu varsayıldığında, hem Y hem de X verildiğinde Y ile ilgili verilerin kayıp koşullu olasılığının, sadece X ile verilen Y üzerindeki kayıp verilerin olasılığına eşit olduğu anlamına gelir. MAR durumunun karşılanıp karşılanmadığını test etmek imkansızdır ve nedeni sezgisel olarak açık olmalıdır. Kayıp verilerin değerlerini bilmediğimiz için, kayıp veriler olan ve olmayanların değerlerini, bu değişken üzerinde sistematik olarak farklılık gösterip göstermediklerini görmek için kıyaslayamayız.

2.4.3 İhmal edilebilir (Ignorable, I)

Kayıp veri mekanizmasının, (a) verilerin MAR olması ve (b) kayıp veri sürecini yöneten parametrelerin tahmin edilecek parametrelerle ilgisiz olması durumunda ihmal edilebilir kayıp (I) olduğu söylenir. İhmal edilebilir kayıp temel olarak, kayıp veri mekanizmasının tahmin sürecinin bir parçası olarak modellenmesine gerek olmadığı anlamına gelir. Bununla birlikte, verileri verimli bir şekilde kullanmak için kesinlikle özel tekniklere ihtiyaç vardır.

2.4.4 İhmal edilemez (Nonignorable, NI)

Veriler MAR değilse, kayıp veri mekanizmasının ihmal edilemez kayıp (NI) olduğunu söylenir. Bu durumda, ilgili parametrelerin iyi tahminlerini elde etmek için genellikle kayıp veri mekanizması modellenmelidir. İhmal edilemez kayıp veriler için yaygın olarak kullanılan bir yöntem, bağımlı değişken üzerinde seçim yanlılığı olan regresyon modelleri için Heckman' ın (1976) iki aşamalı tahmin edicisidir. Ne yazık ki, ihmal edilemez kayıp verilerle etkili bir tahmin için genellikle kayıp veri sürecinin doğası hakkında çok iyi bir ön bilgiye ihtiyaç vardır, çünkü veriler hangi modellerin uygun olacağı hakkında bilgi içermez ve sonuçlar tipik olarak model seçimine karşı çok hassas olacaktır.

2.5 Kayıp Veri Çözümleme Yöntemleri

Kayıp veri durumunda çözümleme yöntemi olarak literatürde karşılaşılan temel iki yöntem, veri silme ve veri atama yöntemleri olarak aşağıdaki gibi tanımlanmıştır.

2.5.1 Veri silme yöntemleri

Kayıp gözlemlerin veri setinden çıkartılması örneklem hacminde kritik ve hassas bir azalmaya yol açabilir ve veri silme durumu analizlerdeki kitle parametresinin kitleyi temsil etme becerisinde azalmaya neden olacaktır. Örneğin, sağlık sektöründeki karşılaşılan zamana ve hastalık çeşitlerine bağlı, az sayıda gözlem imkanı olan hastalık ve tedavi için araştırılan çok değişkenli yapıdaki veride; kayıp verili satıra ait, tüm satırsal gözlemlerin veri setinden çıkartılması örneklem hacminde önemli bir azalmaya yol açabilir ve yeterli büyüklükte düzenlenmiş bir örneklem, yetersiz büyüklükteki bir örnekleme dönüşebilir.

Kitlenin homojenliği fazla ise alınması gereken örneklem büyüklüğünün azaldığı bilinmektedir. Ancak kitledeki birimler arasında farklılaşma büyüdükçe doğru istatistiksel analizler için büyük örneklem hacmi almak gerekmektedir. Büyük örneklem

boyutu, standart hatayı azaltır. Standart hata büyürse, onun kitlenin parametresini tahmin etme ya da kitleyi temsil etme becerisi azalmaktadır. Kayıp veri olması durumunda n örneklem hacmi azalır; bu yanlılığa sebep olabilir, standart hatayı artırabilir bu da kitlenin parametresini tahmin etme ya da kitleyi temsil etme becerisini azaltmaktadır.

Ayrıca, kayıp veri içeren hücrelerin veri setinden çıkartılabilmesi için kayıp verilerin tamamen rastgele olarak dağılıyor olması gerekmektedir. Kayıp verilerin analize dahil edilen başka değişkenlerle ilişkili olduğu durumlarda yapılacak olan kayıp olmayan verilerin silme işlemi önemli bir yanlılığa yol açabilir (Tabachnick ve Fidell 1996, Schafer 1999, Osborne 2013).

2.5.1.1 Liste/Durum düzeyinde veri silme

Bu yöntem, kayıp veriler içeren veri setinde bir indirgeme yapma esasına dayanır. Yani, veri kümesinin her bir satırı tek tek ele alınır ve eğer kayıp veri içeren hücre ya da hücreler varsa bu kayıt tümüyle silinerek veri kümesinden çıkarılır. Özetle, analize alınan değişkenlerin tüm durumlarında kayıp veri olmayan kayıtlar ele alınır. Bu işlem veri setinde yer alan tüm birimler için yinelenir. Böylece, tümüyle gözlenmiş birimlerden oluşan temiz bir veri kümesi elde edilir. Daha sonra yapılacak olan hesaplamalarda ve istatistiksel analizlerde kalan veri kümesi üzerinden işlemler yapılır.

Liste/durum düzeyinde veri silmenin olumsuz yönleri aşağıdaki gibi verilebilir;

- Veri setindeki azalma,
- Tamamıyla rasgele kayıp olmadığı durumlarda yanlı sonuç vermesi.

2.5.1.2 Çiftler düzeyinde veri silme

Bu yöntem liste bazında veri silme yöntemine benzer olarak kayıp veriler içeren veri setinde bir indirgeme yapma esasına dayanır. Veri setinin her bir satırı tek tek ele alınır.

Ancak liste bazında veri silme yönteminden farklı olarak, varyans kovaryans matrisi ve ortalama vektörü tahminleri için değişkenler ikişer ikişer ele alınır; her ikisi de gözlenmiş olan gözlemler dışında kalan gözlemler veri kümesinden silinir. Geri kalan gözlem çiftleri üzerinden kovaryans ve ortalama tahminleri hesaplanır. Çiftler düzeyinde veri silme yöntemi bu yönüyle liste/durum düzeyinde veri silme yönteminden üstündür.

Çiftler düzeyinde veri silme yönteminde karşılaşılan zorluklar;

- Kovaryans matrisi ile çalışıyor olması,
- Korelasyon matrisinin tersinir olmadığı durumlarda yaşanan problem,
- Serbestlik derecesinin işlemsel olarak hesaplanmasının zorluğu,

olarak verilebilir.

2.5.2 Veri atama yöntemleri

Kayıp veriler yerine yaklaşık değer atama yöntemleri, araştırmacıların hem zamandan ve emekten tasarruf edecekleri hem de topladıkları verileri koruyabilmelerini sağlayacak bir yol olarak ortaya çıkmaktadır. Ancak, Little ve Rubin' e göre (1987), kayıp veriler yerine bilinçsizce atanan değerler var olan sorunları ortadan kaldırmadığı gibi ortaya çözümü daha güç olan yeni sorunlar çıkarmaktadır.

Kayıp verilerin makul değerlerle "doldurulması" uygulaması olan imputasyon ile tamamlanmamış verilerin analizinde kullanılan birçok yöntem bulunmaktadır. Bu yöntemler arasında bulunan Ortalama İmputasyon (mean imputation), En Yakın Komşuluk İmputasyonu (nearest neighbor), Regresyon İmputasyon. (regression imputation) gibi yöntemler basit atamaya dayalı yöntemler olarak adlandırılmaktadır. Ayrıca bu yöntemler arasında daha gelişmiş yöntemler olarak nitelendirilen en çok olabilirlik kestirimine dayalı En Çok Olabilirlik Yöntemleri (ML) ve EM-Algoritması (expectation maximization) gibi yöntemler de bulunmaktadır.

Schafer' a (1999) göre, kayıp veri miktarının az sayıda ve tamamen rastgele olarak dağıldığı durumlarda basit atamaya dayalı imputasyon, başlangıçta kayıp veri problemini çözerek çalışmanın ilerlemesini sağlar. Fakat, çoklu atamaya dayalı kayıp veri durumunda çoklu imputasyonun (MI, Multiple Imputation), kayıp değerleri işlemek için en iyi yöntem olduğu söylenemez. Bazı durumlarda, iyi tahminler ağırlıklı bir tahmin prosedürü ile elde edilebilir. Tam parametrik modellerde, maksimum olabilirlik tahminleri genellikle EM algoritması gibi özel sayısal yöntemlerle kayıp verilerden doğrudan hesaplanabilir. Bu prosedürler yoluyla elde edilen tahminler, çoklu imputasyondan elde edilenlerden biraz daha verimli olabileceği belirtilmiştir. Bunun üzerine Osborne (2013), kayıp verilerin oranı küçük olduğunda ve tahmin iyi olduğunda, tek imputasyon muhtemelen yeterli olacağı, ancak çoklu regresyon yoluyla yapılan herhangi bir tahminde olduğu gibi, verileri "geçersiz kılar" ve orijinal verilerin sahip olacağından daha az genelleştirilebilir sonuçlara yol açabileceğini belirtmiştir (Osborne 2000, 2008; Schafer 1999). Çoklu imputasyonun avantajı genelleştirilebilirlik ve tekrarlanabilirliktir, yani kayıp veriyi açıkça modeller ve araştırmacıya tek bir imputasyona güvenmek yerine tahminler için güven aralıkları vereceğini ileri sürer.

Günümüzde, istatistiksel programlama dilleri kullanıcılara kayıp verilerle başa çıkma olanağını çeşitli yöntemlerle sağlamaktadır. Analizler veriye dayalı olduğu için henüz tek veya çok değişkenli veri yapısında hangi miktardaki kayıp veriler için hangi yöntemin kullanılmasının daha uygun olacağına dair net yönlendirmeler mümkün değildir. Kayıp veri durumunda kullanılan imputasyon yöntemlerini özetle aşağıdaki gibi basit atamaya dayalı yöntemler ve daha gelişmiş yöntemler olarak nitelendirilen model tabanlı imputasyon yöntemler başlıkları altında tanımlamaya çalışalım.

2.5.2.1 Basit imputasyon yöntemler

2.5.2.1.1 Tümdengelimli imputasyon (Deductive imputation)

Tümdengelimli imputasyon yöntemle, önceki anketlerdeki benzer öğeler, mevcut anketlerdeki ilgili öğeler vb. gibi mevcut bilgilerden yararlanarak kayıp değer için

imputasyon yapılır. Bu yöntemi uygulamak için, kullanıcının kayıp değer ile diğer kaynaklardan değerler arasında bir miktar belirleyici ilişki bulması gerekir. Örneğin, Cold Deck, önceki benzer anketlerden gelen bilgileri kullanan bir tümdengelimli imputasyon yöntemidir. Genel olarak, bir anketteki tüm kayıp öğeleri tümdengelimli imputasyon kullanarak ispatlamak için yeterli bilgi bulmak imkansızdır ancak, bu yöntem bazı kayıp değişkenleri hesaplamak için kullanılabilir. Mümkün olduğunda, kayıp durumlar için doğru veya yaklaşık olarak doğru imputasyonlar sağladığından, tümdengelimli imputasyon diğer herhangi bir imputasyon yönteminden önce kullanılmalıdır. Ancak, tümdengelimli bir imputasyon yönteminin performansı tamamen mevcut kaynaklara bağlıdır (Hu ve Salvucci 2001).

Tümdengelimli imputasyonları gerçekleştirmek için çok önemli bir varsayımının, verilerin yalnızca kayıp değerleri içermesi, fakat hata içermemesi gerektiği unutulmamalıdır.

2.5.2.1.2 İkame (Substitution)

İkame (yerine koyma), bir araştırmanın saha çalışması aşamasında birim yanıt vermemesi ile başa çıkma yöntemi olup, yanıtsız birimleri örneklemde seçilmeyen alternatif birimlerle değiştirir. Örneğin, bir hane ile iletişim kurulamıyorsa, aynı konut bloğunda önceden seçilmemiş bir hane ikame edilebilir. Elde edilen örneği tam olarak işleme eğilimine karşı koyulmalıdır, çünkü ikame edilen birimler yanıtlayıcılardır ve bu nedenle yanıt vermeyenlerden sistematik olarak farklılık gösterebilir. Bu nedenle, analiz aşamasında, ikame edilmiş değerler, belirli bir modeldeki imputasyon değerleri olarak kabul edilmelidir (Little ve Rubin 1987).

2.5.2.1.3 Ortalama imputasyon (Mean imputation)

Ortalama imputasyon yöntem, kayıp veri tahmini için uygulanması sıkça kullanılan kolay bir yöntemdir. Kayıp veriler yerine diğer tüm verilerin genel örnek ortalaması atanarak kayıp veriler tamamlanır. Ortalama imputasyon yöntemi ilk önce bazı yardımcı

değişkenleri kullanarak imputasyon hücrelerini oluşturur ve sonra her hücredeki kayıp değerleri örnek ortalamasıyla değiştirir. Ayrıca kayıp veriler mod veya medyan değerleri kullanılarak da tahmin edilebilir. Medyanın aykırı değerlerden etkilenmemesi, ortalamaya göre güçlü bir özelliği olduğundan; aykırı değerlerin bulunduğu veri setlerinde ortama yerine medyan kullanmak daha sağlıklı olacaktır.

Ortalama imputasyon yöntem, kayıp değerler tamamen rastgele kayıpsa ($MCAR$) ya da kayıp değerler yalnızca imputasyon hücrelerini oluşturmak için kullanılan yardımcı değişkenlere bağlıysa; kitle ortalamaları veya toplamları için yansız tahminler sağlayabilir. Kayıp verilerin diğer verilerin ortalaması ile tamamlanması sonucunda gözlem değerlerinin ortalama etrafında yığılması durumu ayrı bir tartışma konusudur. Ayrıca, ortalama imputasyonun tek değişkenli analiz için bazı çekici özellikleri vardır, ancak çok değişkenli analiz için sorunlu olacağı söylenebilir. Çok büyük veri kümelerinde kayıp veri oranı da düşük ise ortalama etrafında yığılma durumu söz konusu olmayacaktır.

Bununla birlikte, ortalama imputasyondan sonra verilerin dağılımı önemli ölçüde bozulur ve hücre ortalamasındaki tüm imputasyon edilen değerlerin yoğunluğu dağılımda ani artışlar yaratır. Bu nedenle, quantil tahminleri yanlı olacak ve varyanslar gerçekte olduğundan daha az önemli olacak şekilde tahmin edilecektir. Cohen (1996), kayıp durumlar için daha çeşitli değerler yükleyerek varyans tahminlerini düzenlemenin başka bir yolunu önermiştir. Örneğin, tüm kayıp değerlerin ortalamasını imputasyon yerine, birinci ve ikinci momentleri koruyacak şekilde kayıp değerleri ikiye bölerek imputasyon önermiştir (Hu ve Salvucci 2001).

2.5.2.1.4 Hot/Cold deck imputasyon (Hot/Cold deck imputation)

Hot deck imputasyon, kayıp değerin başka matematiksel ve istatistiksel bilgi olmaksızın veri setinden elde edilmesini sağlayan önemli ve birçok uygulayıcı için sezgisel olarak anlamlı bir yöntemdir. Anında işlenebildiği için "hot" atama olarak düşünebiliriz.

Bu yöntemde, kayıp veriler, mevcut verilerden kayıp değer için tahmini bir dağılımdan elde edilen değer ile doldurulur. Bu işlem bir kümeye ait verilerin ortalaması veya modu hesaplanarak gerçekleştirilir. Rasgele Hot Deck imputasyonda; kayıp veri, aynı kümedeki diğer verilerden biri rasgele seçilerek tamamlanır. Cold Deck imputasyon da hot deck imputasyona benzer yönteme sahiptir; ancak seçim yapılan kaynak mevcut veri kaynağından farklı olmalıdır (Acuna ve Rodrıguez 2004).

Geleneksel hot-deck prosedürü olarak bilinen atama ise; imputasyon sınıflarının spesifikasyonu ile başlar ve her sınıfa, yöntem için bir başlangıç noktası sağlamak üzere y değişkenine tek bir değer atanması ile başlar. Bu başlangıç değerleri, örneğin, her sınıf için bir yanıtlayıcı değeri veya anketin önceki bir turundan sınıf ortalaması gibi temsili bir değer alınarak elde edilebilir. Mevcut anketin kayıtları daha sonra sırayla işlenir. Bir kaydın y değişkeni için bir yanıtı varsa, bu değer daha önce imputasyon sınıfı için depolanan değerin yerini alır. Kaydın kayıp bir yanıtı varsa, imputasyon sınıfı için o anda depolanan değer atanır. Bu prosedürün önemli bir avantajı, bilgi işlem ekonomisidir, çünkü tüm imputasyonlar veri dosyasından tek bir geçişle yapılır. Hot-deck yönteminin bir dezavantajı, anket tahmin edicileri için hassasiyet kaybına yol açan bir özellik olan birden fazla donör kullanımına kolaylıkla yol açabilmesidir. Bu, belirli bir imputasyon sınıfı içinde, kayıp yanıtı olan bir kaydın ardından kayıp yanıtları olan bir veya daha fazla kayıt geldiğinde oluşur; tüm bu kayıtlara daha sonra sınıftaki son katılımcıdan alınan değer atanır. Sınırsız donörlerin örneklemesine sahip sınıf içi rastgele yöntem bu dezavantajı paylaşır. Rastgele sınıf içi yöntemle, bununla birlikte, donörlerin çoklu kullanımı, donörlerin değiştirilmeden örneklenmesi ile en aza indirilebilir (Kalton ve Kasprzyk 1982).

2.5.2.1.5 Son gözlemi ileri taşıma (Last Observation Carried Forward)

Son gözlemi ileri taşıma imputasyonu, hot deck imputasyonun özel bir yöntemi olarak karşımıza çıkar. İleri taşınan son gözlem ($LOCF$, Last Observation Carried Forward) değeri, bir veri kümesinin çeşitli değişkenlerden herhangi birine göre sıralanması ve böylece sıralı bir veri kümesi oluşturulmasıyla elde edilir.

Bu yöntem, ilk kayıp değeri bulur ve kayıp değeri belirlemek için kayıp olan veriden hemen önceki hücre değerini kullanır. İşlem, tüm kayıp değerler atanana kadar, kayıp değeri olan sonraki hücre için tekrarlanır.

Son gözlemi ileri taşıma ataması ölçüm kayıpsa, en iyi tahminin, son ölçüldüğünden beri değişmediği inancını temsil eder. Bu yöntemin, artan yanlılık ve potansiyel olarak yanlış sonuçlara varma riskini artırdığı için kullanılması tavsiye edilmez (Molnar vd. 2008).

2.5.2.1.6 En yakın komşuluk imputasyonu (Nearest Neighbor Imputation)

En yakın komşuluk imputasyonu (Nearest Neighbor Imputation, NNI), genel olarak bazı veri setlerindeki kayıp değerin tüm veri setindeki ilgili durumlardan elde edilen bir değerle değiştirildiği kayıp verileri doldurmak için kullanılan yöntemlerden biridir. En yakın komşuluk imputasyonunun, hot deck imputasyonun deterministik bir formu olduğu söylenebilir. Kayıp verileri gerçek değere mümkün olduğu kadar yakın olan makul değerlerle ikame etme yeteneğinin yanı sıra, imputasyon algoritmaları orijinal veri yapısını korumalı ve imputasyon edilen değişkenin dağılımını bozmaktan kaçınmalıdır.

Bu yöntemin özel bir hali, kayıp verileri tahmin etmek ve değiştirmek için k en yakın komşu imputasyon (k Nearest Neighbor Imputation, $kNNI$) algoritmalarını kullanır. Yöntemin avantajlarından biri, hem niteliksel verileri (k en yakın komşu arasında en sık görülen değer) hem de nicel verileri (k en yakın komşunun ortalaması) tahmin edebilir olmasıdır. Bir başka avantajı olarak, görünür modeller oluşturmasa bile, kayıp verilere sahip her özellik için bir tahmine dayalı model oluşturmak gerekli değildir. Bu yöntem için en büyük sıkıntı etkinlik (efficiency) olduğu düşünülmektedir. k en yakın komşuluk algoritması birbirine en benzer örnekleri bulmaya çalışırken, tüm veri seti araştırılmaktadır. Bununla birlikte, veri setleri genellikle arama için aşırı büyük olabilir. Öte yandan, k değerinin seçilme şekli ve benzerlik ölçeği elde edilecek sonuçları oldukça etkileyecektir (Peng ve Lei 2005).

2.5.2.1.7 Regresyon imputasyon (Regression imputation)

Regresyon imputasyon, bir değişkenin değerinin diğer değişkenler ile doğrusal bir şekilde değiştiğini varsayar. Bu yöntem kayıp değerleri istatistiklerle tamamlamak yerine, doğrusal regresyon modelinden elde edilen tahminleri kullanır. Bu yöntem, değişkenler arasındaki doğrusal ilişki varsayımına dayanır. Ancak çoğu durumda değişkenler arası ilişki doğrusal değildir. Bu sebeple kayıp verileri doğrusal bir şekilde tahmin etmek, yanlı bir model olacaktır (Peng ve Lei 2005).

Regresyon imputasyon, regresyon modellerine dayandığından, X ve Y ilişkileri korunur. Bu nedenle, ortalama imputasyon gibi daha basit imputasyon yöntemlerine göre avantaj sağladığı söylenebilir.

Kestirim regresyon imputasyon yöntemi her bir imputasyon sınıfı içinde de uygulanabilir. Bu durumda, oluşturulan imputasyon sınıflarından yararlanılır ve yöntem bu sınıflar içinde uygulanır. Bu yöntemin dezavantajı, dağılımın ortalama değerini daraltmasıdır. Değişkenliği artırmak için tahmin edilen değerlere, imputasyon olarak küçük rastgele artıklar eklenebilir (Hu ve Salvucci 2001).

2.5.2.1.8 Stokastik regresyon imputasyon (Stochastic regression imputation)

Regresyon imputasyonu iki farklı; deterministik ve stokastik regresyon imputasyon olarak sınıflandırmamız, stokastik regresyon imputasyanonu tanımı için kolaylık sağlayacaktır. Klasik regresyon imputasyon yani deterministik regresyon imputasyon, kayıp değerleri regresyon modelinin tam (exact) tahminiyle değiştirir. Regresyon eğimi etrafındaki rastgele değişim yani hata terimi dikkate alınmaz. Bu nedenle imputasyon değerler genellikle çok kesindir ve X ile Y arasındaki korelasyonun fazla tahmin edilmesine yol açar. Bu deterministik regresyon imputasyon sorununu çözmek için stokastik regresyon imputasyon geliştirilmiştir. X ve Y' nin korelasyonunu daha uygun bir şekilde yeniden hesaplayabilmek için, Stokastik regresyon imputasyonu, tahmin edilen değere rastgele bir hata terimi ekler.

Stokastik regresyon imputasyonu, kayıp değerleri, tahmin edilen değerdeki belirsizliği yansıtmak için çizilmiş hata terimi ilavesiyle regresyon imputasyonu ile tahmin edilen bir değerle değiştirir. Normal lineer regresyon modellerinde, hata terimi, sıfır ortalama ve varyans regresyondaki hata terimi varyansa eşit olacak şekilde doğal olarak normal olacaktır. Lojistik regresyonda olduğu gibi bir ikili sonuçla, tahmin edilen değer 1' e karşı 0 olan bir olasılıktır ve imputasyon değerler, bu olasılıkla çizilmiş bir 1 veya 0' dır. Herzog ve Rubin (1983), hem normal hem de ikili sonuçlar için stokastik regresyon kullanan iki-aşamalı bir prosedürü tanımlar (Little ve Rubin 2002).

2.5.2.2 Model tabanlı yöntemler

2.5.2.2.1 En çok olabilirlik yöntemleri (ML)

En çok olabilirliği ilk kez Edgewort 1908 yılında kullanmıştır. 1921 yılında Fisher, bu yöntem ile bulunan tahmin edicinin varyansı için genel formülü bulmasıyla, bu yöntem daha da önem kazanmıştır (İnal ve Günay 1993).

En çok olabilirlik (ML), birçok tahmin probleminin üstesinden gelmek için yaygın olarak kullanılan istatistiksel tahmine yönelik çok genel bir yaklaşımdır. Örneğin, lojistik regresyon modelini tahmin etmek için ML tercih edilen yöntemdir. Sıradan en küçük kareler doğrusal regresyon, hata teriminin normal dağıldığı varsayıldığında da bir ML yöntemidir. ML' nin özellikle kayıp veri sorunlarını ele alma konusunda daha iyi olduğunu söyleyebiliriz. Allison (2001) geleneksel yöntemlere göre maksimum olabilirlik tahmini ve çoklu veri atama yöntemlerinin daha başarılı olduğunu ortaya koymuştur (Allison 2001).

En çok olabilirlik yöntemi (ML), en iyi birinci ve ikinci dereceden moment tahminlerini oluşturmak için, bir veri setinde gözlemlenen tüm verileri kullanır. Herhangi bir veriye tahmin ataması yapmaz, bunun yerine bir veri setindeki değişkenler arasında bir ortalama-kovaryans matrisi vektörü oluşturur. Bu yöntem, beklenti maksimizasyonu (Expectation Maximization, EM) yaklaşımının geliştirilmiş halidir. Bir avantajı, iyi

bilinen istatistiksel temele sahip olmasıdır. Dezavantajları, orijinal veri dağılımı varsayımını ve tamamlanmamış rastgele kayıp (incomplete missing at random) varsayımını içermesidir (Peng ve Lei 2005).

2.5.2.2.2 EM- Algoritması (Expectation Maximization)

EM algoritması, bazı veriler kayıp olduğunda ML tahminleri elde etmek için çok genel bir yöntemdir (Dempster vd. 1977, McLachlan ve Krishnan 1997). EM olarak adlandırılır çünkü iki adımdan oluşur: bir beklenti adımı (E-adımı (expectation)) ve bir maksimizasyon adımı (M-adımı (maximization)). Bu iki adım, sonunda ML tahminlerine yakınsayan yinelemeli (iterative) bir süreçte birden çok kez tekrarlanır (Allison 2001).

Dempster vd. (1977), EM algoritmasında; E-adımı, gözlenen veriler ve mevcut parametre tahminleri göz önüne alındığında, verilen tam yeterli istatistik verilerinin tamamının beklentisini hesaplar. M-adımı, tam yeterli istatistiklerin mevcut değerlerine dayalı olarak maksimum olabilirlik yaklaşımı aracılığıyla parametre tahminlerini günceller. Algoritma daha sonra, son iki ardışık parametre tahminleri arasındaki fark belirli bir kritere yakınsayana kadar yinelemeli (iteratif) bir şekilde ilerler. Son E-adımı, nihai parametre tahminleri ve gözlenen veriler verilen her bir kayıp değerin beklentisini hesaplar; bu, imputasyon değeri olarak kullanılacaktır (Hu ve Salvucci 2001).

Yinelemeli (iteratif) bir yöntem olan EM algoritması daha açık olarak şu şekilde ifade edilebilir:

(a) Kayıp değerleri tahmini değerlerle değiştirin,
(b) Parametreleri tahmin edin,
(c) Yeni parametre tahminlerinin doğru olduğunu varsayarak kayıp değerleri yeniden tahmin edin (re-estimate),
(d) Parametreleri ve benzerleri yeniden tahmin edin, yakınsama sağlanana kadar yinelemeyi sürdürün (Little ve Rubin 1987).

EM algoritması her bir kayıp verinin imputasyonu için kullanılabilir, ancak daha çok kitle parametreleri için doğrudan tahminler elde etmek için kullanılır. Veriler için normal bir dağılım varsayıldığında, hem E adımındaki yeterli istatistiklerin beklentileri hem de M adımındaki parametrelerin maksimum olabilirlik tahminlerinin elde edilmesi kolaydır. Ancak diğer dağıtımlarla bunu yapmak kolay olmayabilir. Yakınsama yavaş olabilir ve özellikle aralıklı (sparse) verilerle EM algoritması ile yakınsama garanti edilemeyebilir. Her M-adımı aynı zamanda maksimum olabilirlik tahminlerini elde etmek için yinelemeli bir süreç gerektiriyorsa, yakınsama süreci daha da yavaşlayacaktır. Bu yöntem aynı zamanda ortalama değere doğru küçülmeye (büzülmeye) sebep olur. EM algoritmasının avantajı kararlı yakınsamasıdır; yani, yinelemeler her zaman olabilirliği artırır (Hu ve Salvucci 2001).

2.5.2.2.3 Çoklu imputasyon (multiple imputation)

Çoklu imputasyon (MI), rasgelelik içeren veri setinden yardım alarak kayıp veriyi tahmin eder. İşlem tekrarlanarak tamamlanmış veri setleri elde edilir. Elde edilen veri setlerinin ortalamaları alınarak tek veri seti oluşturularak sonuca ulaşılır. MI basit imputasyon yöntemlerine göre genellikle daha iyi ve yansız sonuçlar verir. MI nin olumsuz bir özelliği verinin işlenmesi için ve tahminlerin hesaplanması için daha çok işlem yüküne ve zamana gerek duymasıdır.

Schafer' a (1999) göre, MI , kayıp değerleri ele almak için tek yöntem değildir ve herhangi bir sorun için mutlaka en iyisi değildir, ancak kullanışlı bir araçtır. Bazı durumlarda, ağırlıklı tahmin prosedürü (weighted estimation procedure) ile iyi tahminler elde edilebilir. Tam parametrik modellerde, maksimum olabilirlik tahminleri genellikle doğrudan tamamlanmamış verilerden (incomplete data), EM algoritması gibi özel sayısal yöntemlerle hesaplanabilir. Bu tür prosedürler yoluyla elde edilen tahminler, simülasyon içermedikleri için MI den elde edilen tahminlerden biraz daha verimli olabilir. Yeterli zaman ve kaynak verildiğinde, herhangi bir özel problem için belki de MI den daha iyi bir istatistiksel prosedür elde edilebilir.

Osborne' a (2013) göre, MI, masaüstü bilgisayar gücünün yaygınlığı ile kayıp veri yönetiminde daha yaygın modern seçeneklerden biri olarak ortaya çıkmıştır. Kayıp verilerin oranı küçük ve tahmin iyi olduğunda; çoklu regresyon yoluyla herhangi bir tahminde olduğu gibi, orijinal verilerin sahip olacağından daha az genelleştirilebilir sonuçlara yol açarak verilere "fazla uyuyor (overfits)" olsa da, tek imputasyon muhtemelen yeterlidir (Osborne 2000, 2008; Schafer 1999). MI nin avantajı genelleştirilebilirlik ve tekrarlanabilirliktir, yani kayıp veriyi açıkça modeller ve araştırmacıya tek bir imputasyona güvenmek yerine tahminler için güven aralıkları verir. Son olarak ve önemli olarak, bazı MI prosedürleri, verilerin rastgele kayıp olmasını gerektirmez (örneğin, SAS ta, kayıp verilerin etrafındaki varsayımlara bağlı olarak değerleri tahmin etmek için birkaç seçenek vardır). Diğer bir deyişle, bazı önemli yanlılıklardan kaynaklanan önemli bir kayıp veri parçasının en kötü senaryosu altında, bu prosedür iyi bir alternatif olmalıdır (Schafer 1999).

MI, Bayes ve sıklık paradigmaları, prosedürleri oluşturmak için Bayes model tabanlı yaklaşım ve prosedürleri değerlendirmek için sıklık (rastgele-tabanlı) yaklaşım kullanarak hedefleri karşılamak için tasarlanmıştır. Her şeyden önce, bilimsel tahminler için istatistiksel geçerlilik elde etmek için, nokta tahmini, bilimsel tahminler için yaklaşık olarak yansız olmalı, örnekleme ve varsayılan cevapsızlık mekanizmaları üzerinden ortalama alınmalıdır. İkinci olarak, aralık tahmini ve hipotez testi geçerli olmalıdır. MI, çoklu imputasyon amacı (bazen tekrarlanan atama olarak da adlandırılır), (a) nihai kullanıcıların ve veritabanı kurucularının, farklı analizler, modeller ve yeteneklere sahip farklı varlıklar olduğu ve (b) tipik olarak kayıp veriler için kabul edilen bir nedenin olmadığı, zor gerçek dünya durumunda istatistiksel olarak geçerliliğe sahip sonuç çıkarım sağlamaktır. MI, çoklu imputasyon varyans tahminleri genellikle tek imputasyon varyans tahminlerinden daha büyüktür, çünkü çoklu atama, arada-atama (between-imputation) varyasyonunu ekler. Çoklu atama varyansı tahminleri yansız olarak kabul edilebilir (Hu ve Salvucci 2001).

2.5.2.2.4 Bayesci imputasyon (Bayesian imputation)

Peng ve Lei' e (2005) göre, Naive Bayes Sınıflandıcısı, sadece iyi performansı için değil, aynı zamanda basit formu için de popüler bir sınıflandırıcıdır. Kayıp verilere duyarlı değildir ve hesaplama verimliliği çok yüksektir. Bayes Yineleme İmputasyon, kayıp verileri atamak için Naive Bayes Sınıflandırıcısı' nı kullanır. İki aşamadan oluşur: (a) Bilgi kazanımı, kayıp oranı, ağırlıklı indeks vb. bazı ölçümlere göre işlenecek özelliğin sırasına karar vermek; (b) Kayıp verileri tahmin etmek için Naive Bayes Sınıflandırıcıyı kullanma. İteratif ve tekrar eden bir süreçtir. Algoritmalar, birinci aşamada tanımlanan ilk öznitelikteki kayıp verileri değiştirir ve ardından doldurulan özniteliklerin temelinde bir sonraki özniteliğe döner. Genel olarak, tüm kayıp verilerin değiştirilmesi gerekli değildir ve yineleme süreleri azaltılabilir (Liu vd. 2004).

Bayes sonsal dağılımı genellikle doğrudan simülasyon veya Markov Zinciri Monte Carlo teknikleri kullanılarak hesaplanabilir ve bunlar bilgi matrisini hesaplamaya ve tersine çevirmeye gerek kalmadan standart hataların tahminlerini sağlar. Sonsal dağılıma dayalı çıkarımlar genellikle ML ye dayalı asimptotik çıkarımlardan daha iyi sıklık özelliklere sahiptir. Ayrıca, Bayes hesaplamalarının bir yan ürünü olarak çoklu imputasyonlar oluşturulabilir ve kayıp veri problemlerini çözmek için pratik olarak kullanışlı ve esnek bir araç sağlar. Kayıp veri setinde ML den Bayes yöntemlerine geçiş, kayıp verilerle çok değişkenli normal modelin parametrelerinin sonsal dağılımını oluşturmak için veri büyütmeyi tanımlayan Tanner ve Wong (1987) tarafından başlatılmıştır. Bir diğer önemli gelişme, Rubin' in (1978, 1987, 1996) Bayesci fikirlerle motive edilen çoklu imputasyon (MI) ile kamuya açık veri kümelerindeki kayıp verileri ele alma önerisiydi. Başlangıç aşamasında, bu öneri çok hesaplama açısından pratik görünmüyordu, artık $MCMC$ (Markov Chain Monte Carlo) tekniği, Bayes çoklu imputasyonu kolaylaştırır ve şimdi hem Bayes hem de sıklık türlerinin kullanıcılarının rahatlığı için halka açık yazılımlarda uygulanmaktadır (Little 2011).

2.5.2.2.5 Sağlam imputasyon (robust imputation)

Kayıp veri durumunda, verinin kayıp olmayan kısmından elde edilen tahmin edicilerle veri setinin tamamlandığı yöntemlerde, sağlamlıktan söz edemeyiz. Çünkü veriler aykırı değerlere sahip olabilir. Bu aykırı değerlerin varlığı, kayıp değerler için impute edilen değerler üzerinde olumsuz etkilere sahip olabilir. Kayıp değerin impute sonrasında, bu tamamlanan (completed) veriler üzerindeki herhangi bir istatistiksel analizin sonucu, aykırı değerlerin varlığı sebebiyle yanlış sonuçlara yol açabilir. Bu durumda amaç, kayıp veri atamasında tahmin edici elde ederken, aykırı değerin etkilerini en aza indirgemeye çalışmaktır.

Örneğin, kayıp değerleri atamak için en popüler teknik, kayıp değerlerin bir verinin k en yakın komşusuna (k Nearest Neighbor Imputation, $kNNI$) dayalı bir tahminle değiştirildiği kNNimpute yöntemidir (Troyanskaya vd. 2001). Öklid mesafesini kullanarak, bu yaklaşım aykırı değerlerin varlığını hesaba katmaz ve bu nedenle sonuçlar aykırı değerlerin varlığından kötü şekilde etkilenebilir. Önerilen bir başka teknik, SEQimpute (Verboven vd. baskıda) adı verilen, verilerin kovaryansının determinantını en aza indirerek bir veri setindeki kayıp değerleri sırayla tamamlar. Kayıp değerler, verilerin ortalaması ve kovaryansı dikkate alınarak hesaplandığından, verilerde aykırı değerler mevcut olduğunda bu yaklaşım da zarar görecektir. Kayıp değerlerin hesaplanmasında verilerin ortalamasını ve kovaryansını kullanan SEQimput yöntemi, verilerde aykırı değerlerin bulunması durumunda sağlam değildir. Bununla birlikte, SEQimpute' deki ortalama ve saçılımın (scatter) hesaplanmasındaki bir değişiklik, aykırı değerlerin bu kötü etkisini ortadan kaldıracaktır. Branden ve Verboven (2009) tarafından bu yan etkilerin üstesinden gelmek için ROBimpute adı verilen sağlam bir atama tekniği önerilmiştir. ROBimpute yöntem, SEQimput dikkate alınarak ve bu yöntemdeki bazı önemli adımların sağlamlaştırılmasıyla elde edilir. Bu sağlam atama yöntemi, ortalama ve kovaryans hesaplamasını ayarlayarak SEQimpute yönteminden elde edilir (Branden ve Verboven 2009).

Öncelikle, SEQimpute (sequential imputation) yönteminin tanımı yapılacak olursa, SKNNimpute (Sequential K Nearest Neighbour)' ye benzer sıralı bir atama tekniğidir (Verboven vd. 2007). Bu yöntem, bir kovaryans kriterini en aza indirerek kayıp değerleri sırayla tahmin eder. Bunun için kayıp değerleri olmayan X_c adlı bir tam matris yapısı gereklidir. x^* en az kayıp değerlere sahip bir veri yapısı ile çözümleme başlar, x^* dahil tam veri kümesinin kovaryansının determinantını en aza indirerek tahmin edilir. Verboven vd.' nde (2007) gösterildiği gibi, bu, x^*' nin kayıp parçasının aşağıdaki gibi değiştirilmesiyle ilgilidir:

$$D(x^*) = (x^* - \bar{x}_c)' - (cov(X_c)^{-1}(x^* - \bar{x}_c)$$

x_m^* için determinantın minimize edilmesiyle $\frac{\partial D(x^*)}{\partial x_m^*} = 0$ minimizasyon probleminin çözümü;

$$x_m^* = (\bar{x}_c)_m - \left(C_{m,m}\right)^{-1} C_{m,o}(x_o^* - (\bar{x}_c)_o)$$

burada $\bar{x}_c = [(\bar{x}_c)_m' \ (\bar{x}_c)_o' \ ]'$, x^* 'deki kayıp ve gözlemlenen verilere göre iki parçayı gösterir, $X^* = [\ X_c' \ x^*]'$ ile $C = (cov(X_c))^{-1}$ gösterilir. Burada

$$C = \begin{bmatrix} C_{m,m} & C_{m,o} \\ C'_{m,o} & C_{o,o} \end{bmatrix}$$

C kare matrisinde $C_{m,m}$, C 'nin x^* 'de kayıp olan değişkenlere karşılık gelen kısmını gösterir ve $C_{m,o}$, C 'nin satırlarda x^* 'in kayıp değişkenleri olan kısmını ve sütunlarda x^* 'in gözlenen değişkenlerini gösterir.

x^*' nin kayıp değerleri tahmin edildikten sonra, veriler tam $X_c = [\ X_c' \ x^*]'$ alt kümesine eklenir ve kayıp değerleri tahmin etmek için en az kayıp değerlere sahip bir sonraki veri dahil edilir. Diğer yöntemlerden farklı olarak, bu yaklaşım bir parametrenin belirtilmesine ihtiyaç duymaz. Verboven vd. (2007) bu prosedürün, diğer yöntemleri geliştirdiği gösterilmiştir.

ROBimpute adı verilen sağlam bir atama tekniğiyle, konum ve dağılım (scatter) için sağlam bir tahmin edici ekleyerek daha sağlam bir atama yöntemi elde ederiz. Hesaplamalı olarak uygulanabilir bir algoritmaya sahip olmak için, Stahel (1981) ve Donoho' da (1982) tanımlandığı ve Hubert vd.' de (2005) aykırı verileri saptamak ve ortadan kaldırmak için kullanılan aykırılık ölçüsü önerilmiştir. Aykırılık sadece tamamen gözlemlenen veriler için ölçülebildiğinden, ilk olarak sadece X_c verilerinin tam setindeki verilerin aykırılığını tanımlayabiliriz. Tamamen bilinen veri x_i' nin aykırılığı, X_c' nin bir parçası

$$outl_i = \max_{v \in B} \left(\frac{\left| x_j'v - med(x_j'v) \right|}{mad(x_j'v)} \right)$$

olarak tanımlanır, burada B, $\mathbb{R}^s$' deki tüm sıfır olmayan vektörleri içerir, $med(x_j'v)$, $\{x_j'v,\ x_j \in X_c\}$' nin medyanıdır ve medyan mutlak sapma (median absolute deviation)

$$mad = med(\left| x_l'v - med(x_j'v) \right|; x_l \in X_c)$$

olarak tanımlanır (Branden ve Verboven 2009).

ROBimpute yöntemini başlatmak için, eksiksiz (complete) veri kümesi X_c' nin bir başlangıç sağlam kovaryans matrisi ve sağlam ortalaması gereklidir. Tam verilerin bir kısmının aykırı değer olduğu varsayıldığından, bu istatistikler başlangıçta X_c' deki en küçük aykırılık ile tam verilerin $(\alpha \times 100)\%$ kovaryansı ve ortalaması olarak tanımlanır. Aykırı değerlerin oranı olmak üzere, α parametresi, yöntemin dayanabileceği aykırı değerlerin miktarını kontrol eder. Örneğin %10' dan fazla bozulmuş verilerden şüphe varsa $\alpha \leq 0.9$ ve %10' dan az aykırı değerler varsayılıyorsa $\alpha \geq 0.9$‘ dır.

X_c' nin ortalaması ve kovaryansı için sağlam tahminler elde edildikten sonra, en az kayıp x^* değerine sahip veri, SEQimpute için tanımlanan yaklaşımla tamamlanır. Bu veri bir kez tamamlandığında, uzaklığı

$$outl^* = \max_{v \in B} \left(\frac{\left| (x^*)'v - med(x_j'v) \right|}{mad(x_j'v)} \right)$$

olarak hesaplanabilir.

Burada B kümesi medyan ve mad tam veriler kümesini güncelledikten sonra değişmez ve bu nedenle, uygulanamaz bir zaman komplex algoritmasından kaçınmak için her zaman ilk tam veriler kümesine dayanır. Daha sonra bu aykırılık değerine göre bu değerin aykırı olup olmadığına karar verilir. Değerin aykırılığının küçük olması aykırı değer demek değildir; yani $outl^* \leq (outl_i)_{100\times\alpha\times c:c}$, tamamlanmış verinin aykırılığının $(outl_i)_{i=1}^{c}$' in $(\alpha \times 100)$ üncü yüzdelik diliminden daha küçük olması gerektiği anlamına gelir. Formülde $outl_i$, $i.$ verinin aykırılığıdır ve c, prosedürün başlangıcındaki tam verilerin sayısını temsil eder. Aykırılık ölçüsü, belirlenmiş olan aykırılık değerini aşarsa analizlere dahil edilip edilmeyeceğine karar verilebilir. Veride bozulma olmadığı düşünülürse kovaryans ve ortalama, x^* veriyi dahil edilerek güncellenir. Bu prosedür daha sonra tüm tamamlanmamış (incomplete) veriler için tekrarlanır. Son derece yüksek bir aykırılığa sahip veriler, ROBimpute tarafından aykırı olarak işaretlenir (0 etiketi alırlar) ve veri analizinin sonraki adımlarında yok sayılabilir**.**

Veri setinin ortalamasını ve dağılımını (scatter) sağlamlaştırmak için diğer bir yöntem olarak örneğin Minimum Kovaryans Determinant (MCD) tahmin edicisi gibi düşünülebilir (Rousseeuw, 1984). Ancak, veriler üzerinde herhangi bir dağılımın belirtilmesini gerektirmediği, yöntemin kolayca uygulanabildiği ve hesaplama süresinin çok hızlı olduğu için ROBimpute yaklaşımı daha avantajlıdır. Benzer şekilde sağlam teknikler de, örneğin Hubert ve Engelen' de (2004) aykırı değer bulunan veriler için sağlam değildir, çünkü burada örneklerin çoğunluğunun aykırı değerlerden bağımsız olması gerektiği varsayılır. Verilerin her gözleminde aykırı değer bulunabileceğini için ROBimpute yaklaşımının avantajlı olduğu söylenebilir.

2.5.2.2.6 Oran imputasyon (ratio imputation)

Oran imputasyon, kayıp değerlere sahip bir değişkenin değerlerini tahmin etmek için kesişmeyen basit bir lineer regresyon modelini varsayar. Regresyon parametresinin ağırlıklı en küçük kareler tahmini, bu model altında en iyi doğrusal yansız tahmin edici olduğu söylenebilir.

Oran imputasyon yöntemi, Y' nin kayıplığı büyük ölçüde ilişkili bir yardımcı değişken X' e bağlıysa, çok doğru imputasyonlar sağlayabilir. Ancak bu çok kısıtlayıcı bir varsayımdır. Pratikte, kayıp değerlerin birkaç yardımcı değişkene bağlı olması daha olasıdır. Oran imputasyonu yalnızca bir yardımcı değişken kullanabildiğinden, birçok durumda tam olarak verimli değildir. Bunu aşmanın bir yolu, bazı yardımcı değişkenleri sınıflandırma değişkenleri olarak kullanmaktır, ancak bu, yardımcı değişkenlerin verimli kullanımı üzerindeki sınırlamaya hala tatmin edici bir çözüm değildir. Sınıflandırma değişkenlerinin sayısı arttıkça, imputasyon sınıflarının sayısı hızla büyük hale gelir ve daha sonra bazı imputasyon sınıfları oldukça doğru oran tahminleri elde etmek için yeterli örneğe sahip olmayabilir (Hu ve Salvucci 2001).

3. AYKIRI DEĞER ve AYKIRI DEĞER YÖNTEMLERİ İÇİN GENEL BİLGİLER

3.1 Aykırı Değerin Tarihi Kronolojisi

Aykırı değer tarihçesi araştırıldığında, ilk defa Bernoulli (1777) tarafından uyumsuz (discrepant) gözlemler ortaya atılmıştır. Bu yıllarda yapılan bilimsel çalışmalarda çoğunlukla uyumsuz yani aykırı değer olduğu düşünülen verilerin veri setinden atılması gerektiği fikri yaygındı. Ancak Bernoulli, birkaç gözlemin aynı ağırlık veya momentte olduğunu veya her hataya eşit derecede eğilimli olduğu varsayımını şüphe ile karşılamış, çok yaygın olan tutarsız gözlemleri atma uygulamasını kınamıştır, atılan gözlemin belki diğerlerine en iyi düzeltmeyi sağlayacak olan gözlem bile olabileceğini ve her gözlemin kalitesi ne olursa olsun kabul edilmesi gerektiğini vurgulamıştır. Pierce' nin (1852) çalışmasında olasılık hesabının temel ilkelerinden kurallara uygun olarak, çıkarılacak olan gözlemlerin reddedilmesi için çözmenin önerildiği ilke; şüpheli verilerin tutularak elde edilen hataların olasılığı, artık anormal gözlem olmadığında önerilen gözlemlerin hatalarının olasılığından daha az olduğunda, önerilen gözlemler reddedilerek gözlem kümesinden atılması şeklindedir.

Chauvenet (1863) belirlediği kriterlere uymayan değerlerin reddedilmesini önermiştir. Stone (1868) ve Wright (1884) da aykırı değerlerin reddedilmesi için test istatistikleri önermişlerdir. Ayrıca regresyon modellerinde aykırı değerlerin belirlenmesi ve reddedilmesi için çeşitli istatistiksel testler önerilmiştir. Anscombe (1960), Anscombe ve Tukey (1963), Tietjen, Moore ve Beckman (1973), Lund (1975), Rosner (1975), Ellenberg (1976), Cook ve Weisberg (1980, 1982), Cook ve Prescott (1981), Barnett ve Lewis (1984) bunların başlıcalarıdır (Çil 1990).

Bunlar dışında aykırı değerlerin belirlenmesi Grubbs (1969), Kale (1979), Hawkins (1980), Prescott (1980) ve Rider (1933) tarafından verilmiştir. Bu yöntemlerin bir tarihsel incelemesi Stigler (1973, 1975) ve Harter (1978) tarafından verilmiştir (Beckman ve Cook 1983).

Glaisher (1872-73) aykırı değerleri reddetmek yerine, yaklaşık en olası bir sonuç bulduktan sonra; hiçbir zaman hiç olmayan şekilde, bu değerlere diğer değerlerden daha küçük bir ağırlıklandırmayı önermiştir. Stone (1873), Glaisher' ın yöntemine alternatif bir ağırlıklandırma yöntemi önermiştir. Newcomb (1886) tarafından da başka bir ağırlıklandırma yöntemi önerilmiştir.

20. yüzyılda da aykırı değerlerin veri setinden atılması ve atılmaması çalışmaları karşımıza çıkmıştır. Fisher (1922) tutarlılık kriterinin üzerinde verdiği örnekte aykırı değerlerin reddedilmesinin savunulamayacağı üzerinde durulmuştur, Irwin (1925) örneklemeye dayanan bir çalışmayla, aykırı değerlerin çalışma gurubuna ait olmadığını düşünerek atılmasını tercih etmiştir ve Student (1927) ise bunu doğru bulup, okuyucularının reddetme gerekliliğinin ortaya çıktığı rutin analiz hatalarının tanımıyla ilgilenebileceğini önermiştir. Thompson (1935) aykırı değerlerin reddedilmesine yönelik kriterler için, tüm gözlemlerin belirli bir kısmını veya örneğin boyutuna göre değişen bir oranı reddetmek üzere tasarlanabileceğini düşünmüş ve Student türü kritere dayalı test istatistiği önermiştir. Pearson ve Chandra Sekar (1936) ise Thompson' un önerisinin yararlı ve pratik bir kriter olduğunu söylemiş ancak birden fazla aykırı değer varlığında etkisiz olabileceğini belirtmiştir, ayrıca aykırı değerlerin atılması sorununda başka zorlukların olduğunu göstermeye çalışmıştır. Grubbs (1950) aykırı değerlerin belirlenmesi için başka bir test istatistiği önermiştir. Rosner (1975) bir örnekte birden fazla aykırı değeri algılayabilen "çok sayıda aykırı değer" prosedürleri önerilmiştir ve güç karşılaştırmaları yardımıyla, "çok sayıda aykırı değer" tespitinde "bir aykırı değer" prosedüründen çok daha üstün olduklarını önermiştir.

Çok değişkenli veri setinde bir durumu ya da olayı düşündüğümüzde bunu veri matrisinde satırsal olarak düşünebiliriz. Ancak bu veri matrisinde tüm sıranın etkilenmediği sadece bir veya birkaç hücrenin aykırı olduğunu düşündüğümüzde, "çok sayıda aykırı değer" yapısını daha detaylı bir tanıma ihtiyacı olduğu açıkça görülmektedir. 1960' lı yıllarda, bu tür aykırı sıraları (satırları) tespit edebilen uygun yöntemleri geliştirmek için çok fazla araştırma yapılmıştır. Hücresel bozulma modeli ilk olarak bir veri madenciliği bağlamında Alqallaf vd. (2002) tarafından araştırılmış ve

daha sonra kapsamlı olarak Alqallaf vd. (2009) tarafından tanımlanmıştır. Bunun sonucunda çok değişkenli yapıda aykırı değer mevcut ise yapının hücresel aykırı değer ve satırsal aykırı değer tanımları çalışmalara konu olmaya başlamıştır.

21. yüzyılda verilerin bilgiye dönüştürülmesinin önemi üzerinde duran araştırmacılar, aykırı değerlerin gözlem kümesinden atılmasını doğru bulmayıp; bu değerlerin yerine bazı tahmin değerlerinin kullanılmasını (imputasyon yapılmasını) veya sağlam tahmin yöntemleri ile aykırı değerlerin parametre tahminine olan etkilerinin azaltılmasını önermektedirler.

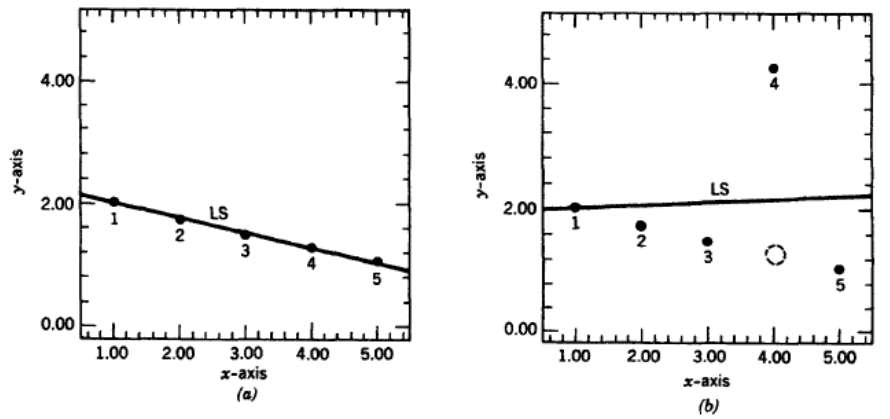
3.2 Aykırı Değer Tanımı

Aykırı değer; bir veri kümesinde, bu veri setinin geri kalanıyla tutarsız görünen bir gözlem olarak tanımlanmıştır (Barnett ve Lewis 1978). Araştırmalarda veri setlerindeki gözlemlerin aynı yapıya sahip olduğunu bir başka deyişle aynı dağılımdan geldiğini varsayılır. Veri setinde bulunan ve "diğerlerinden farklı" olarak adlandırılan aykırı gözlemleri belirlemek ve bu gözlemlere ilişkin gerekli çalışmayı yapmak ve verinin yapısını belirlemek araştırmacıların daha doğru sonuçlar elde etmesi için gereklidir.

Aykırı değerlerin matematiksel kesin bir tanımını vermek zor olduğundan Rousseeuw ve Leroy' un (1987) belirttiği gibi aykırı değer türlerini şu 2 başlık altında sınıflandırabiliriz:

3.2.1 Y- Yönlü aykırı değer

Şekil 3.1' de 5 nokta ile tahmin edilen EKK regresyon doğrusu ve noktalardan y_4' ün değiştirilmesiyle elde edilen yeni EKK regresyon doğrusu verilmiştir. y_4 noktası aykırılık göstermiştir ve bu nokta y yönlü aykırı değer ya da dikey aykırı değer (vertical) olarak adlandırılmıştır.

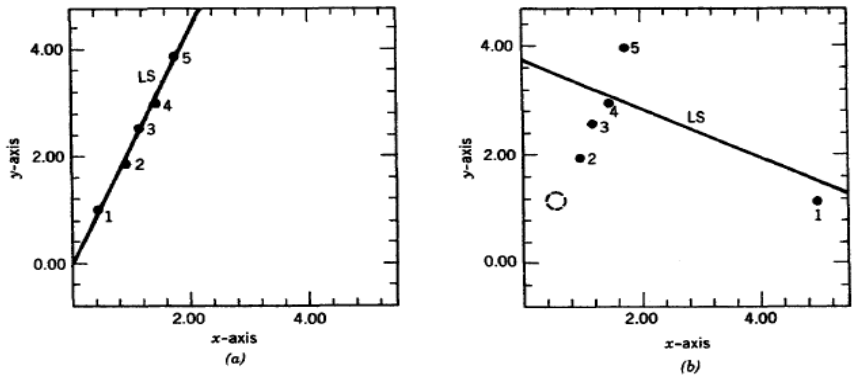


Şekil 3.1 Y-Yönlü aykırı değer (Rousseeuw ve Leroy 1987)

Bu nokta EKK doğrusu üzerinde büyük bir etkiye sahip olacak ve doğruyu tamamen değiştirebilecektir. Açıklayıcı değişkenlerinden birinde bir aykırı değer olması muhtemeldir, genellikle p-value 1' den büyüktür. Dikey aykırı değerler büyük pozitif ya da büyük negatif hata terimlerine sahiptirler. Bu tip aykırı değerler hata terimleri listesi ya da grafiğinin çizimi yardımı ile bulunabilirler.

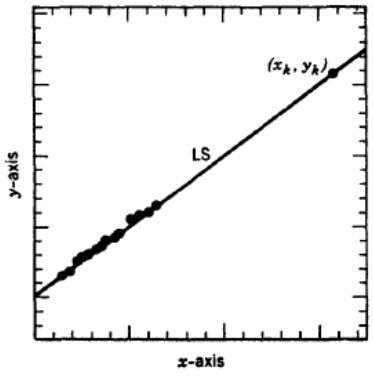
3.2.2 X- Yönlü aykırı değer (Kaldıraç Noktaları, leverage point)

Şekil 3.2‘ de 5 noktaya uydurulan EKK regresyon doğrusu ve x_1’ in değiştirilmesiyle noktalara uydurulan yeni EKK doğrusu verilmiştir. Veri noktasının merkezinden oldukça uzakta olan nokta regresyon doğrusunu büyük bir güçle çektiğinden kaldıraç nokta (leverage point) olarak adlandırılırlar.



Şekil 3.2 X-Yönlü aykırı değer (Rousseeuw ve Leroy 1987)

Bu örnekteki x_1 noktası, verilerin çoğunluğu tarafından belirlenen regresyon çizgisine yakın olduğunda, **Şekil 3.3**' teki gibi "iyi" bir kaldıraç noktası olarak kabul edilebilir.



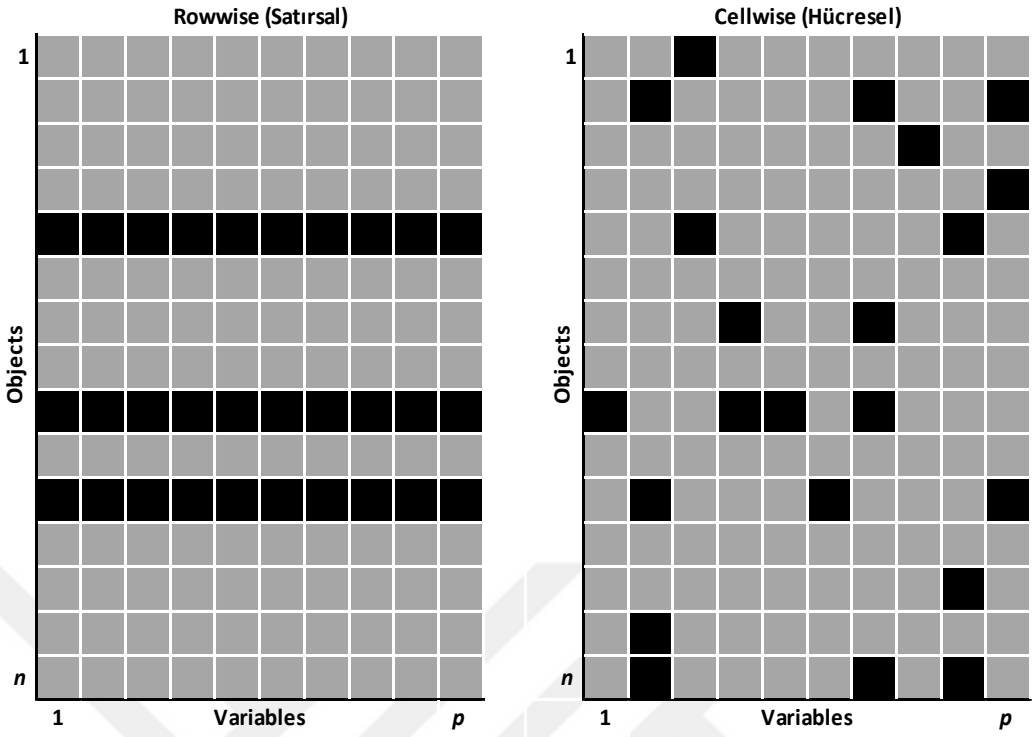
Şekil 3.3 İyi X-Yönlü aykırı değer (**Rousseeuw ve Leroy 1987**)

Veri setlerinde yer alan aykırı değerler, değişkenler tek tek ele alınarak kolaylıkla belirlenebilir. Gerek grafiksel yöntemler gerekse istatistiksel testler yardımıyla aykırı gözlemler hakkında bilgi sahibi olunabilir. Ancak birden fazla değişken ele alındığında çok değişkenli bir yapıda verilerin genel eğilimine uymayan gözlemlerin belirlenmesi oldukça zordur. Çok değişkenli bir veri seti ile çalışıldığında aykırı değerleri belirlemek için farklı yöntemler kullanmak gerekmektedir.

Son zamanlarda çok değişkenli veri setinde satırsal ve hücresel aykırı değer kavramı içeren iki çeşit yapı ile karşılaşmaktayız. Yani, çok değişkenli yapıda sadece veride aykırı değer varlığından söz edemeyiz. Bu aykırı değerin veri setindeki gözlemlenme durumuna göre iki farklı satırsal ve hücresel aykırı değer tanımı için en sade tanım aşağıdaki gibi verilebilir.

Rousseeuw ve Bossche' nin (2017) tanımına göre hücresel (cellwise) ve satırsal (casewise) aykırı değer tanımında veri seti için tanımlanan veri matrisi: n satır ve p sütun içeren dikdörtgen bir X matrisi biçimindedir. X satırları durumlara (olaylara) karşılık gelirken sütunlar değişkenlerdir. İstatistikler ve veri analizinde, aykırı değer sözcüğü tipik olarak **Şekil 3.4**' ün sol panelinde olduğu gibi veri matrisinin bir sırasına (satırına) karşılık gelir. 1960' lı yıllardan beri, bu tür aykırı sıralara (satırlara) daha az duyarlı olan ve aykırı değerleri bu uyumdan büyük artık (veya mesafe) ile tespit edebilen uygun yöntemleri geliştirmek için çok fazla araştırma yapılmıştır.

Son zamanlarda araştırmacılar, satırsal aykırı değer modelinin modern yüksek boyutlu veri setleri için artık yeterli olmadığını fark etmişlerdir. Çoğu zaman bir satırdaki çoğu veri hücresinin (girişler) düzenli olduğu ve sadece birkaç tanesinin anormal olduğu görülür. Hücresel bozulma modeli ilk olarak bir veri madenciliği çalışmalarında Alqallaf vd. (2002) tarafından araştırılmış ve daha sonra kapsamlı olarak Alqallaf vd. (2009) tarafından tanımlanmıştır. Rastgele konumlarda bozulmuş hücrelerin bir oranı ϵ verildiğinde, bozulmuş satırların beklenen oranı $\bar{\epsilon} = 1 - (1 - \epsilon)^p$ olmak üzere, aykırı değerlerin nasıl yayıldığını ve **Şekil 3.4**' ün sağ panelinde gösterildiği gibi ϵ yi arttırmak ve/veya p boyutunu arttırmak için çabucak % 50' yi aştığını belirttiler. Ancak, satırsal yöntemlerin, bazı temel değişmezlik (invariance) özellikleri varsa, bozulmuş satırların % 50' sinden fazlasını işleyemediği Lopuha ve Rousseeuw' de (1991) gösterilmiştir.



Şekil 3.4 Satırsal ve Hücresel aykırı değer modelleri (siyah aykırı değer anlamına gelir) (Rousseeuw ve Bossche 2017)

İki model oldukça farklıdır. Aykırı satır modeli, örneğin farklı bir topluluğun üyeleri oldukları için veri kümesine ait olmayan durumlarla ilgilidir. Temel birimleri durumlardır ve eğer bunları p-boyutlu uzayda noktalar olarak yorumlanırsa gerçekten bölünemezler. Ancak bir durumu bir matristeki satır olarak görünürse, o zaman hücrelere bölünebilir. Hücresel model, bu hücrelerin bazılarının, belki de kaba ölçüm hataları nedeniyle sahip olmaları gereken değerlerden saptığını varsayarken, aynı satırdaki geri kalan hücrelerin hala yararlı bilgiler içerdiğini varsayar. Mevcut araçlar yeterli olmadığından, aykırı (anormal) veri hücreleri probleminin oldukça zor olduğu kanıtlanmıştır. **Şekil 3.4**' ün sağ panelinde, başlangıçta hangi (varsa) hücrelerin siyah olduğunu (aykırı) bilmiyoruz, bu durum kayıp veri yapısıyla tam bir tezat oluşturmaktadır (Rousseeuw ve Bossche 2017).

3.3 Aykırı Değer Yapıları

Aykırı değerlerin belirlenmesinde kullanılacak çözümleme yöntemleri belirlemeden önce, verinin yapısı dikkate alınmalıdır. Çünkü verinin yapısı, değişken sayısı ve verideki şüpheli gözlemlerin sayısı hangi yöntemin kullanılacağını belirlemektedir. Veri setini etkileyen aykırı değeri, bulunduğu yapıya bağlı olarak belirleme yöntemlerinin bazı örneklerini vermeye çalışalım. Tek değişkenli verilerde aykırı değerlerin belirlenmesi için sıklıkla kullanılan iki yöntem vardır: Z Skoru ve Çeyreklikler Arası Dağılım Aralığı Yöntemi (IQR). Çok değişkenli verilerde ise; Cook Uzaklığı, Mahalanobis Uzaklığı, Sağlam Uzaklık Fonksiyonları ve Minumum Kovaryans Determinant Yöntemi gibi klasik belirleyiciler ile belirlenebilir. Regresyon hata terimlerine dayanarak gösterilen Standartlaştırılmış Hata Terimleri, Student Türü Hata Terimleri, Şapka Matrisin Köşegenleri, $DFBETAS$ Ölçüsü, $DFFITS$ Ölçüsü ve Kovaryans Oranı gibi aykırı değer belirleme yöntemleri de mevcuttur. Regresyon artıklarına dayanarak gösterilen yöntemler aykırı değerleri belirlemede oldukça etkindir.

3.3.1 Tek değişkenli veri setlerinde aykırı değerlerin belirlenmesi

Tek değişkenli verilerde aykırı değerlerin belirlenmesi için sıklıkla kullanılan iki yöntem vardır.

3.3.1.1 Z Skoru

En temel yöntem, Z skorlarından yararlanmaktır. Eğer üzerinde çalışılan veri tek tepeli simetrik olarak dağılıyorsa, bu durumda Z skorlarından yararlanılarak aykırı değerler belirlenebilir. Z skorları $Z = (x - \mu)/\sigma$ ile verilen formül yardımıyla hesaplanabilir. x gözlenen değeri, μ aritmetik ortalamayı, σ da standart sapmayı ifade etmektedir. Hesaplanan Z skoru $(-3,3)$ arasında olan gözlemler normal, bu sınırların dışında olan veriler ise aykırı değer olarak belirlenebilir.

3.3.1.2 Çeyreklikler arası dağılım aralığı yöntemi (IQR, Inter Quantile Range)

Bu yöntemde çeyreklikler arası dağılım aralıkları belirlenerek, bir güven sınırı oluşturulur ve bu sınırlar dışında olan veriler, aykırı değer olarak tanımlanır. Bu işlemlerin aşamalarını aşağıdaki gibi gösterebiliriz.

(a) Çeyreklik Değerlerinin belirlenmesi: Q_1 ve Q_3 olarak gösterilecek bu değerleri belirlemek için veriler küçükten büyüğe doğru sıraya dizilir. Daha sonra ilk ve üçüncü çeyrek değere karşılık gelen gözlemler 1. ve 3. çeyreklik (sırasıyla, Q_1 ve Q_3) olarak belirlenir.
(b) $Q_3 - Q_1$ değeri bulunur. Bu değer çeyreklikler arası dağılım aralığı değeridir ve IQR ile gösterilir.
(c) Q_1 değerinden 1.5* IQR değerini çıkartarak alt sınır, Q_3 değerine 1.5* IQR değerini ekleyerek üst sınır değeri belirlenir.
(d) $Q_1 -$ 1.5* IQR, Q_3 + 1.5* IQR güven aralığının dışında kalan gözlemler aykırı değer olarak belirlenir.

3.3.2 Çok değişkenli veri setlerinde aykırı değerlerin belirlenmesi

Çok değişkenli veri setlerinde, değişkenler arasında bağımlılık yapısı var ise, aykırı değerlerin belirlenmesinde regresyon çözümlemesi kullanılabilirken, bağımlılık yapısı yok ise farklı yöntemler kullanılarak aykırı değerler belirlenebilir.

Birden fazla değişkenimiz olduğunda; çeyreklikler arası dağılım aralığı gibi tek değişken için uygun olan yöntemleri kullanmak çok yanıltıcı olabilir. Bu tür yöntemler ile belirlenen uç değerler her zaman aykırı değer olarak nitelendirilmemelidir. Bu tür farklılıkları açıklayabilmek için regresyon artıkları incelenebilir. Regresyon artıklarına dayanarak gösterilen yöntemler aykırı değerleri belirlemede oldukça etkindir. Öncelikle uzaklık ölçüleri ile aykırı değer belirleme yöntemlerini inceleyelim.

3.3.2.1 Bazı aykırı değer belirleme yöntemleri

Literatürde karşılaşılan uzaklık ölçüleriyle tanımlanmış, en çok kullanılan aykırı değer belirleme yöntemlerinden bazıları aşağıdaki gibi hesaplanmaktadır.

3.3.2.1.1 Cook uzaklığı

Cook (1977), β' nın en küçük kareler tahmininin belirlenmesine her veri noktasının etkisini değerlendirmek için güven elipsoidlerine dayalı bir ölçü geliştirdi. Aykırı değerlerin saptanmasında, uygulama kolaylığından dolayı diğer yöntemlere göre pratik bir aykırı değer belirleme aracıdır.

Cook uzaklığı, $\beta \in \mathbb{R}^{p\times 1}$ tahmin edilecek parametrelerin vektörü, $\hat{\beta}$ parametre vektörünün tahmini, $\hat{\beta}_{(-i)}$ $i.$ gözlem veriden çıkartıldıktan sonra hesaplanan parametre vektörü, $X \in \mathbb{R}^{n\times p}$ bağımsız değişken matrisi, n gözlem sayısı, p parametre sayısı, ve $e \in \mathbb{R}^{n\times 1}$ hata vektörü, artıkların kovaryans matrisi $R = r_i$ ve $s^2 = \frac{R\prime R}{n-p}$ olmak üzere,

$$D_i \equiv \frac{\left[(\hat{\beta}_{(-i)} - \hat{\beta}\,)'\mathrm{X}'\mathrm{X}(\hat{\beta}_{(-i)} - \hat{\beta}\,)\right]}{p.s^2}$$

ya da eşitlik olarak verilirse;

$$e = y - \hat{y} = (I - H)y \text{ ve } s^2 = \frac{e\prime e}{n-p} \text{ iken } D_i = \frac{e_i^2}{p.s^2}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right]$$

olarak tanımlanmıştır (Cook 1977, 1979). D_i betimsel anlamlılık düzeyleri açısından $\hat{\beta}_{(-i)}$ ve $\hat{\beta}$ arasındaki uzaklığın bir ölçüsünü sağlar. Ölçek (scale) değişimlerinde D_i' nin invaryant olduğu kolayca görülür. Cook, D_i test istatistiğini karşılaştırmak için belirli bir α düzeyinde $F > (1-\alpha; p, n-p)$ tablo değerlerinin kullanılabileceğini

belirtmiştir. $D_i > F(1-\alpha)$ ise, şüphelenilen gözlem değerinin aykırı değer olduğuna karar verilmektedir.

Güncel çalışmalarda genellikle $4/n$ değeri ile karşılaştırılarak karar verilir. D_i nin $4/n$ den büyük değerlerinin aykırı değer olduğu söylenir (İşçi Güneri vd. 2021).

3.3.2.1.2 Mahalanobis uzaklığı

1936 yılında Mahalanobis tarafından önerilmiştir. Mahalanobis uzaklığı, bir gözlemin, bağımsız değişkenlerin oluşturduğu verilerin merkezinden uzaklığını ölçer.

Mahalanobis uzaklığı verinin varyans-kovaryans yapısını da göz önüne alan ve tek bir aykırı gözlemin testi için bir uzaklık formülü tanımlar. Tek bir aykırı gözlemin teşhisi yöntemlerinden farklı olarak Mahalanobis uzaklığı bir regresyon denkleminden hesaplanan artıklar üzerinden değil çok değişkenli veri üzerinden hesaplanır. X, p boyutlu gözlemlerden oluşan veri matrisi; $\bar{x}$, veriden hesaplanan ortalama vektörü ve S' de aynı veriden hesaplanan örneklem varyans-kovaryans matrisi olmak üzere Mahalanobis uzaklığı i. gözlem için

$$MD_i = \sqrt{(x_i - \bar{x}\ )'S^{-1}(x_i - \bar{x}\ )}$$

Şeklinde tanımlanır. Özellikle küçük örneklem boyutları söz konusu olduğunda, genellikle bir F dağılımının χ_p^2 dağılımından daha uygun olduğu öne sürülür. Ancak, uygulamada $\frac{p(n-1)}{(n-p)}F_{n,n-p}$ küçük örneklerde çok değişkenli aykırı değerleri test ederken uygun değildir (Penny 1996).

Mahalanobis uzaklığı, çok değişkenli normal dağılımlı bir örnekte, tek aykırı değerleri belirlemek için kullanılabilir. Ancak birden fazla aykırı değer veya aykırı değer kümesi içeren veri kümeleri, maskeleme (maskeleme: küçük bir aykırı değer kümesi $\bar{x}$ i çekecek ve S' yi kendi yönünde şişirerek MD_i için küçük değerler vermesi) ve bataklık

(swamping-bataklık**:** küçük bir aykırı değer kümesi $\bar{x}$ ' i çekecek ve S' yi kendi yönünde ve gözlemlerin çoğunluğu tarafından önerilen modele ait olan diğer bazı gözlemlerden uzağa şişirecek ve böylece bu gözlemler için büyük MD_i değerleri vermesi) sorunları nedeniyle, Mahalanobis uzaklıkları doğru sonuçlar vermeyebilir. Maskeleme ve bataklık sorunları, tanımı gereği aykırı değerlerden daha az etkilenen, sağlam şekil ve konum tahminleri kullanılarak çözülebilir (Hardin ve Rocke 2005).

3.3.2.1.3 Sağlam uzaklık fonksiyonları

Aykırı değer belirlemede kullanılan yöntemlerden birinin uzaklık fonksiyonlarından yararlanmak olduğu ve bir örnek olarak Mahalanobis uzaklığı ölçüsü tanımı verilmişti. Ancak birden fazla aykırı değer veya aykırı değer kümesi içeren veri kümelerinin, maskeleme ve bataklık sorunları nedeniyle, Mahalanobis uzaklıkları doğru sonuçlar vermeyebilir. Maskeleme ve bataklık sorunları, $\bar{x}$ ve S' nin sağlam olmamasından kaynaklanmaktadır. Bu tür sorunlardan kaçınmanın bir yolu, konum ve kovaryans matrisinin daha sağlam kestirimlerini kullanmaktır. Bu şekilde hesaplanan uzaklık fonksiyonlarına Sağlam Uzaklık Fonksiyonları (Robust Distances) adı verilir.

Sağlam uzaklıklarına ilişkin Hadi' nin (1992) önerdiği uzaklık fonksiyonu formülü

$$RD_i = D_i(C_R, S_R) = \sqrt{(x_i - C_R)' {S_R}^{-1} (x_i - C_R)}$$

şeklinde tanımlanır, burada C_R ve S_R değerleri, sağlam konum ve ölçek tahmin edicileridir. Hadi (1992) uygun sağlam parametre kestirimlerini aşağıdaki gibi göstermiştir.

Önce, sağlam bir kestirici olan medyan vektörleri C_M kullanılarak kovaryans S_M değerleri hesaplanır.

$$S_M = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - C_M)(x_i - C_M)'$$

Daha sonra gözlemleri $D_i(C_M, S_M)$‘ ye göre artan sırada yeniden düzenleyip, v_i ağırlık fonksiyonu aşağıdaki gibi tanımlanır

$$v_i = \begin{cases} 1, & eğer\ i \leq ((n+p+1)/2)'\ nin\ tamsayı\ kısmı, \\ 0, & dd, \end{cases}$$

Burada $C_R = C_v$ ve $S_R = S_v$' yi ayarlayarak RD_i denklemi elde edilir, burada $C_v = \frac{\sum_{i=1}^{n} v_i x_i}{\sum_{i=1}^{n} v_i}$ ve $S_v = \frac{\sum_{i=1}^{n} v_i (x_i - C_v)(x_i - C_v)'}{\sum_{i=1}^{n} v_i - 1}$ ile tanımlanır.

Verboven ve Hubert (2004) sağlam uzaklıklara ve Mahalanobis uzaklıklarına bağlı olarak, aykırı değerleri görselleştirmek ve klasik sonuçları karşılaştırmak için birkaç grafiksel saçılım grafiği çizdirmişlerdir. Çizilen dağılım grafiğinde her iki uzaklık değeri için de aykırı olan gözlemler aykırı değer olarak adlandırmışlardır. Sağlam uzaklıklara göre aykırı iken Mahalanobis uzaklıklara göre aykırı olmadığı görülen değerler gözlemlemiş olup, klasik bir yaklaşımla bu aykırı değerlerin fark edilemeyeceğini belirtmişlerdir.

3.3.2.1.4 Minimum kovaryans determinant yöntemi

Rousseeuw' un (1984) minimum kovaryans determinantı (MCD, Minimum Covariance Determinant) yöntemi, çok değişkenli konum ve saçılımın sağlam bir tahmincisidir. MCD konum ve şekil tahminleri, aykırı değerlere karşı oldukça dirençlidir ve genellikle tek başına ve diğer sağlam çok değişkenli yöntemler için bir yapı taşı olarak uygulanır. Verilen n veri noktasının MCD si, kovaryans matrisinin determinantını en aza indiren h $(h \leq n)$ boyutundaki örnekleme dayalı ortalama ve kovaryans matrisidir.

$MCD = (\bar{X}_J^*, S_J^*)$ olmak üzere

$$J = \{h\ noktaları\ kümesi: |S_J^*| \leq |S_K^*|, \forall K\ \ |K| = h\}$$

$$\bar{X}_J^* = \frac{1}{h}\sum_{i \in J} x_i$$

Ve

$$S_J^* = \frac{1}{h}\sum_{i \in J} (x_i - \bar{X}_J^*)(x_i - \bar{X}_J^*)'$$

Roussseeuw tarafından 1984 yılında önerilen yöntemde; n gözlemden oluşan ve p değişkene sahip olan $X_{n\times p}$ veri matrisinde, h değeri aykırı değer olmaması gereken minimum nokta sayısı $n/2 \leq h \leq n$ olmak üzere, klasik kovaryans matrisi en küçük olan h gözlemi belirlemek amaçlanır. $X_{n\times p}$ veri setine ait tüm h alt veri setine ait kovaryans matrisleri incelenerek, en düşük determinanta sahip alt veri seti seçilir. Konum parametresi, seçilen alt veri setinin ortalamasıdır. Dağılım parametresi de alt veri setinin kovaryans matrisi kullanılarak hesaplanır. MCD, $h = \left[\!\left[\frac{(n+p+1)}{2}\right]\!\right]'$ de mümkün olan en yüksek bozulmaya sahiptir (Rousseeuw ve Leroy 1987).

Aykırı değer tespiti için h yi olası en yüksek kırılımında kullanıp; h boyutundaki bir örneklemi yarı örneklem olarak adlandırılmıştır. MCD, en yakın yarı örneklemden hesaplanır ve bu nedenle aykırı değerlerin MCD konumu veya şekil tahmini üzerinde çarpıklığa sebep olmayacaktır. MCD konumu ($\bar{X}^*$) ve şekil (S^*) tahmini kullanılarak Mahalanobis kare uzaklıkların (MSD, Mahalanobis Squared Distance) büyük değerleri, sağlam uzaklık tahminleri olacak ve noktaları aykırı değer olarak doğru bir şekilde tanımlayacaktır.

Minimum determinantlı kovaryans yönteminde mümkün olan bütün h alt kümelerini hesaplamak, özellikle yüksek boyutlu veri kümelerinde zaman alıcı ve zor olabilir. Örneğin, h değerini örneklem sayısının yarısı olarak kabul edersek, n=20 iken yaklaşık 184756, n=100 iken yaklaşık 10^{29} determinant hesaplamak gerektirir. Bu hesaplama zorluğunu aşmak için çeşitli algoritmalar geliştirilmiştir. Bu algoritmalar yardımıyla doğru alt kümeye en yakın alt küme bulunmakta ve böylece gerçeğe en yakın sonuçlara

ulaşılmaktadır. MCD yi tahmin etmek için kullanılan algoritma bir anlamda tahmin edicidir (Hardin ve Rocke 2005).

3.3.2.2 Regresyon artıklarına bağlı aykırı değer belirleme yöntemleri

Regresyon artıklarına dayanarak gösterilen yöntemler aykırı değerleri belirlemede oldukça etkindir. Literatürde regresyondan hesaplanan tahmin değerleri ve artıklar üzerinden izlenecek süreç veya kural dizileri dolaysız yöntemler olarak adlandırılır. LMS gibi sağlam regresyon artıkları üzerinden veya MVE ve MCD gibi sağlam konum ve yayılım ölçüsü tahmincileri üzerinden hesaplanan uzaklıklardan faydalanılarak geliştirilen algoritmalar da dolaylı yöntemler olarak adlandırılırlar. Sağlam konum ve yayılım ölçüleri için geliştirilen algoritmalar genel olarak bir yaklaşımdır**.** Bu yüzden algoritmalarda bir miktar hata payı vardır.

3.3.2.2.1 Standartlaştırılmış ve Student türü artık terimleri

Farklı gözlemler için artıkları karşılaştırırken, artık varyanslarının farklı olabileceği gerçeğini hesaba katmak için r_i/s olarak tanımlanan artık terimlerinin standart sapmalarının tahminlerine bölümünü kullanabiliriz. y_i' de ölçüm hatalarının, bağımsız ve sıfır ortalama ve σ standart sapmayla normal dağıldığı bilindiğinde $s^{2'}$ nin $\sigma^{2'}$ nin yansız bir tahmincisi olduğunu hatırlayalım.

e, en küçük kareler regresyonunda artık vektörü, n gözlem sayısı, p parametre sayısı ve s^2 hata teriminin varyansının yansız bir tahmincisi olmak üzere

$$s^2 = \frac{1}{n-p}\sum_{i=1}^{n} e_i^2$$

olarak verilmişken standart artıklar

$$t_i = \frac{e_i}{s\sqrt{1-h_{ii}}} \qquad (3.1)$$

olarak tanımlanabilir. Standartlaştırılan hata terimlerinin 0 ortalama ve 1 varyans ile normal dağılıma uyması beklenir. Standartlaştırılan hata terimlerinden aykırı değerleri belirlemek için faydalanılabilir. Mutlak değeri 2' den büyük standartlaştırılmış hata terimine sahip gözlemler aykırı değerler olarak tanımlanmaktadırlar. Standartlaştırılmış artıkların, aykırı değerlerden etkilenebileceği ve dolayısıyla tespitten kaçabilecek olan bir standart sapma tahminine bağlı olmaları karşılaşılan bir problemdir. Bu problemin çözümü için $s(i)$, tüm regresyon i' inci gözlem olmadan tekrar çalıştırıldığında σ' nın tahmini iken hesaplanan

$$t(i) = \frac{e_i}{s(i)\sqrt{1-h_{ii}}} \tag{3.2}$$

değeri ise student artık olarak adlandırılır. Bazı araştırmacılara göre $t(i)$ değerleri jack-knifed artıklar olarak adlandırılmıştır. Bazı araştırmacılar da (3.1) ile gösterilen eşitliğe dahili (internally) student artıklar, (3.2) ile gösterilen eşitliğe ise harici (externally) student artıklar olarak adlandırmıştır (Rousseeuw ve Leroy 1987). Ayrıca (3.2) eşitliği ile ifade edilen $t(i)$ değerleri $n-p-1$ serbestlik derecesi ile t-dağılımına yakın olarak dağılır. $t(i)$ bağımlı iken, bu sayede i. gözlemin regresyon üzerindeki etkisini kolayca değerlendirebiliriz (Belsley vd. 1980). Mutlak değeri 2' den büyük studentlaştırılmış hata terimine sahip gözlemler aykırı değerler olarak tanımlanmaktadırlar.

3.3.2.2.2 Şapka matrisin (Hat Matrix) köşegenleri

Doğrusal regresyon modeli; n gözlem sayısı, p tahmin edilecek parametre sayısı olmak üzere $X \in \mathbb{R}^{n\times p}$ bağımsız değişkenler matrisi, $y \in \mathbb{R}^{n\times 1}$ bağımlı değişkenlerin vektörü, $\beta \in \mathbb{R}^{p\times 1}$ tahmin edilecek parametrelerin vektörü ve $e \in \mathbb{R}^{n\times 1}$ hata vektörü iken

$y = X\beta + e$

şeklinde tanımlanabilir. β' nın en küçük kareler tahmini $\hat{\beta}$

$\hat{\beta} = (X'X)^{-1}X'y$

iken tahmin vektörü

$\hat{y} = X\hat{\beta}$

olacaktır. $\hat{\beta}$ değeri yerine konursa,

$\hat{y} = X(X'X)^{-1}X'y$

olur.

$H = X(X'X)^{-1}X'$

denilirse H değeri yerine konursa

$\hat{y} = Hy$

şeklinde yazılabilir. $n \times n$ boyutlu şapka matrisi H görüldüğü gibi y vektörü ile çarpıldığında tahmin vektörü elde edilmiştir, diğer bir deyişle H nin y ile çarpımı $\hat{y}$ verdiğinden; H matrisi y vektörüne şapka takmıştır. Bundan dolayı bu matris şapka matrisi olarak bilinir.

Şapka matrisi için *Rank*(H) = *Iz*(H) = p olduğundan şapka matrisinin diagonal köşegenindeki elemanların (average value of the h_{ii}) ortalama değeri p/n' dir. Ayrıca şapka matrisi idempotent ($HH = H$), simetrik ($H' = H$) ve diagonal köşegen elemanları 0 ile 1 arasında ($0 < h_{ii} < 1$ tüm i için) yer alan bir matristir. Bu sınırlar, kaldıraç ölçüleri ve tahmin üzerindeki etkinin ölçüsünü yorumlamak için kullanışlıdır, ancak h_{ii}' nin ne zaman aykırı olduğunu tam olarak söylemez. Çoğu araştırmada $h_{ii} > 2p/n$ oranını aşan diagonal köşegen elemanına denk gelen gözlem değeri x-uzayında aykırı değer olarak nitelendirilebilir (Hoaglin ve Welsch 1978, Henderson ve Velleman 1981, Cook ve Weisberg 1982, Hocking 1983, Paul 1983, Stevens 1984). Bazı araştırmacılar da daha küçük örneklemler için $h_{ii} > 3p/n$ oranını aykırılık için kesme değeri olarak kullanır. Rousseeuw, ve Leroy‘ a göre, şapka matrisi söz konusu olduğunda, gerçek aykırı değerler, iyi kaldıraç noktalarının etkisiyle maskelenir. Regresyon analizinde aykırı değerleri keşfetmeye çalışırken h_{ii}' yi tek başına incelemek yeterli değildir, çünkü bunlar y_i' yi hesaba katmazlar. Şapka matrisi, veri seti x_i aykırı değeri ile yalnızca tek bir gözlem içerdiğinde kullanışlıdır (Rousseeuw ve Leroy 1987).

3.3.2.2.3 $DFBETA$ ve $DFBETAS$ ölçüleri

$DFBETA$ istatistiği $i.$ gözlemin parametre tahmini üzerindeki etkisini ölçmek için hesaplanan bir etki ölçüsüdür. Regresyon modellerinden tahmin edilen parametre tahminleri ile $i.$ satır silinirse tahmin edilen regresyon parametre tahmini arasındaki farktır. X açıklayıcı değişken matrisi, e artık vektörü, $h_i = x_i(X'X)^{-1}x_i'$ şapka matrisinin $i.$ köşegen elemanı ve x_i, X matrisinin $i.$ satırı olmak üzere, $b(i)$ ile silinen $i.$ satır ile tahmin edilen katsayıları ifade eden bu değişiklik,

$$DFBETA_i \equiv b - b(i) = \frac{(X'X)^{-1}x_i'.e_i}{1-h_i}$$

şeklinde tanımlanmıştır.

Benzer bir şekilde $i.$ gözlemin $j.$ parametre değerini ne kadar değiştirdiğinin ölçüsü olarak $DFBETAS$ kullanılabilir. $DFBETAS$, $DFBETA$' dan farklı olarak $i.$ gözlemin çıkartılmasıyla elde edilen standart hatayı da göz önüne alır. i' nci satırın silinmesinden kaynaklanan b' nin j' inci bileşeni olan b_j' deki değişikliğin büyük veya küçük olup olmadığı genellikle en faydalı şekilde b_j' nin varyansına, yani $\sigma^2(X'X)_{jj}^{-1}$' e göre değerlendirilir.

$C = (X'X)^{-1}X'$ olmak üzere $b_j - b_j(i) = \frac{c_{ji}e_i}{(1-h_i)}$

$$DFBETAS_{ij} \equiv \frac{b_j - b_j(i)}{s(i)\sqrt{(X'X)_{jj}^{-1}}} = \frac{c_{ji}}{\sqrt{\sum_{k=1}^{n} c_{jk}^2}} . \frac{e_i}{s(i)(1-h_i)}$$

olarak tanımlanmıştır. Örneklem sayısı arttıkça, gözlemlerden hesaplanan $DFBETA$ değerleri aynı oranda azalmaktadır. Örnek boyutu arttıkça bir gözlemi silmek daha az etkiye sahip olmalıdır. Öte yandan $DFBETAS_i$ değerleri, $\sqrt{n}$ le orantılı şekilde azalırlar. Büyük $\left|DFBETAS_{ij}\right|$ değeri j' inci katsayının, b_j' nin belirlenmesinde etkili olan

gözlemleri gösterir (Belsley vd. 1980). n gözlem sayısı olmak üzere $\left|DFBETAS_{ij}\right| > 2/\sqrt{n}$ veya $|DFBETA_i| > 2/\sqrt{n}$ olan gözlemler aykırı değer olarak düşünülebilir.

3.3.2.2.4 *DFFIT* ve *DFFITS* ölçüleri

$DFFIT$, $DFBETA$' ya çok benzer fakat $DFFIT$ istatistiği bir gözlemi ayrı tutarak parametre tahminlerinde meydana gelen değişiklik yerine, bir gözlemin dışlanması ile modelin fit ($\hat{y}$) değerlerinde meydana gelen değişikliği ölçer. İlk olarak tüm gözlemler kullanılarak model tahmin edilir ve daha sonra şüpheli gözlem dışlanarak tahmin tekrarlanır. Her gözlem için tekrarlanan bu süreç, şüpheli i. gözlemin modelde bulunduğu ve modelden çıkarıldığı durumda hesaplanan tahmin değerleri arasındaki farktır.

$$DFFIT_i \equiv \hat{y}_i - \hat{y}_i(i) = x_i[b - b(i)] = \frac{h_i . e_i}{1 - h_i}$$

olarak tanımlanır. $DFFIT$ ölçütü regresyon modelini oluşturmak için kullanılan belirli koordinat sistemine bağlı olmaması avantajına sahiptir (Belsley vd. 1980). Büyük $DFFIT$ değerlerine karşılık gelen değerler aykırı değer olarak düşünülebilir. Ölçeklendirme amacıyla, fit ($\hat{y}$) in standart sapması olan $\sigma\sqrt{h_i}$ ile bölünebilir, burada σ, $s(i)$ ile tahmin edilmiştir. Aslında, i. gözlemin çıkartılmasıyla elde edilen standart hatayı da göz önüne alan $DFFITS_i$, $\hat{y}_i - \hat{y}_i(i)$' nin i' nci bileşeninin standardizasyonundan kaynaklanır ve bir gözlem silindiğinde tahmin üzerindeki etkiyi ölçer.

$$DFFITS_i \equiv \frac{\hat{y}_i - \hat{y}_i(i)}{s(i)\sqrt{h_i}} = \left[\frac{h_i}{1 - h_i}\right]^{1/2} \frac{e_i}{s(i)\sqrt{1 - h_i}}$$

p parametre sayısı, n gözlem sayısı olmak üzere $|DFFIT_i| > 2\sqrt{p/n}$ veya $|DFFITS_i| > 2\sqrt{p/n}$ olan veriler aykırı değer olarak kabul edilebilir (Rousseeuw ve Leroy 1987).

3.3.2.2.5 $COVRATIO$ ve $FVARATIO$ ölçüleri

Regresyonda; katsayılar, y' nin tahmin edilen (fit) değerler ve artıklar ele alınır, bunların yanı sıra regresyonun bir diğer önemli çalışması, tahmin edilen katsayıların kovaryans matrisidir. $COVRATIO$ ölçütü $i.$ hücrenin veri setinden çıkarılmasıyla hesaplanacak olan kovaryans matrisinin, $i.$ hücrenin de dahil edilerek hesaplanacak olan kovaryans matrisine oranı şeklinde hesaplanır. Tüm verileri kullanan kovaryans matrisi tahmini $s^2[X'X]^{-1}$ ve $i.$ satır silindiğinde ortaya çıkan kovaryans matrisi tahmini $s^2(i).[X'(i)X(i)]^{-1}$ olmak üzere,

$$COVRATIO \equiv \frac{det|s^2(i).[X'(i)X(i)]^{-1}|}{det|s^2.[X'X]^{-1}|} = \frac{1}{\left[\frac{n-p-1}{n-p} + \frac{e_i^{*2}}{n-p}\right]^p (1-h_i)}$$

$COVRATIO$ ölçütü, kabaca söylemek gerekirse h_i büyük olduğunda büyük ve student artıklar büyük olduğunda küçük olacaktır. p parametre ve n örneklem sayısını göstermek üzere $COVRATIO$ oranı için $|COVRATIO - 1| > 3p/n$ olan veriler aykırı değer olarak kabul edilebilir (Belsley vd. 1980). Ayrıca $i.$ hücrenin silinmesine bağlı olarak $\hat{y}_i$' nin varyansındaki değişim de incelenebilir. Bunun için verilerden hesaplanan $FVARATIO$ ölçütü faydalı olacaktır.

$$FVARATIO \equiv \frac{s^2(i)}{s^2(1-h_i)}$$

ölçüm olarak, yukarıda $COVRATIO$ için açıklandığı gibi h_i ve student artıkların farklı biçimlerine göre aynı davranış kalıplarını sergileyecektir.

3.4 Aykırı Değer Çözümleme Yöntemleri

Bir veri setindeki aykırı değerler belirlendikten sonra yapılacak işlemler değişiklik gösterebilmektedir. Örneğin araştırmacılar bu verileri;

- Veri setinden çıkartabilir,
- Yerlerine yeni değerler atayabilirler veya etkilerini azaltacak istatistiksel yöntemler kullanabilir.

Aykırı değer tespit edilip belirlendikten sonra çözümleme aşamasında, son zamanlarda yapılan incelemeler sonucunda araştırmacılar tarafından, aykırı değerlerin veri setinden atılması yerine, imputasyon veya sağlam tahmin yöntemleri ile aykırı değerlerin parametre tahminine olan olumsuz etkilerinin en aza indirilmesi önerilmektedir.

3.4.1 Veri silme yöntemleri

Aykırı değerlerin veri setinden çıkartılması, söz konusu verinin doğru olduğu durumlarda veri kaybına yol açabilir. Bu nedenle hatalı bir veri olduğu kesinleşmeyen hiçbir veri, veri setinden çıkarılmamalıdır.

Zamana bağlı, az sayıda gözlem imkanı olan örneğin kanser ve tedavisi için araştırılan çok değişkenli yapıdaki veri ele alındığında; aykırı gözlem içeren satıra ait, gözlemlerin veri setinden çıkartılması gözlem sayısında ciddi bir azalmaya yol açabilir. Aykırı değerin veri setinden silinmesi durumunda n örneklem hacmi azalır; bu yanlılığa sebep olabilir, standart hatayı artırabilir bu da kitlenin parametresini tahmin etme ya da kitleyi temsil etme becerisini azaltmaktadır.

Aykırı değerlerin veri setinden atılması yerine**,** imputasyon veya sağlam tahmin yöntemleri ile aykırı değerlerin parametre tahminine olan olumsuz etkilerinden kurtulmaya çalışmalıdır.

3.4.2 İstatistiksel yöntemler

Veri seti üzerinde yapılacak analizler için istatistiksel yöntemlerde veya sağlam tahmin yöntemlerde, aykırı değer varlığında, bu aykırı değerlerin parametre tahminine olan

olumsuz etkilerini en aza indirilmesi amaçlanır. Literatürde karşılaşılan bu istatistiksel yöntemlerde karşılaşılan kavramsal bazı tanımları aşağıda verelim.

Alqallaf vd. (2009) tarafından bozulma (kontamine) modelinin tarihsel gelişimi ele alınmıştır. İstatistiksel yöntemlerin çoğu belirli bir model bağlamında oluşturulur ve bu nedenle bu model için iyi performans gösterecek şekilde tasarlanır (örneğin, en iyi olmak için). Klasik modeller verilerin "normal" gürültüden etkilendiğini varsayar: ölçüm hataları, madde-madde farklılıkları ve diğer "iyi davranılmış" rasgelelik kaynaklarından kaynaklanan küçük ölçekli dalgalanmalar, örneğin Gauss rasgele değişkenleri, Gamma rasgele değişkenleri, Poisson süreçleri vd. gibi. Bozulma modelleri ise verilerin anormal gürültüden de etkilenebileceğini varsayar: veri bozulması, eşit olmayan veri kalitesi, karışık popülasyonlar, kaba hatalar, vb. kaynaklanan büyük ölçekli dalgalanmalar. En iyi bilinen ve en önemli bozulma modeli Tukey-Huber modelidir (Tukey 1962, Huber 1964). Bu model, ortalama olarak, verilerin büyük bir kısmının $(1-\epsilon)$ klasik yalnızca normal hata modelinden üretildiğini varsayar. Bununla birlikte, kalan veriler anormal gürültüden etkilenebilir. Başka bir deyişle, Tukey-Huber modeli, tam olarak tanımlanmış bir baskın bileşen ve belirtilmemiş bir azınlık bileşeni ile bir karışım dağılımını varsayar. Karışım oranı ϵ serbest bir şekilde belirtilmiş bir rahatsızlık parametresidir (örneğin $0 \leq \epsilon < 0.25$). Sağlam bir istatistiksel analizin amacı, azınlık bileşeni tarafından üretilen olası anormal gürültüyü filtreleyerek karışımın baskın kısmı üzerinde çıkarım yapmaktır. Aykırı durumları (azınlık karışımı bileşeninden gelenleri) tanımlayarak ve etkilerini hafifleterek Tukey-Huber bozulma modeli, genel stratejide en sağlam istatistiksel prosedürlerin altında yatan derin bir etkiye sahipti. Tukey – Huber bozulma modeli

$$X = (1-B)Y + BZ$$

ilk olarak tek değişkenli konum ölçeği kurulumunda tanıtılmıştır. Gözlenemeyen Y, Z ve B değişkenleri bağımsızdır, $Y \sim F$ [$N(\mu, \sigma^2)$ gibi iyi bir konum- ölçek dağılımı], $Z \sim G$ (dağıtım oluşturan belirtilmemiş bir aykırı değer) ve $B \sim Binom(1, \epsilon)$ (rastgele bir bozulma göstergesi). Sonuç olarak, gözlenen değişken X, karışım dağılımına $(1-$

$\epsilon)F + \epsilon G$ sahiptir. Model daha sonra genişletilmiş, regresyon ve çok değişkenli konum dağılım modelleri gibi diğer konularda kullanılmıştır (Alqallaf vd. 2009).

Machkour vd. (2017) tarafından En Küçük Mutlak Çekme ve Seçim Operatörü (Least Absolute Shrinkage and Selection Operatör, $Lasso$) tanımlanmıştır. Burada veri analizindeki birçok problem, p açıklayıcı değişkenlerinin sayısının n örnek boyutundan daha büyük olduğu durumlar da dahil olmak üzere, büyük veri kümelerinin model yorumlaması açısından ele alınması için dağılım model tahmini gerektirir. Bu ayarlarda aynı anda doğru parametre tahmini ve tutarlı değişken seçimi gereklidir. Klasik bir yöntem en küçük mutlak çekme (küçültme) ve seçme operatörüdür,

$$\hat{\beta}_{lasso} = \underbrace{argmin}_{\beta} \|\boldsymbol{y} - \boldsymbol{X}\beta\|_2^2 + \lambda\|\beta\|_1$$

burada $y \in \mathbb{R}^{n\times 1}$ bağımlı değişkenlerin vektörü, n gözlem sayısı, p tahmin edilecek parametre sayısı olmak üzere $X \in \mathbb{R}^{n\times p}$ bağımsız değişken matrisi, $\beta \in \mathbb{R}^{p\times 1}$ tahmin edilecek parametrelerin vektörü ve λ negatif olmayan gerçek bir sayıdır. λ ayarlama (akort, uydurma, uyarlanmış) parametresinin seçimine dayanarak, bu yaklaşım katsayıları sıfıra doğru çeker, bu da sapma-varyans değişimine izin verir (Machkour vd. 2017, Tibshirani 1996).

Literatür taramasında çok değişkenli veri setinde hücresel ve satırsal aykırı değer durumunda yapılan çalışmalarda kullanılan; korelasyon yapılarını dikkate alan yöntemler, kovaryans ve kovaryans matrisinin tersinin tahminine ilişkin yöntemler, konum ve dağılım matrisinin tahminleri ve bunların sağlam tahminleri için genelleştirilmiş S-tahmin ediciler, M-tahmin ediciler, $Lasso$ tipi değişken seçim operatörleri vb. bir çok önerilen yöntemlerle karşılaşılmıştır.

Rocke (1996) çalışmasında; çok değişkenli konum ve ölçeğin sağlam tahmini problemi için, S-tahmin edicilerinde iyileşme sağlayan tahmin ediciler tanımlamıştır. S-tahmin edicilerinin ρ fonksiyonu açısından tanımlanması açıkça ifade edilmemesine rağmen, ρ

fonksiyonunun, herhangi bir boyut için $\rho = 0.5x^2$ olan çok değişkenli normal dağılıma benzer şekilde boyutla değişmemesinin amaçlandığı açıktır. Konum için tek uygun boyut, d Mahalanobis mesafesidir, böylelikle $\rho(d/c)$ dikkate alınır ve c istenen bozulma boyutuna göre değişir. Bunun yüksek boyuttaki tahmin edicinin, az da olsa ana nokta kütlesinden çok uzakta olan aykırı değer noktaları olarak reddetme yeteneğini kaybettiğini göstermiştir. Ayrıca bunun gerekli olmadığını ve yüksek bozulma ve daha iyi aykırı değer reddetme özelliklerine sahip iki yeni S-tahmincisi örneği sağlamıştır. Bu fikrin uygulanmasının, standart iki kere ağırlıklandırılmış (biweight) S-tahmin edicisinden daha sağlam davranışa sahip bir M-tahmin edicisinin oluşturulmasına izin verdiği bir örnek verilmiştir. Genellikle tanımlandığı gibi %50 bozulma S- (ve M-) tahminlerinin, esnek olmayan bir ρ fonksiyonları ailesinin kullanılması nedeniyle yüksek boyuttaki bariz aykırı değerlere sıfır ağırlık uygulamadığı gösterilmiştir.

Danilov vd. (2012) araştırmalarında, aykırı değer ve kayıp veri sorununu eş zamanlı ele almıştır. Çok değişkenli konum ve dağılımın S-tahmincileri gibi popüler sağlam yöntemler, aykırı değerlere karşı koruma sağlarken, kayıp verilerle baş edemez. Kayıp veri ve aykırı değerlerle eş zamanlı olarak başa çıkmak için çok değişkenli konum ve dağılımın S-tahmin edicilerinin tanımını genelleştirmişlerdir. Veriler tamamen rastgele kayıp veri olduğunda, önerilen tahmin edicilerin eliptik modeller altında güçlü bir şekilde tutarlı olduğunu göstermişlerdir. Önerilen tahmin edicileri hesaplamak için Beklenti-Maksimizasyon algoritmasına benzer bir algoritma türetmişlerdir.

Agostinelli vd. (2015) tarafından yapılan araştırmada, aykırı değer türlerinin üstesinden gelmek için, iki adım içeren yeni bir yöntem önerilmiştir. $THCM$ (Tukey–Huber Contamination Model) sağlamlığına ulaştığı kanıtlanmış bir özellik olan benzeşik eşdeğerlik, aykırı değerlerin yayılımı nedeniyle ICM (Independent Contamination Model) altında bir engel haline gelir. ICM nin pratik ve teorik önemini savunup, sadece $THCM$ modeline güvenmenin tehlikelerine ve sakıncalarına işaret etmişlerdir. ICM ve $THCM$ altında sağlamlık elde etmek için iki aşamalı bir prosedür tanımlanmıştır. Prosedürlerinin ilk adımı, aykırı değerlerin yayılımının etkisini azaltmayı ve ICM nin yarattığı boyutsallık probleminin üstesinden gelmeyi amaçlamaktadır. İkinci adım

$THCM$ altında sağlamlık kazanmayı amaçlamaktadır. Prosedürleri benzeşik eşdeğerli değildir, ancak yine de hem ICM hem de $THCM$ aykırı değerlerine karşı oldukça yüksek direnç sağlar. Prosedürleri, en iyi performans gösteren güçlü benzeşik eşdeğer tahmin edici çerçevesinde karşılaştırıldığında $THCM$ altında bir miktar sağlamlık kaybı göstermektedir. Tahmin edicilerinin etki fonksiyonun, eksiksiz veri için S-tahmincisinin etki fonksiyonuyla aynı olduğunu düşünmüşlerdir. Benzer şekilde türetilmiş sağlam regresyon tahmin edicisinin, en küçük kareler tahmincisi ile aynı etki fonksiyonuna sahip olduklarını göstermişlerdir. Ayrıca, genel olarak etki fonksiyonunun çok bilgilendirici bir sağlamlık ölçütü olmadığına inandıklarını dile getirmişlerdir. Sınırlı bir etki fonksiyonu $THCM$ ve ICM altında sağlamlık için gerekli veya yeterli bir koşul olmadığını belirtmişlerdir. Bu konularda daha fazla araştırmaya ihtiyaç olduğunu söylemişlerdir.

Rousseeuw ve Bossche (2017) tarafından yapılan araştırmada, değişkenler arasındaki korelasyonları dikkate alan çok değişkenli bir örnekte aykırı veri hücrelerini saptamak için temiz satır sayısı üzerinde bir kısıtlama olmayan ve yüksek boyutlarla başa çıkabilen bir yöntem önermişlerdir. Önerilen yöntemin avantajı, kayıp verileri impute ederken, bununla birlikte aykırı hücrelerin tahmin edilen değerlerini sağlamasıdır. Önerilen yöntem, belirli bir satırın neden dışarıda olduğunu teşhis etmeye yardımcı olabilir. Aynı zamanda çok değişkenli konum ve saçılma matrislerini tahmin etmek için bir ilk adım olarak hizmet eder. Bu yöntemle, matrisin satırlarını temel birimler olarak kabul eden mevcut aykırı değer bulma yöntemlerinin aksine, veri matrisinin tek tek hücreleri (girişleri) düzeyinde aykırı değeri tespit etmektedir. Ana yapısı her hücrenin tahminidir. Simülasyonlarda, Detect Deviating Cells (DDC) yöntemlerinin, sütunlar arasında çok az korelasyon olduğunda sütunsal olarak aykırı değer tespiti gerçekleştirdiğini, daha yüksek korelasyonlar olduğunda ise çok daha iyi olduğu görülmüştür. Ayrıca, Agostinelli vd. (2015) yönteminin ilk adımı olarak kullanılması, sütunsal bir başlangıçtan daha iyi sonuç verdiği düşünülmüştür.

Croux ve Öllerer (2015) yorumlarında, hem hücresel hem de satırsal bozulmaya dayanıklı, çok değişkenli konum ve dağılım için tahmin edici önermelerinde genel fikir:

(i) büyük tek değişkenli aykırı değerleri kontrol edip ve bu hücresel aykırı değerleri NA ile değiştirip burada (ii) Danilov vd. (2012) deki kayıp veriler için S-tahmin edicisi uygulanması yönündedir. Adım (i)' de saptanan hücresel aykırı değerler yoksa, önerilen tahmin edici normal S tahmin edicisine eşittir ve afin (benzeşik) eşdeğerlik özelliğini paylaşır. Araştırmada, tahmin edicinin ana gücünün satırsal (veya çok değişkenli) aykırı değerlerinin Mahalanobis mesafesinin sağlam bir versiyonu kullanılarak tespit edildiği ikinci adımdan geldiği kabul edilmiştir. Her gözlem için, bu mesafe kayıp olmayan bileşenlerin sayısıyla verilen boyutta hesaplanmaktadır. Bu yorumlamada (a) $TSGS$ nin (Two-Step Generalized S-estimator) kovaryans matrisinin ters tahmin edicisi olarak performansı bir simülasyon çalışması ile araştırılmış, (b) $TSGS$ nin düzenli bir versiyonunu tartışıp, ve (c) Öllerer ve Croux (2014) tarafından yakın zamanda GGQ (Glasso Gauss Q_n, Gaussian rank correlation, Q_n-estimator as a scale) tahmin edicisi olarak önerilen bir tahmin ediciyle bir karşılaştırma yapılmıştır. Önerilen iki aşamalı genelleştirilmiş S-tahmin edicisinin, hem hücresel hem de satırsal bozulma altında var olan tam ve sağlam bir tahmin edici olduğunu söylemişlerdir.

Tarr vd. (2016), kovaryans matrisleri ile ters kovaryans matrisler arasında bir bağlantı olsa da, biri için iyi (sağlam) bir tahmin edicinin diğeri için iyi bir tahmin edici olamayacağına dikkat çekmişlerdir. Bununla birlikte, klasik kovaryans matrisinin aykırı değerlere karşı sağlam olmadığı ve karakter yapısında yüksek boyutlarda bozulma yaşandığı hatırlatılmıştır (Johnstone 2001). Çalışmalarında sağlam, ters kovaryans matrislerin tahminini kolaylaştırmak için çeşitli düzenleme teknikleri için başlangıç noktası olarak sağlam çift kovaryans matrisleri kullanılmıştır. Sağlam çift kovaryans tahminini NPD (nearest positive definite, belirli bir simetrik matrise en yakın pozitif tanımlanmış matris) yöntemiyle ve $CLIME$ (constrained ℓ_1-minimization for inverse matrix estimation), $QUIC$ (Quadratic Inverse Covariance) veya $GLASSO$ (Graphical $Lasso$) gibi düzenleyici tekniklerle birleştirmenin, hücresel bozulmaya dayanıklı ters kovaryans matrisler verdiğini göstermişlerdir. Giriş kovaryans matrisinin pozitif yarı-belirli olmasını sağlamak için simetrik bir çift kovaryans matrisini en yakın kovaryans matrisine dönüştüren bir yöntem kullanılmıştır. Sonuç olarak orta seviyedeki hücresel bozulmalara dayanıklı kovaryans matrisinin tersi elde edilmiştir. Bu prosedür alt

örneklemeye dayalı olmadığından, değişkenlerin sayısı arttıkça iyi ölçeklendiği önerilmiştir.

Leung (2016) tarafından yapılan çalışmada, hücresel ve satırsal aykırı değerleriyle başa çıkmak için iki adımlı bir prosedür önermiştir. İlk olarak, hücresel aykırı değerleri tanımlayıp ve kayıp verilerle değiştirmek için bir filtre kullanılmış, daha sonra tamamlanmamış verilere, düşük ağırlıklı satırsal aykırı değerler (down-weight casewise outliers) için sağlam bir tahmin edici uygulanmıştır. Filtrenin uygun şekilde seçilmesi koşuluyla, iki aşamalı prosedürün merkezi model altında tutarlı olduğunu göstermiştir. Daha sonra konum ve dağılım matrisini tahmin etmek için önerilen iki aşamalı prosedürle, ikinci aşamada elde edilen tahmini, sağlam çok değişkenli konum ve dağılım matrisinden güçlü regresyon katsayılarını hesaplayan üçüncü bir adım eklenerek, sürekli ortak değişkenler (covariates) durumunda regresyonda uygulanmıştır. Üç aşamalı tahmin edicinin, sürekli değişkenler için merkezi modelde tutarlı ve asimptotik olarak normal olduğu gösterilmiştir.

Leung vd. (2016) çalışmalarında yüksek bozulma noktası afin (benzeşik) eşdeğer sağlam tahmin edicilerin bağımsız bozulma modelinde (ICM, independent contamination model) ne etkili ne de sağlam olduğunu belirlemişlerdir. Klasik sağlam tahmin ediciler ICM altında verimsiz olup, tek bir bileşenin bozulmasıyla tüm bir sıraya düşük ağırlık verebilme ihtimalinin, verilerde yer alan bazı yararlı bilgileri kaybedebileceğini söylemişlerdir. Ayrıca, klasik yüksek bozulma noktası, afin eşdeğeri sağlam tahmin ediciler ICM altında çökebilir. Hücresel aykırı değerlerin küçük bir kısmı çoğalabilir ve bu da durumların büyük bir kısmını etkiler. Klasik sağlam tahmin edicilerin bu eksikliklerinin üstesinden gelmek için, hücresel ve satırsal aykırı değerleriyle başa çıkabilen üç adımlı bir regresyon tahmin edicisi önermişlerdir. Tahmin edicinin ilk adımı, ICM nin ortaya çıkardığı aykırı yayılımın etkisini azaltmayı amaçlamaktadır. İkinci adım $THCM$ (Tukey–Huber Contamination Model) altında sağlamlık kazanmayı amaçlamaktadır. Son olarak, üçüncü adımdan elde edilen sağlam regresyon tahmininin verimli (verilerin temiz kısmını kullanma yeteneği) ve ICM ve $THCM$ altında sağlam olduğu gösterilmiştir.

Leung vd. (2017) tarafından çok değişkenli konum ve yayılım matrisinin hücresel ve satırsal aykırı değerlerin varlığında sağlam bir tahmin gerçekleştirmesi için iki aşamalı bir yaklaşım önerilmiştir. İlk adımda hücresel aykırı değerlerin çıkarılması için tek değişkenli bir filtre uygulanmıştır. İkinci adımda satırsal aykırı değerlerin azaltılması için genelleştirilmiş bir S-tahmin edicisi kullanılmıştır. Bu öneri için ilk adımda tek değişkenli filtre ile birlikte kullanılacak tutarlı bir iki değişkenli filtrenin tanıtılmasıyla, ikinci adımda genelleştirilmiş S-tahmin edicisi için başlangıç noktaları oluşturmak üzere yeni bir hızlı alt örnekleme prosedürünün önerisi yoluyla, üçüncü adımda genelleştirilmiş S-tahmin edicisi için satırsal aykırı değerlerin yüksek boyutta daha iyi ele alınması için monoton olmayan bir ağırlık fonksiyonu kullanılarak geliştirilebileceği düşünülmüştür. Orijinal iki aşamalı prosedürün aksine, değiştirilmiş iki aşamalı yaklaşımın yüksek boyutta iyi performans gösterdiğini ve ölçeklendiğini göstermişlerdir. Önerilen iki adımlı prosedür sınırlamalara sahiptir. Sayısal deneyleri, önerilen iki adımlı prosedürün ilk adımda filtrelenen veri miktarına bağlı olarak $n \geq 5p$ olduğunda iyi çalıştığını, ancak $2p < n < 5p$ olduğunda iyi çalışmadığını göstermektedir.

Maronnaa ve Yohai (2017) tarafından yapılan araştırmada, çok değişkenli konum ve dağılımın en iyi tahmin edicilerini seçmek için uğraşmışlardır. En sık kullanılan tahmin edicilerin bu bakımdan tatmin edici olmadığını düşünmüşlerdir. Minimum Hacim Elipsoidi (MVE, Minimum Volume Ellipsoid) ve Minimum Kovaryans Determinantı (MCD, Minimum Covariance Determinant) tahmin edicilerinin çok düşük bir verime sahip oldukları bilinmektedir. Bisquare gibi monoton ağırlık fonksiyonuna sahip S-tahmin edicileri, p boyutu küçük olduğunda düşük bir verime sahiptir ve verimleri artan p ile bire eğilimlidir. Ne yazık ki, bu avantaj, büyük p için sağlamlıkta ciddi bir kayıptan daha ağır basmaktadır. Orta ile büyük p performansı şimdiye kadar incelenmemiş, kontrol edilebilir verimlilikleri olan dört (monoton olmayan ağırlık fonksiyonuna sahip S tahmin edicileri, MM tahmin edicileri, τ tahmin edicileri ve Stahel-Donoho tahmin edicisi) tahmin edici ailesi incelenmiştir. Alt örnekleme yoluyla hesaplanan MVE ve maksimum ve minimum basıklık gösterimine dayanarak daha önce aykırı değeri tespit için önerilen yarı-deterministik bir prosedür kullanmışlardır. Bir

simülasyon çalışmasıyla, monoton olmayan ağırlık fonksiyonuna sahip bir S-tahmin edicisinin $p \geq 15$ için eşzamanlı olarak yüksek verimlilik ve yüksek sağlamlığa ulaşabildiğini gösterirlerken, $p < 15$ için belirli bir ağırlık fonksiyonuna sahip bir MM tahmincisi önerilebileceğini söylemişlerdir.

Rousseeuw ve Bossche (2015) tarafından yapılan araştırmada yazarlar, hücresel ve satırsal aykırı değerlerin kombinasyonunu ele almak için özel olarak tasarlanmış bir tahmin edici önermişlerdir. Önerilen $2SGS$ yöntemi, aykırı değerler koordinatsal işaretlenip (konum ve ölçeğin sağlam tek değişkenli tahminlerine göre) ve daha sonra NA olarak ayarlandığı bir başlangıç adımıyla başlar. Ancak $2SGS$ nin asıl işi, Danilov vd. (2012) genelleştirilmiş S-tahmincisi (GSE), kayıp verilerin varlığında çok değişkenli konum ve dağılım hakkında sağlam bir tahmin sağlayan ikinci adımdır. $2SGS$ nin şu anda hücresel ve satırsal aykırı değerlerin kombinasyonu ile başa çıkmak için mevcut en iyi yöntem olduğunu düşünmüşler. Çok sınırlayıcı görünen ilk husus, ilk adımda yalnızca söz konusu tek değişken için değerlerine dayanarak hücresel aykırı değerleri işaretler, bu nedenle bu adımda hiçbir korelasyon yapısı dikkate alınmaz. Öte yandan, herhangi bir koordinatı dışarıda olmadan aykırı değer olabileceği iyi bilinmektedir. Bu nedenle, ilk adımın korelasyon yapısı hakkında bazı bilgiler kullanabilmesi daha iyi olduğu belirtilmiştir.

Machkour vd. (2017a) tarafından yapılan araştırmada, model seçimi için hücre yönelimli aykırı değerleri dikkate alan sağlam ağırlıklı ve uyarlanabilir $Lasso$ tipi düzenleyici bir terim önerilmiş ve analiz edilmiştir. Önerilen tür düzenleyici terimi, önerilen MM-Sağlam Ağırlıklı Adaptif Lasso yu ($MM - RWAL$) veren MM-tahmin edicisi amaç fonksiyonuna entegre edilmiştir. Ağırlıkları hesaplamak için bir algoritma önerilmiş ve analiz edilmiştir. Monte Carlo deneyleri kullanılarak mevcut sağlam $Lasso$ tahmin edicileri ile bir performans karşılaştırması sağlanmıştır. $MM - RWAL$, sayısal deneylerde mevcut sağlam tahmincileri geride bıraktığı ve gürültülü gözlemler ve hücresel ve satırsal aykırı değerleri içeren kesin olarak tahmin edilen koşulsuz ve dağılım modelinin verildiği, gerçek bir tahmin verisi uygulamasındaki yararlılığı kanıtlanmıştır. Önerilen $RWAL$, MM tahmincisi ile sınırlı olmayıp, $RWAL$ bunları

hücresel bozulma çerçevesine genişletmek için diğer satırsal sağlam tahmin edici fonksiyonlarına kolayca entegre edilebilir olduğunu önermişlerdir.

Machkour vd. (2017b) tarafından yapılan araştırmada, MM veya $MM\text{-}Lasso$ tahmincisi gibi yüksek bozulma noktası tahmin edicileri, X' teki satırlara karşılık gelen belirli bir veri noktalarının belirli bir bölümünün bozulma olduğunu varsayan Tukey-Huber bozulma modelinde ($THCM$) iyi performans gösterdiği hatırlatılmıştır. Ne yazık ki birçok veri kümesi, yüksek oranda ve bağımsız olarak bozulmuş tahmin ediciler içerir. Bu tip bozulmalar bağımsız bozulma modeli (ICM) ile temsil edilebilir. $THCM$ nin aksine, her bir tahmin edicideki girişlerin belirli bir bölümünde bozulma olduğunu varsayar. Böylece, ICM de veri noktalarının çoğu bozulmuş olabilir ve bu da klasik sağlam tahmin edicilerin bozulmasına neden olabilir. Bu nedenle, yeni bir sağlam ve seyrek tahminci olan Aykırı Düzeltilmiş Veri (Ayarlanabilir) $Lasso$ (Outlier Corrected Data (Adaptive) $Lasso$, $OCD-(A)\,Lasso$) önerilmiştir. Bu tahmin edici, klasik $THCM$ yi takip eden aykırı değerlerin yanı sıra X regresyon matrisinin tek tek hücrelerini bozan ICM aykırı değerlerle de başa çıkabilir. Bozulan hücreler tespit edilir ve enterpole edilen değerlerle değiştirilir. Bunun için, Stahel Donoho Outlyingness (SDO) ve Predictor Outlyingness (PO) özelliklerini birleştiren bir aykırılık ölçüsünü sunulmuştur. Bir p-boyutlu birim hiper küreden örneklemeye dayalı ilişkili ağırlıkları hesaplamak için bir algoritma önerilmiştir. Önerilen $OCD-(A)\,Lasso$ yönteminin, klasik $OLS\,Lasso$, $MM\,Lasso$ ve uyarlanabilir $A-MM\,Lasso$ yönteminden daha iyi performans gösterdiğini gösteren sayısal deneyler sunulmuştur.

Toka vd. (2021) tarafından yapılan araştırmada, aykırı değerlerin varlığında sağlam regresyon tahmin yöntemleri, aynı anda sağlam tahmin ve değişken seçimine ulaşmak için değişken seçim yöntemleriyle birleştirilmiştir. Sağlam değişken seçim yöntemleri, verideki hücresel aykırı değerlerle başa çıkamayacağı için, hücresel ve satırsal aykırı değerleri ile birlikte mevcut olduğunda bazı ekstra özen gösterilmesi gerektiği vurgulanmıştır. Çalışmalarında, verilerde hem hücresel hem de satırsal aykırı değerlerle başa çıkabilmek için sağlam bir tahmin ve değişken seçim yöntemi önermişlerdir. Önerilen yöntemin üç adımı vardır. İlk adımda, hücresel aykırı değerler tanımlanır,

silinir ve her açıklayıcı değişken NA işareti ile belirlenir. İkinci adımda, NA işaretli hücrelere sağlam veri atama yöntemi kullanılarak yerine veri atanır. Son adımda, regresyon parametrelerini tahmin etmek ve önemli açıklayıcı değişkenleri seçmek için, güçlü regresyon tahmin yöntemleri, değişken seçim yöntemi $Lasso$ (Least Absolute Shrinkage and Selection Operator) ile birleştirilmiştir.

4. HÜCRESEL AYKIRI DEĞER ve KAYIP VERİ DURUMUNDA SAĞLAM VERİ ANALİZİ

4.1 Sağlam Yaklaşım

İstatistiksel sonuç çıkarımlar kısmen gözlemlere dayanmaktadır. En basit durumlarda bile, rastgelelik ve olayların bağımsızlığı, dağılım modelleri, bilinmeyen parametreler için önceki dağılımlar vb. hakkında birçok varsayım vardır. Bu varsayımlar veri analizi veya istatistiksel modelleme problemi hakkında bilinenleri veya tahmin edilenleri resmileştirmeyi amaçlar. Kopyalama veya tuş vuruşu hataları gibi büyük hatalar, genellikle veri yığınından çok uzakta olan ve birçok klasik istatistiksel prosedür için tehlikeli olan aykırı değer olarak ortaya çıkarlar. Sağlam teknikler, bu varsayımların sağlanamaması durumunda güvenli sonuçlar bulmak ve aykırı değerlerin etkisini azaltmak amacıyla kullanılarak, klasik yönteme bir alternatif olarak ortaya çıkmıştır. En yaygın istatistiksel prosedürlerden bazılarının, varsayımlardan küçük sapmalara karşı aşırı duyarlı olduğunun giderek daha fazla farkına varılmış ve çok sayıda alternatif sağlam prosedürler önerilmiştir.

"Robust" sözcüğünün kelime anlamına bakıldığında bu kelimenin Türkçede; güçlü, sağlam, istikrarlı anlamlarına karşılık geldiği görülmektedir. İstatistiksel sağlamlık ise dar bir anlamda Huber' ın (1981) tanımına göre "sağlamlık, varsayımlardan küçük sapmalara karşı duyarsızlığı" ifade eder.

Hampel vd. (1986) sağlam istatistiği, geniş gayri resmi anlamda, istatistikteki idealize edilmiş varsayımlardan sapmalarla ilgili olarak kısmen "sağlamlık teorileri" şeklinde resmileştirilmiş bir bilgi bütünü olarak tanımlamıştır. Varyans analizimiz, güçlü regresyonumuz veya çok değişkenli istatistiklerdeki yüksek kapsamlı kovaryans matrislerimiz gibi tanıdık ve basit modellerimizden vazgeçmememizi; ancak bunları sağlam tahminciler ile biraz değiştirmenin avantajlı olduğunu belirtmişlerdir. Verilere bakmak ve göze çarpan gözlemleri yeniden kontrol etmek, sağlamlığa doğru bir adımdır; normalden sapan değerleri hariç tutmak gayri resmi bir sağlam prosedürdür.

Örneğin Medyan sağlamdır. Uzun kuyruklu veriler açısından ortalamadan (mean) medyana (median) geçiş sağlam bir yöntemdir. En olası değer ikinci en olası değerden açıkça daha mümkünse; Mod (ayrık, soyut veya gruplanmış veriler için) sağlam olabilir (Hampel vd. 1986).

Yaklaşık normal dağılım varsayımı, merkezi yayılımı normal bir şekle sahip, ancak kuyrukları normal dağılımdan daha ağır veya kalın kuyruk olan yapı türleri varsa aykırı değer oluşumuna sebep olabileceği dikkat edilmesi gereken bir durumdur. Verilerin normal dağıldığı varsayılır, ancak gerçek dağılımının kalın kuyrukları varsa o zaman maksimum olabilirlik ilkesine dayalı tahminler yalnızca "en iyi" olmaktan çıkmakla kalmaz, aynı zamanda kuyruklar simetrikse kabul edilemez derecede düşük istatistiksel verimliliğe sahip olabilir ve kuyruklar asimetrikse çok büyük önyargıya sahip olabilir (Maronna vd. 2006).

Sağlam teknik terimi ilk 1953' te George E. P. Box tarafından sözcük olarak ortaya atılmış ve sağlamlığı pratikte dağılıma ait varsayımlardaki sapmalara karşı sağlam sonuçlar veren istatistiksel yöntemler için kullanmıştır. Huber' ın (1964) sağlamlık ile ilgili temel çalışmaları sayesinde sağlamlık artık saygınlık katmak için yardım alınılan sihirli bir kelime haline gelmiştir. Hampel (1971) de sağlamlığın genel niteleyici yapısını Hampel (1968), Tukey (1960), Huber (1964), (1968a) ile karşılaştırarak tanımlamıştır. Huber (1981), esas dağılım formunun varsayılan modelden (genellikle Gauss yasası) küçük sapmalar göstermesini; dağılımsal sağlamlık (distributional robustness) olarak ve modelde bilinmeyen regresyon, konum ve ölçek parametrelerinin tahmin edilmesinde varsayımlardan küçük sapmaların tahmin edicileri etkilemediği durumu ise tahmin edicilerin sağlamlığı olarak tanımlamıştır. Dağılımsal sağlamlık için dirençli (resistant) istatistik terimini, dağılımsal olmayan örneğin olasılık modeli veya tahmin edicisi hakkındaki varsayımların ihlaline karşı kullanılan yöntemler için sağlamlık terimini tercih eden araştırmacılar da vardır. Örneğin, Mosteller ve Tukey (1977) tarafından tahmin edicinin (veya test istatistiğinin) değeri, temel alınan örneklemdeki küçük değişikliklere karşı duyarsızsa, istatistiksel bir prosedür dirençli (resistant) olarak adlandırılmıştır.

Sağlamlık sorununa yönelik çok çeşitli yaklaşımlar mevcuttur. Bazıları sağlamlığın oldukça genel ve soyut kavramlarını düşünür; bazıları, matematiksel problemleri uğruna sağlamlıkla ilgili farklı topolojileri veya diğer matematiksel yönleri inceler; birçoğu, simetri gibi kavramları kullanarak, kapsamı bir tür parametrik olmayan istatistiklerle genişletmeye çalışır. Yaygın bir yaklaşım, belirli bir parametrik modeli başka bir modelle değiştirmek, özellikle daha fazla parametre ekleyerek onu bir "süper model" haline getirmektir (Hampel vd. 1986).

Hampel vd. (1986) sağlam istatistiklerin temel amaçlarını aşağıdaki gibi tanımlamıştır:

(a) Veri kütlesine en uygun yapıyı tanımlamak.
(b) Daha ileri işlemler için sapan veri noktalarını (aykırı değerleri) veya sapan altyapıları belirlemek.
(c) Etkili veri noktalarını ("kaldıraç noktaları") belirlemek ve bunlarla ilgili uyarı vermek.
(d) Şüphelenilmeyen seri korelasyonlarla veya daha genel olarak varsayılan korelasyon yapılarından sapmalarla ilgilenmek.

Hampel (1974) sağlam tahmin ediciler için temel sezgisel koşulları şu şekilde özetlemiştir: Niteliksel sağlamlığa karşılık gelen küçük bozulmalara çok az tepki vermeli ve yüksek kırılma (breakdown) noktasına karşılık gelen yüksek bozulma (veya birçok büyük hata) varlığında bile güvenli olmalıdırlar. Düşük bir brüt hata duyarlılığına karşılık gelen herhangi bir sabit miktardaki bozulmanın maksimum göreceli etkisi üzerinde bir sınır tutmalıdırlar. Ve muhtemelen yuvarlama ve gruplandırmaya (düşük bir yerel kaydırma duyarlılığı anlamına gelen) da sorunsuz bir şekilde tepki vermelidirler ve açıkça aşırı gözlemleri veri yığınından (düşük bir reddetme noktası anlamına gelen) ayırmalıdırlar. Bu yan koşullar altında, daha sonra Fisher' in orijinal anlamındaki tutarlılığı zorunlu tutularak şekillendirilerek doğru miktarı yine de tahmin etmeli ve ideal parametrik model altında mümkün olduğu kadar küçük bir varyansa sahip olmalıdırlar, bu da maksimum olabilirlik tahmin edicisi ile yüksek oranda ilişkili olmak anlamına geldiği belirtilmiştir.

Bir kestiricinin sağlamlığını ve dirençliliğini değerlendirmede bazı sağlamlık ölçütleri: Duyarlılık eğrisi, Büyük hata duyarlılığı, Yerel kayma duyarlılığı, Reddetme noktası, Etki eğrisi veya Etki fonksiyonu olarak adlandırılmaktadır. Bu ifadelerin literatürde karşılaşılan tanımlarına bakalım.

Sağlamlık ölçütü olarak, yalnızca verilere güvenmek yerine, rastgele değişkenlerin dağılımını kullanabiliriz. Verilerin dağılımını biraz değiştirdiğimizde bir tahmin ediciye ne olduğunu görmek için duyarlılık eğrisi, eklenen gözlemin kestirici üzerindeki etkisini gösterir. Nitekim, örneklemde meydana gelebilecek olası bir hatanın etkisini tanımlamak için kullanılır**.** Tukey' nin (1970-1971) duyarlılık eğrisi (sensitivity curve): eklemeli ve değiştirmeli olmak üzere iki versiyonu bulunmaktadır. Ek bir gözlem olması durumunda, $\{T_n; n \geq 1\}$ tahmin edici dizisi ve $n-1$ gözlemin bir $(x_1, \dots x_{n-1})$ örneklemi ile başlanır ve duyarlılık eğrisi $SC_n(x) = n[T_n(x_1, \dots x_{n-1}, x) - T_{n-1}(x_1, \dots x_{n-1})]$ olarak tanımlanır (Hampel 1974, Hampel vd. 1986).

Etki fonksiyonu $IF(x; T, F)$, F dağılımına sahip bir örneklemin verilen herhangi bir x noktasındaki ek bir gözleminin, bir T istatistiği üzerindeki etkisini tanımak üzere, F deki T nin büyük hata duyarlılığı (gross-error sensitivity), $IF(x; T, F)$ nin mevcut olduğu tüm x' ler için $\gamma^*(T, F) = \sup_x |IF(x; T, F)|$ şekilde tanımlanır. Merkezi yerel sağlamlık ölçüsü olarak büyük hata duyarlılığı, tahmin edicinin değeri üzerinde sabit boyuttaki küçük bir bozulma miktarının sahip olabileceği en kötü (worst influence) etkiyi ölçer. Bu nedenle, tahmin edicinin asimptotik yanlılığı için bir üst sınır olarak kabul edilebilir. $\gamma^*(T, F)$ nin sonlu olması arzu edilen bir özelliktir. Büyük hata duyarlılığı tipik olarak, özellikle M-tahmin edicileri için, sağlamlığın iyi bir nicel ölçüsü olmasına rağmen, bunda yanıltıcı veya doğrulanmamış iddialara şüpheyle bakılması gereken durumlar vardır. Özellikle, R-tahmin edicileri (rank testlerinden türetilen tahmin ediciler, bkz. Andrews vd. 1972) sonsuz bir brüt hata duyarlılığına sahip olabilir ve yine de oldukça kararlı davranırken, diğer yandan L-tahmin edicileri sınırlı bir etki eğrisine sahip olabilir ve yine de bir anda bozulabilir. Bu nedenle, tahmin edicilerin bir süreklilik (niteliksel sağlamlık) özelliğini dikkate almak gerekli hale gelir. Böyle genel bir "nitel sağlamlık" kavramı Hampel' de (1968, 1971) tartışılmış ve tanımlanmıştır.

Kabaca söylemek gerekirse, eğer iki ampirik kümülatif (empirical cumulatives) birbirine "yakın" ise (zayıf anlamda, örneğin merkezi limit teoreminde kullanılır), bu örneklemlere dayalı tahminler birbirine yakın olmalıdır. Niteliksel sağlamlığın yanı sıra en önemli sağlamlık koşulları, yüksek kırılma noktası ve düşük büyük hata duyarlılığıdır (Hampel 1974, Hampel vd. 1986).

Gözlemlerdeki küçük değişimler için de bir ölçü gerekebilir. Bazı değerler biraz değiştiğinde, bunun tahmin üzerinde belirli bir ölçülebilir etkisi vardır. Sezgisel olarak, y' de bir gözlem eklenip ve x' te bir gözlem daha çıkarılarak, bir gözlemi x noktasından komşu bir y noktasına hafifçe kaydırmanın etkisi ölçülebilir. Böyle bir ölçü, etki eğrisinin uyduğu en küçük *Lipschitz* sabiti, yerel-kayma-duyarlılığı (local-shift sensitivity) $\lambda^* = \sup_{x \neq y} |IF(y;T,F) - IF(x;T,F)|/|y - x|$ tarafından sağlanır. Yuvarlatma veya gruplandırmanın kısmi etkileri düşünüldüğünde önemlidir, ayrıca aykırı değerlerin reddedilmesine ilişkin ortaya çıkar. Bununla birlikte, λ^* nın doğru yorumlanması için, bunun yalnızca tahmin edicinin değerindeki standartlaştırılmış yerel değişikliklere atıfta bulunduğu akılda tutulmalıdır. İlginç bir şekilde, aritmetik ortalama, tüm konum tahmin edicileri arasında en düşük yerel kaydırma duyarlılığına sahiptir ve bu özel anlamda "en sağlam" tahmin edicidir, ancak diğer çok daha önemli anlamlarda büyük ölçüde sağlam değildir (Hampel 1974, Hampel vd. 1986).

Tahmin edicinin aykırı değeri reddedip reddetmediğini ve eğer öyleyse hangi mesafede reddedileceğini veren ölçü $p^* = \inf_{r>0}\{r > 0; IF(x;T,F) = 0\ when\ |x| > r\}$ ile tanımlanır. IF nin belirli bir alanın dışında kaybolduğu anlamına gelir. Yani, IF bazı bölgelerde sıfıra eş ise, bu noktalardaki bozulmanın hiç bir etkisi olmaz. Etki eğrisinin aynı şekilde sıfır olduğu bir dış bölge olabilir ve bu özelliğe sahip en küçük bölgenin sınırı özel bir rol oynar. En uzak noktasının uygun bir dağılım merkezinden (örneğin simetri merkezi) uzaklığı "red noktası (rejection point)" p^* olarak adlandırılabilir. Böylece, reddetme noktasından daha uzaktaki tüm gözlemler (ve muhtemelen diğerleri de) tamamen reddedilir. p^* nin sonlu olması arzu edilen bir özelliktir. Sonlu reddetme noktası ve sonlu yerel-kayma duyarlılığının her zaman uygun olmayabileceğini unutmamalıyız (Hampel 1974).

Hampel vd.' e (1986) göre en zengin nicel sağlamlık bilgisi, etki eğrisi (influence curve, IC) veya etki fonksiyonu (influence function, IF) ve türetilmiş sayılar tarafından sağlanır. Başlangıçta, geometrik yönlerini buluşsal olarak bakılması ve yorumlanması gereken bir şey olarak vurgulamak için "etki eğrisi" veya IC olarak adlandırılmış; daha sonra, daha çok yüksek boyutlu uzaylara genelleştirilmesi nedeniyle, daha sık olarak "etki fonksiyonu" veya IF olarak adlandırılmıştır.

Etki fonksiyonu (IF), F dağılımına sahip (büyük) bir örneklem verilen herhangi bir x noktasındaki ek bir gözlemin, bir T istatistiği üzerindeki etkisini tanımlar. Kabaca söylemek gerekirse, etki fonksiyonu $IF(x;T,F)$, bir T istatistiğinin, F altında dağılımındaki ilk türevidir; burada x noktası, olasılık dağılımlarının sonsuz boyutlu uzayında koordinat rolünü oynar. Yani, bir $\{T_n; n \geq 1\}$ tahmin edici dizimiz ve $n-1$ gözlemlerinden bir örneklem $(x_1, \dots, x_{n-1})$ olduğunu varsayalım. Tahmin edicinin ampirik IF değeri, $T_n(x_1, \dots, x_{n-1}, x)$ nin x' in bir fonksiyonu olarak grafiğidir. Etki fonksiyonu, kestiricinin herhangi bir noktadaki küçük bozulum miktarına nasıl karşılık vereceği hakkında kesin bir bilgi verir (Hampel vd. 1986).

Etki eğrisi (IC), bazı dağılımlarda fonksiyonel olarak görülen bir tahmin edicinin esasen ilk türevidir ve sadece asimptotik varyansları türetmek için değil, aynı zamanda tanımlanan ve sezgisel olarak yorumlanan birkaç yerel sağlamlık özelliklerini incelemek için de nasıl kullanılabileceğini gösterir. Etki eğrilerinin incelenmesi, tahmin ediciler hakkındaki anlayışımızı derinleştirmeye hizmet eder. Aynı zamanda, önceden belirlenmiş sağlamlık özelliklerine sahip yeni tahmin edicilerin türetilmesine de hizmet eder (Hampel 1974).

Kırılma noktası, IF tarafından sağlanan yerel lineerleştirmenin modelden ne kadar uzakta kullanılabileceği konusunda olan ihtiyacımızı karşılar. Kabaca söylemek gerekirse, istatistiği tüm sınırlar boyunca taşıyabilen, büyük hataların en küçük oranıdır. Daha kesin olarak, istatistiğin tamamen güvenilmez ve bilgisiz hale geldiği model dağılımından olan mesafedir (Hampel vd. 1986).

Yijun Zuo‘ ya (2001) göre, kırılma noktası kavramı ilk olarak Hodges' de (1967) konumun tek boyutlu tahminiyle sınırlı olarak ortaya çıkmış ve daha sonra Hampel (1968, 1971) tarafından çok daha genel bir formülasyon verilmiştir. Kırılma noktasının sonlu örneklem versiyonları, etki fonksiyon yaklaşımı tarafından yakalanan tahmin edicilerin yerel sağlamlığının değerlendirmesini tamamlayan, tahmin edicilerin küresel sağlamlığının en yaygın sayısal değerlendirmeleri haline gelmiştir (bkz. Hampel vd. 1986; Rousseeuw ve Leroy 1987). Aslında, sağlamlık konusundaki son araştırmaların birincil hedeflerinden biri, Rousseeuw (1984) tarafından doğrusal modelde örneklenen yüksek kırılma yöntemlerinin ve Lopuhaa ve Rousseeuw (1991), Tyler (1994) ve Donoho ve Gasko (1992) tarafından konum ve dağılım (scatter) modellerinin oluşturulması olmuştur. Kırılma noktası ile ilgili çalışmalar Davies (1987), Maronna ve Yohai (1991), Stromberg ve Ruppert (1992) ve Stromberg (1993), Muller (1995), Chen (1995), Zhang ve Li (1998), Singh (1998), He ve Fung (2000), He vd. (2000), Ghosh ve Sengupta (1999), Chang vd. (1999) olarak sayılabilir.

Hampel vd. (1986) tarafından önemli bir küresel sağlamlık yönüyle etki fonksiyonlarına dayalı yaklaşım, “kırılma noktasından (breakdown point)” oluşmasına rağmen, “bölünemeyecek kadar küçük yaklaşım (infinitesimal approach)” olarak da adlandırılmıştır. İnfinitesimal yaklaşım: niteliksel (qualitative) sağlamlık, etki fonksiyonu ve kırılma noktası olarak üç temel kavrama dayanmaktadır. Bir fonksiyonun sürekliliğine ve ilk türevine ve en yakın ucunun mesafesine açık bir şekilde karşılık gelirler. 1976 ve 1977' de başlıca optimallik sonucu tek boyutlu durumdan genel parametrik modellere genişletilmiş, Hampel' de (1978a) ve Krasker' de (1980) yayınlanmıştır (bkz. Stahel 1981a). Ronchetti (1979), Rousseeuw (1979) ve Rousseeuw ve Ronchetti (1981) yaklaşımı tahmin edicilerden testlere genelleştirmiştir.

$\{T_n\}$ bir tahmin ediciler dizisi olmak üzere, bazı F olasılık ölçülerinde $\{T_n\}$ nin kırılma noktası (breakdown point) δ^* şu şekilde tanımlanır; $\delta^* = \delta^*(\{T_n\}, F) = \sup\{\delta \leqq 1; \exists\ kompakt\ küme\ K = K(\delta)\ parametre\ uzayının\ alt\ kümesi\ öyle\ ki\ \pi(F, G) < \delta \Rightarrow G\{T_n \in K\} \to 1\ olarak\ n \to \infty\}$. Kırılma noktası, tahmin edicinin parametrik

modelden hangi Prokhorov uzaklığına kadar bize parametrik model içindeki orijinal dağılımın bazı göstergelerini verdiğini söyler (Hampel 1971).

Kırılma noktasının (breakdown point) basit bir sonlu örneklem versiyonu olarak; orijinal veri noktalarının herhangi bir m sinin keyfi değerlerle değiştirilmesiyle elde edilen tüm olası bozulmuş Z' örneklerini ele alalım. $bias(m;T,Z) = \sup_{Z'} \|T(Z') - T(Z)\|$ sonsuzsa, m aykırı değerlerinin T üzerinde keyfi olarak büyük bir etkiye sahip olabileceği anlamına gelir ve tahmin edicinin "bozulduğu (breaks down)" söylenerek ifade edilebilir. Bu nedenle, Z örneklemindeki tahmin edici T nin kırılma noktası $\delta_n^*(T,Z) = \min\{\frac{m}{n}; bias(m;T,Z)\ sonsuz\}$ olarak tanımlanır. Başka bir deyişle, tahmin edici T nin $T(Z)$ den keyfi olarak uzak değerler almasına neden olabilecek en küçük bozulma oranıdır. En küçük kareler için, T yi tüm sınırlar boyunca taşımak için bir aykırı değerin yeterli olduğu bilinmektedir. Bu nedenle, kırılma noktası $\delta_n^*(T,Z) = 1/n$' e eşittir ve bu, artan örneklem büyüklüğü n için sıfır olma eğilimindedir, dolayısıyla LS nin %0' lık bir kırılma noktasına sahip olduğu söylenebilir. Bu yine LS yönteminin aykırı değerlere aşırı duyarlılığını yansıtır (Rousseeuw ve Leroy 1987).

4.2 Aykırı Değer Durumunda Sağlam Tahmin Ediciler

Klasik kestirim yöntemleri, örneğin en küçük kareler yöntemi kullanıldığında tek bir aykırı gözlem bile kestirimlerimizin yanıltıcı çıkmasına neden olabilirken, sağlam yöntemler kullanıldığında aykırı gözlemin etkisi en aza indirilmiş olur. Sağlam kestirimin temel amacı veride aykırı değerler olduğunda da anlamlı kestirimler elde etmektir.

Çok değişkenli veri setinde klasik sağlam yöntemler, satırsal aykırı değerler olduğunda anlamlı kestirimler elde etmektedir. Ancak, satırsal aykırı değer üzerine daha iyi olan klasik sağlam yaklaşımlar için, hücresel aykırı değer üzerinde de aynı etkiyi gösterdiğini söyleyemeyiz. Çok değişkenli hücresel aykırı değer durumunda sağlam kesirim yöntemleri üzerinde çalışmalar devam etmektedir.

Sağlam bir regresyon tahmin edicisi için ilk adım, Boscovich' in bir önerisini geliştirerek Edgeworth' ten (1887) gelmiştir. Aykırı değerlerin bilinen LS (Least Squares) üzerinde çok büyük bir etkiye sahip olduğunu savunmuştur çünkü $r_i = y_i - \hat{y}_i$ artıkların L_2 de kareleri alınmaktadır. Bu nedenle, L_1 ile belirlenen en küçük mutlak değerler (least absolute) regresyon tahmin edicisini önermiştir. Önerilen teknik genellikle $LA = L_1 : \underset{\hat{\beta}}{\text{Minimize}} \sum_{i=1}^{n} |r_i|$ regresyonu olarak adlandırılır, ayrıca en küçük kareler $LS = L_2 : \underset{\hat{\beta}}{\text{Minimize}} \sum_{i=1}^{n} r_i^2$ ' dir. En küçük kareler için, aykırı değer varlığında kırılma noktası $1/n$' e eşittir ve büyük n-örneklem değerleri için sıfıra eğilimlidir; bu nedenle LS nin %0' lık bir kırılma noktasına sahip olup, yine LS yönteminin aykırı değerlere karşı aşırı duyarlı olduğu söylenebilir (Rousseeuw ve Leroy 1987).

Aykırı değerlere karşı geçici olarak bir miktar koruma sağlayan sağlam istatistiksel yöntemleri göstermek için, bir konum parametresi μ için çeşitli sağlam tahmin yöntemleri ele alınır. Amaç, örneklemdeki aykırı değerlerin μ tahmini değeri üzerindeki etkisini azaltmaktır. Genel olarak karşılaşılan bu yöntemler: L-tahmin edicileri sıralı istatistikleri kullanarak oluşturulan lineer eşitlikler ile tahmin vermektedir, M-tahmin edicileri ise en çok olabilirlik tahmin edicilerini kullanarak tahmin değerini elde ederler, R-tahmin edicileri sıralama test yöntemlerini kullanırlar. Ayrıca diğer tahmin edici tipleri: A, D, P, S, W de litaritürde kullanılmaktadır (Hampel vd. 1986).

Aykırı değer durumunda literatürde karşılaşılan başlıca kestirim yöntemlerinin hangi durumlarda kullanılacakları ve nasıl uygulanacakları aşağıda kısaca anlatılmıştır.

4.2.1 L-Tahmin ediciler (Linear order statistics estimators, L-estimators)

L-tahmin edicileri, sıralanmış örneklem değerlerinin doğrusal kombinasyonları biçimindeki tahmin edicilerinin kullanımının, budanma (α-trimmed mean) ve winsorizasyon (Winsorized mean) yaklaşımlarının bir genelleştirmesi olduğu söylenebilir.

Gözlemlerin sıralandığını varsayalım, $x_1 \leqq x_2 \leqq \cdots \leqq x_n$ o zaman $T = n^{-1}(gx_{g+1} + x_{g+1} + x_{g+2} + \cdots + x_{n-h} + hx_{n-h})$ istatistiğine "Winsorized ortalama" denir, g en soldaki ve h en sağdaki gözlemin winsorize edilmesiyle elde edilir (Huber 1964).

Burada amaç, örneklemdeki aykırı değerlerin μ tahmini değeri üzerindeki etkisini azaltmak için, c_i ağırlıklarının uçlarda veri setinin ortada olanlarınkine göre daha düşük olduğu sıralı örneklem değerlerinin $\tilde{\mu} = \sum c_i x_i$ lineer kombinasyonları şeklinde tahmin ediciler kullanmaktır. Bu tür "lineer sıralı istatistik tahmin edicilerinin" örneği (Huber, 1972 tarafından 'L-tahmin edicileri' olarak adlandırılır), orta veya iki orta sıralı gözlemler hariç tümü için $c_i = 0$ olan örneklem medyandır (Barnett ve Lewis 1978).

4.2.2 M-Tahmin ediciler (Maximum likelihood type estimators, M-estimators):

Huber' in (1972) maksimum olabilirlik tipi tahmin ediciler olarak adlandırdığı M-tahmin edicileri, $x_1, x_2, \ldots, x_n$' nin rastgele bir örnek olduğu ve $\psi(u)$' nun istenen özelliklere sahip bir ağırlık fonksiyonu olduğu bir $\tilde{\mu}$ tahmin edici elde etmek için

$$\sum_{i=1}^{n} \psi(x_i - \tilde{\mu}) = 0$$

biçimindeki bir denklemin çözülmesiyle elde edilir. Örneğin, $|\psi(u)|$ büyük $|u|$ için küçükse, $\tilde{\mu}$ örneklemdeki uç değerlerin etkisini azaltır ve aykırı değerlere karşı koruma sağlayacaktır. Özel $\psi(u) = f'(u)/f(u)$ seçimi, burada dağılımın $f(x - \mu)$ biçimindeki olasılık yoğunluk fonksiyonuna sahip olması durumunda, elbette olabilirlik denkleminin bir çözümü olarak $\tilde{\mu}$ verecektir (Barnett ve Lewis 1978).

Özetle, bilinen klasik en küçük kareler (LS) denkleminde, artıkların karesi, artıkların başka bir artık fonksiyonu ile değiştirilmesi fikrine dayanır ve bu seçim fonksiyonu, simetrik bir fonksiyondur.

4.2.3 R-Tahmin ediciler (Rank test estimators, R-estimators)

İlk olarak Hodges ve Lehmann (1963) μ tahmin edicilerinin Wilcoxon testi gibi sıra testi yöntemleriyle elde edilebileceğini belirtmişlerdir. Bu tür R-tahmin edicileri genellikle sağlamlık özelliği gösterir. Huber (1972) tarafından konum kayması için iki örneklem sıralama testi için bir örnek verilmiştir. Eğer örneklemler $x_1, x_2, \dots, x_n$ ve $y_1, y_2, \dots, y_n$ ise test istatistiği

$$W(x_1, x_2, \dots, x_n; y_1, y_2, \dots, y_n) = \sum_{i=1}^{n} J[i/(2n+1)]V_i$$

olur, burada $J[i/(2n+1)]$ birleşik örneklem için ampirik dağılım fonksiyonunun uygun olarak seçilmiş bir fonksiyonudur. Eğer birleşik örneklemdeki i. nci sıralı değer x değerlerinden biri ise $V_i = 1$' dir (aksi durumda $V_i = 0$).

$$W(x_1 - \tilde{\mu}, x_2 - \tilde{\mu}, \dots, x_n - \tilde{\mu}; -x_1 + \tilde{\mu}, -x_2 + \tilde{\mu}, \dots, -x_n + \tilde{\mu}) = 0$$

eşitliğinin çözümü olarak bir $\tilde{\mu}$ tahmin edicisi türetebiliriz ve $\tilde{\mu}$ nın asimptotik davranışı testin güç fonksiyonundan elde edilir. Simetrik dağılımlar için, uygun bir $J(.)$ seçimi için R-tahmin edicileri asimptotik olarak etkili ve normal dağılmış olabilir (Barnett ve Lewis 1978).

4.2.4 LMS ve LTS Kestirim yöntemleri

Literatürde bilinen en küçük kareler toplamı $LS : \underset{\hat{\beta}}{\text{Minimize}} \sum_{i=1}^{n} r_i^2$ yöntemi için, birkaç araştırmacı kareyi başka bir şeyle değiştirerek, toplama işaretine dokunmadan bu tahmin ediciyi sağlam hale getirmeye çalışmıştır. Bu yöntemdeki toplamı, çok sağlam olan bir medyanla değiştirme fikrine dayalı bir yöntem olan en küçük medyan kareler (Least Median of Squares, LMS) tahmin edicisi Rousseeuw tarafından 1984 de

tanımlanmıştır. Hem x hem de y ekseninde bulunan aykırı değerlere karşı duyarlı bir kestirim yöntemin gösterimi

$$\underset{\hat{\beta}}{Minimize}\ \underset{i}{med}\ r_i^2$$

şeklindedir. LMS yönteminin kırılma noktasının %50 mümkün olan en yüksek değer olduğu gösterilmiştir (Rousseeuw ve Leroy 1984). LMS için kırılma noktası $([n/2] - p + 2)/n$' dir. Anormal derecede yavaş bir yakınsama oranına sahip olduğunu kanıtlayarak, LMS yönteminin asimptotik etkinlik açısından zayıf olması sebebiyle Rousseeuw tarafından 1984 yılında En Küçük Budanmış Kareler (Least Trimmed Squares, LTS) yöntemi tanıtılmıştır. LTS yönteminin gösterimi

$$\underset{\hat{\beta}}{Minimize} \sum_{i=1}^{h} (r^2)_{i:n}$$

şeklindedir; burada $(r^2)_{1:n} \leq \cdots \leq (r^2)_{n:n}$ sıralı kare artıklardır ve yalnızca ilk h' si toplanır (artıkların önce karesinin alındığını ve sonra sıralandığını unutmamalıdır). $h = [n/2] + 1$ alındığında LTS, LMS ile aynı kırılma noktasına ulaşır. Ayrıca, $h = [n/2] + [(p+1)/2]$ için LTS kırılma noktası $([(n-p)/2] + 1)/n$ olarak tanımlanır. LTS, LS ye çok benzer, tek fark, en büyük karesi alınmış artıkların toplamda kullanılmamasıdır. En iyi sağlamlık özellikleri, h yaklaşık olarak $n/2$ olduğunda elde edilir, bu durumda kırılma noktası %50' ye ulaşır (Rousseeuw ve Leroy 1987).

LTS yöntemi gerek istatistiksel etkinliği, gerekse hesaplama kolaylığı nedeniyle LMS yöntemine göre daha fazla kullanılmaktadır.

4.2.5 S-Kestirim yöntemi

LTS ve LMS artıkların saçılımının sağlam bir ölçümünü minimum yapmaya çalışan yöntemi genelleştiren Rousseeuw ve Yohai (1984); $S(\beta)$' ın $r_1(\beta), \dots, r_n(\beta)$ artıklarının ölçeğinin belirli bir tür sağlam M-tahmini olduğu

$$\underset{\widehat{\beta}}{Minimize}\ S(\beta)$$

'ye karşılık gelen S-tahmin edicilerini tanımlamıştır. S-tahmin edicilerinin, ilgili sabitlerin uygun bir seçimiyle kırılma noktalarının da %50' ye ulaşabildiği gösterilmiştir.

Ölçek tahmini $\hat{\sigma} = s(r_1(\hat{\beta}), \dots, r_n(\hat{\beta}))$ ile

$$\underset{\widehat{\beta}}{Minimize}\ s(r_1(\beta), \dots, r_n(\beta))$$

artıkların dağılımının minimizasyonu ile tanımlanırlar. Bu tahmin ediciye S-tahmin edicisi denir çünkü bir ölçek (scale) istatistiğinden türetilmiştir.

Dağılım $s(r_1(\beta), \dots, r_n(\beta))$

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{r_i}{s}\right) = K$$

eşitliğinin çözümü olarak tanımlanır. Aslında, bu eşitlikte verilen s, ölçeğin (scale) bir M-tahmin edicisidir. S-tahmin edicilerinin temel olarak regresyon M-tahmin edicileri ile aynı asimptotik performansa sahip olduğu söylenebilir (Rousseeuw ve Leroy 1987).

4.2.6 MM-Kestirim yöntemi

Yohai (1985), LMS ve LTS gibi yüksek kırılma (high-breakdown) tahmin ediciler için daha yüksek etkinliğe (efficiency) yönelik MM-tahmin edicileri adını verdiği bir öneri sunmuştur. Yohai' nin tahmin edicileri üç aşamada tanımlanır. İlk aşamada LMS veya LTS gibi bir yüksek kırılma tahmini β^* hesaplanır. Daha sonra, $r_i(\beta^*)$ artıkları üzerinde, %50 kırılımlı s_n ölçeğinin bir M-tahmini, sağlam fit-uyumdan hesaplanır. Son olarak, MM-tahmin edicisi $\hat{\beta}$, $\sum_{i=1}^{n}\psi(r_i(\beta)/s_n)x_i = 0$' ın $S(\beta) \leq S(\beta^*)$' yi sağlayan herhangi bir çözümü olarak tanımlanır, burada $S(\beta) = \sum_{i=1}^{n}\rho(r_i(\beta)/s_n)$ dır. Yani, çok yüksek artık değerlerine sıfır ağırlık veren bir fonksiyon ile M-tahmin edicisi hesaplanır. Verilen denklemlerde, ρ simetriktir ve sürekli türevlenebilir ve $\rho(0) = 0$. c > 0, öyle ki ρ, $[0, c]$ üzerinde kesin olarak artan ve $[c, \infty)$ üzerinde sabittir. $E_H[\|x\|] < \infty$ ve ρ, türevi $\rho' = \psi$ ile $\psi(u)/u$, $u > 0$ için artmayan fonksiyondur (Rousseeuw ve Leroy 1987).

5. HÜCRESEL AYKIRI DEĞER ve KAYIP VERİ DURUMUNDA TAHMİN EDİCİ DEĞERLERİNİN KARŞILAŞTIRILMA ÖRNEĞİ

Çok değişkenli veri setinde kayıp veri olması durumu, hücresel aykırı değer olarak değerlendirilmesi planlanmıştır. Fiyat İstatistikleri Sekreterlikler Arası Çalışma Grubu (IWGPS[5]) kuruluşları tarafından yayınlanan klavuzdaki kayıp verileri içeren veri seti örnek olarak ele alınmıştır. Kılavuzda çok değişkenli veri setinde kayıp veri olması durumunda karşılaşılan bu sorun ele alınmış, kayıp veriyi ve aykırı değeri eş zamanlı aynı veri setinde değerlendirebilmek için buradaki kayıp veri bir aykırı değer olarak değerlendirilmiştir. Diğer bir deyişle, kayıp veriye sıfır (0) fiyat ataması ile fiyat genel düzeyinde sıfır fiyat değeri ile aykırı değere dönüştürülen bu kayıp veri seti için, bilgisayar programlama diliyle sağlam aykırı değer ve kayıp veri imputasyon hesaplamaları yapılmıştır. Ayrıca, IWGPS kuruluşları tarafından TÜFE[6] hesaplamalarında kullanılan imputasyon yöntemlerine alternatif yöntemler önerilmiştir ve imputasyonları hesaplanmıştır. Aynı veri seti için, IWGPS kuruluşlarında kullanılan TÜFE hesaplamalarındaki imputasyon yöntemlerinin sonuçları, istatistiksel programlama dilindeki sağlam aykırı değer ve kayıp veri imputasyon sonuçları ve önerilen yöntemlerin imputasyon sonuçları karşılaştırılmıştır.

Veri yapısı olarak, belirlenmiş mal ve hizmet değişkenlerin belirli zaman periyotlarında aldığı değerler olduğu düşünülürse; hem kesit hem de zaman serisi olması sebebiyle TÜFE nin panel veri olduğu söylenebilir. Ayrıca mutlak farkın değil de oransal farkın önemli olduğu bu hesaplamada, birim fiyatların temsilinin, belirlenmiş mal ve hizmet değişkenlerin geometrik ortalaması ile hesaplanması tavsiye edilmektedir. Fiyatlar genel düzeyinde aynı ürün fiyat değişim takibi varsayımı sebebiyle, bir maddenin fiyat yelpazesindeki kalite kaynaklı fiyat farklılılarının yüksek olması aykırı değer gibi değerlendirilmemelidir. TÜFE de belirlenmiş aynı ürünün zaman içerisindeki fiyat değişim takibi yapılıp, kayıp veri anlamında; geçici ürün olmaması ve kalıcı ürün olmaması kaynaklı olarak iki problem ile karşılaşılmaktadır. Kalıcı ürün olmaması

[5] Fiyat İstatistikleri Kuruluşları Sekreterlikler Arası Çalışma Grubu (Intersecretariat Working Group on Price Statistics, IWGPS); Avrupa Birliği İstatistik Ofisi (Eurostat), Uluslararası Çalışma Örgütü (ILO), Uluslararası Para Fonu (IMF), Ekonomik Kalkınma ve İşbirliği Örgütü (OECD), Birleşmiş Milletler Avrupa Ekonomik Komisyonu (UNECE) ve Dünya Bankası (WB)' dan oluşmaktadır.

[6] TÜFE; Tüketici Fiyat Endeksi.

durumunda birçok yöntem kullanılmakta olup, genellikle ürün değişimi yani ikame (substitution), zincir yapı bozulmayacak şekilde imputasyon ile yapılır. Bu çalışmada geçici ürün olmaması durumundaki imputasyon yöntemine alternatif yöntem önerilecektir.

TÜFE hesaplamalarında, anket döneminde ürün bulunamaması yani kayıp veri durumda uygulanan yöntemler ve metedoloji için IWGPS kuruluşları tarafından ortaklaşa yayınlanan klavuzu inceleyelim. Bu klavuza göre TÜFE hesaplamalarındaki bazı metedolojik tanımları aşağıda tanımlayalım (Anonymous 2020).

2020 Tüketici Fiyat Endeksi Kılavuzu, Kavramlar ve Yöntemler (2020 Consumer Price Index Manual, Concepts and Methods), bir tüketici fiyat endeksinin (TÜFE, CPI, consumer price index) derlenmesi hakkında kapsamlı bilgiler ve açıklamalar içerir. Kılavuz, ulusal istatistik ofislerinin (NSO, national statistical offices) bir TÜFE nin derlenmesindeki çeşitli sorunların nasıl ele alınacağına ilişkin kararlar alırken göz önünde bulundurması gereken yöntem ve uygulamalara genel bir bakış sağlar (Anonymous 2020).

Kılavuz, TÜFE yi derlemek ve eğitim amacıyla bir referans olarak istatistik ofisleri (veya diğer derleyici kurumlar) için tasarlanmıştır. İşverenler, işçiler, politika yapıcılar ve araştırmacılar gibi TÜFE kullanıcıları da hedef alınır. Kılavuz, onları yalnızca veri toplama ve bu tür endeksleri derlemede kullanılan farklı yöntemler hakkında bilgilendirmekle kalmayıp, aynı zamanda sonuçların doğru yorumlanabilmesi için TÜFE sınırlamaları hakkında da bilgi sağlamaktadır (Anonymous 2020).

Tüketici Fiyat Endeksi TÜFE, tüketim malları ve hizmet fiyatlarının bir dönemden diğerine değişme oranını ölçen bir endekstir. Fiyatlar mağazalardan veya diğer perakende satış noktalarından alınır. Genel hesaplama yöntemi, farklı ürünler için dönemler arası fiyat değişimlerinin ortalamasını almak ve hanehalklarının bunlara harcadığı ortalama miktarları ağırlık olarak kullanmaktır. TÜFE ler genellikle ulusal istatistik ofislerinin, çalışma bakanlıkları veya merkez bankaları tarafından üretilen

resmi istatistiklerdir. Mümkün olan en kısa sürede, genellikle referans döneminden sonraki dört hafta içinde yayınlanırlar (Anonymous 2020).

TÜFE, kısmen üretildiği sıklık ve güncellik nedeniyle, bir bütün olarak ekonomi için bir enflasyon ölçüsü olarak yaygın şekilde kullanılır. Ekonomik politika oluşturma amaçları, özellikle para politikası için önemli bir istatistik haline gelmiştir. Ödemeleri (ücretler, kiralar, faiz, sosyal güvenlik, diğer yardımlar ve emekli maaşları gibi) enflasyonun etkilerine göre ayarlamak için uygun önlem olarak genellikle mevzuatta ve çok çeşitli sözleşmelerde belirtilir. Bu nedenle, hanehalkları için olduğu kadar hükümetler ve işletmeler için de önemli ve geniş kapsamlı finansal sonuçları olabilir (Anonymous 2020).

Tüketici Fiyat Endeksi Sınıflandırma Sistemi (The Consumer Price Index Classification System), adından da anlaşılacağı gibi, "amaç" ilkesi üzerine kuruludur. Amaç tipi bir sınıflandırmadır çünkü birleştirme programı boyunca ürünler genellikle yerine getirdikleri taşıma, beslenme, barınma vb. amaca (veya işleve) göre gruplandırılır. Bu nedenle amaca dayalı bir sınıflandırma, bir TÜFE için mantıksal sınıflandırma sistemi gibi görünmektedir (Anonymous 2020).

Resmi "Uluslararası Bireysel Tüketimin Amaca Göre Sınıflaması (COICOP, Classification of Individual Consumption According to Purpose)", beş basamaklı bir sınıflandırmadır. Ulusal istatistik ofisleri, kullanımları için daha fazla ayrıntı elde etmek için COICOP u altı ve yedi haneye genişletir. Sınıflandırmanın bir üst seviyesinde ürünler amaca göre gruplandırılır. Hanehalkları tüketim amaçlarını (yani barınmak için bir daire kiralamak veya beslenmek için bir elma yemek) gerçekleştirmek için çeşitli mal ve hizmetleri seçeceklerdir. Bu mal ve hizmetler çeşitli gruplara ayrılır ve yalnızca amaç ilkesine göre değil, aynı zamanda ürün türüne göre de değerlendirebilir. Örneğin portakal ve elma "Meyve" kategorisine dahildir (Anonymous 2020).

Veri kaynakları (data sources), nüfus kapsamına bağlı olarak, bir TÜFE için ağırlıklar, ya Hanehalkı Bütçe Anketi (Household Budget Survey, HBS) kapsamındaki bir

örneklemden alınan tahminlere dayanan harcama verilerinden ya da hanehalkı nihai tüketim harcamalarının ulusal hesap tahminlerinden elde edilir (Anonymous 2020).

Fiyatların Toplanması ve Düzenlenmesi (Collecting and Editing the Prices) için iki temel fiyat toplama yöntemi vardır:

(a) Yerel fiyat toplama (Local price collection): Fiyatların ülke genelinde bulunan satış noktalarından alındığı yerel fiyat toplamaya, lisanslı ve lisanssız pazarlar ve sokak satıcıları ile dükkanlar da dahildir. Normalde fiyat toplayıcının fiyatları gözlemlemek için satış noktasını ziyaret etmesi gerekecektir, ancak bazı kalemlerin fiyatları telefon ve fiyat listeleri dahil olmak üzere başka yollarla da alınabilir,

(b) Merkezi fiyat toplama (Central price collection): genellikle fiyatların saha çalışmasına gerek kalmadan merkez tarafından toplanabildiği durumlarda kullanılır. Bu, düzenleyici makamlardan alınabilen merkezi olarak düzenlenen veya merkezi olarak sabitlenen fiyatları da içerebilir, ancak bu durumlarda söz konusu mal ve hizmetlerin gerçekten mevcut olduğundan ve belirtilen fiyattan satıldığından emin olmak için kontrollerin yapılması gerekecektir (Anonymous 2020).

Veri Düzenleme (Data Editing) sürecinde fiyat bilgisi toplanıp kaydedildikten sonra düzenlenmesi gerekir. Veri düzenleme, temel kitle fiyat endekslerinin hesaplanması için doğru ve kullanılabilir verilerin sağlanması sürecidir. Veri düzenlemeye bazen girdi düzenleme denir. Bu süreçte üç adım vardır:

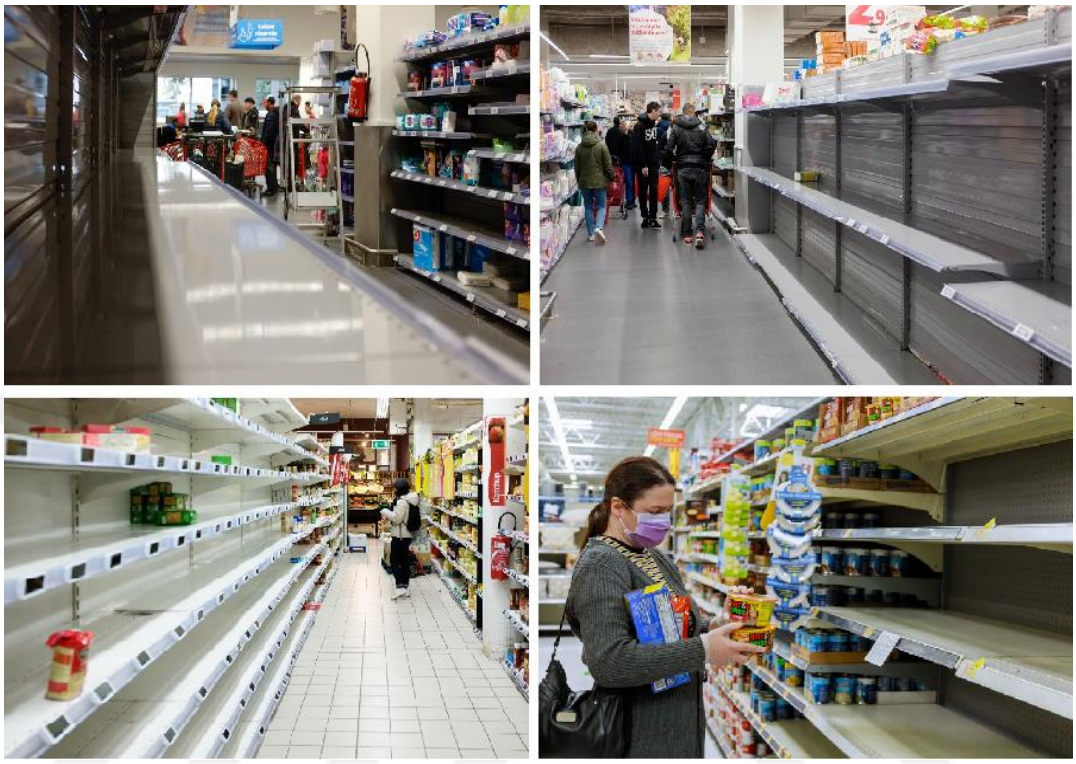
(a) Olası hataların ve aykırı değerlerin tespiti

(b) Verilerin doğrulanması ve düzeltilmesi

(c) Çıktı incelemesi

Fiyat bilgilerinin toplanması ve kaydedilmesindeki hataların tespiti, bilgiler toplandıktan sonra **mümkün olan en kısa sürede yapılmalıdır**. Tespit genellikle fiyat hareketlerini inceleyerek ve önceden tanımlanmış bazı limitleri aşan veya mevcut tüm bilgilerin analizine dayalı olarak gerçekçi görünmeyenleri kontrol ederek

gerçekleştirilir. Çıktı incelemesiyle derleyici, endekslerin gerçeği yansıtmasını sağlar (Anonymous 2020).



Şekil 5.1 Ürünlerin alanda geçici olmaması görselleri

Düzenli olarak meydana gelen üç problem durumu şunlardır:

(a) Bir madde (item), ürün (product) veya satış noktasının kaybolduğu (kapandığı) durumlarda, yeni ürünlerin getirilmesi de dahil olmak üzere ikame prosedürleri,

(b) Bir ürün geçici olarak stokta kalmadığında fiyat imputasyonu (mevsimlik ürünler hariç),

(c) Ürün değişikliğinin fiyat belirleyici özelliklerinde değişiklikler içerdiği durumlarda kalite ayarlaması.

Ürün (Products): Seçilen bir ürün geçici olarak kayıp ise ve hiçbir fiyat kaydedilmemişse, fiyat toplayıcısı (price collector) tarafından bu hususta bir not yazılmalıdır. Geçici ve kalıcı olarak kayıp ürünlerle ilgili sorunlar farklı olduğundan ve bunların çözümlemesi farklı olduğundan, fiyat toplayıcısı için bir ürünün mevcut olmamasının geçici bir kayıp ürün mü yoksa kalıcı bir kayıp ürün mü olduğunu belirlemek önemlidir. Aynı ürünün makul bir süre içinde piyasaya geri dönme olasılığı varsa, bir fiyat geçici olarak kayıp olarak kabul edilebilir. **Geçici olarak kayıp olan bir ürün için bir fiyat imputasyon edilmesi zorunludur.** Mevsimsel olmayan ürünler ve çeşitler, önceden tanımlanmış bir süreden daha fazla kayıp ise değiştirilmelidir. Örneğin, art arda üç ay boyunca stokta yoksa, toplayıcıya ürün tanımına mümkün olduğunca yakın bir ikame seçmesi talimatı verilmelidir (Anonymous 2020).

Geçici (Mevsimsel Olmayan) Kayıp Ürünler (Temporarily (Nonseasonal) Missing Products): Kayıp bir ürünün makul bir süre içinde tekrar bulunacağına inanılıyorsa, fiyat istatistikçisinin üç seçeneği vardır.

(a) Eşleşen çiftler kullanılarak önceki dönem ile benzeri (like-for-like) bir karşılaştırmanın sürdürülmesi için fiyatın kayıp olduğu çeşidi atlanır. Temel endeks, yalnızca fiyat toplayıcının mevcut ve önceki dönemlerde tam olarak aynı çeşitlilikte fiyatlar elde ettiği gözlemleri kullanır. Bu yaklaşımda, silinen ürünün kaybolmadan hemen öncesine kadar kaydedilen fiyat değişikliği o andan itibaren dikkate alınmayacaktır. Bu, örneğin örneklemin dengesini bozarsa sorunlara neden olabilir.

(b) Gözlenen son fiyatı ileriye taşıma (carryforward): Son gözlenen fiyatın ileriye taşınması, yalnızca sabit veya düzenlenmiş fiyatlar olması durumunda tavsiye edilir. Bu durum gözlem yapılamayan dönemlerde fiyat sürekliliği sağlasa da, fiyatların mevcut olmadığı durumlarda söz konusu alt endeksler herhangi bir değişiklik gösteremeyeceğinden, endeksteki kısa vadeli hareketlerin yanlı olması muhtemeldir. Fiyatlar genel olarak yükseliyorsa eğilim (bias) aşağı, fiyatlar düşüyorsa eğilim yukarı olacaktır. Özellikle yüksek enflasyon olduğunda veya fiyat endeksinde (yıllık hareketlerin aksine) dönem dönem hareketlerin önemli olduğu durumlarda, ileriye

taşıma (carryforward) önerilmez. İleriye taşıma yöntemi, yalnızca fiyatın değişmediğine inanmak için bir neden varsa uygundur. Tipik olarak, fiyat sabitlenmedikçe veya düzenlenmedikçe, fiyat istatistikçisinin fiyatın değişmediği inancını doğrulaması zor olacaktır.

(c) İmputasyon (imputation): Şimdiye kadarki en iyi çözüm, bir fiyat impute etmektir. İmputasyon, fiyat hareketinin yansız bir tahminini sağlamak için mevcut en iyi bilgileri kullanır. Temelde iki seçenek vardır:

(i) Kayıp fiyatı, temel toplamda **(genel ortalama imputasyon**, overall mean imputation) mevcut olan fiyatlar için ortalama fiyat değişikliğine atıfta bulunarak impute etmek. Bu, kayıp ürünün fiyat değişiminin, eğer mağazada mevcut olsaydı, temel toplamdaki fiyatlardaki ortalama değişime eşit olacağını varsayar. Temel toplam oldukça homojen ise, bu makul bir varsayım olabilir.

(ii) Kayıp fiyatı, başka bir benzer satış noktasındaki "karşılaştırılabilir" çeşitlerin fiyatları için ortalama fiyat değişikliğine (**sınıf ortalama imputasyon**, class mean imputation) atıfta bulunarak impute etmek. Bu, kayıp ürün ile tahmini fiyat sağlayan ürünler arasında daha kesin bir eşleşmeyi temsil eder (Anonymous 2020).

Sabit veya düzenlemeye tabi fiyatlar haricinde, son gözlenen fiyatın ileriye taşınması önerilmez. Enflasyonun yüksek olduğu dönemlerde veya yüksek oranda yenilik ve ürün devir hızı nedeniyle pazarların hızla değiştiği dönemlerde özel dikkat gösterilmelidir. Uygulaması basit olsa da, son gözlemlenen fiyatı ileriye taşımak, sonuçta ortaya çıkan endeksi sıfır değişime doğru yönlendirir. Ek olarak, kayıp çeşidin fiyatı tekrar kaydedildiğinde, endekste uygun değerine dönmek için telafi edici bir adım değişikliği olması muhtemeldir. Ürün bir süre fiyatlandırılmadan kalırsa, endeks üzerindeki olumsuz etki giderek daha şiddetli olacaktır. Genel olarak, ileri taşımak kabul edilebilir bir prosedür veya kayıp fiyatlar sorununa çözüm değildir (Anonymous 2020).

Çizelge 5.1 Tüketici Fiyat Endeksi Kılavuzu 2020 de Tablo 6.3 deki Geçici Olarak Kayıp Fiyat Gözlemleri ve İmpute Edilen Fiyatlar (Anonymous 2020)

Outlets	Price Reference Period							
	Dec.-19	Jan.-20	Feb.-20	Mar.-20	Apr.-20	May.-20	Jun.-20	Jul.-20
A Supermarket	5.25	5.25	5.49	5.49	5.49	5.49	5.49	5.49
B Supermarket	5.10	5.10	5.10	5.25	5.25	5.25	5.25	5.25
C Supermarket	5.20	5.20	5.20	5.20	5.20	5.25	5.25	5.25
D Independent Trader	5.49	5.49	5.49	5.65	5.75	5.80	5.80	6.00
E Independent Trader	5.99	6.50	6.50	6.90	6.90	6.90	6.90	7.00
F Independent Trader	5.99	5.99	5.99	**6.13**	**6.15**	**6.17**	6.25	6.25
Overall Mean Imputation: F Mar:May								
Geometric Mean: A:E			5.54	5.67	5.69	5.71		
Geometric Mean: A:F	5.49	5.57	5.61	5.74	5.76	5.78	5.79	5.84
Short-Term Price Relatives: A:F	1.00000	1.01371	1.00748	1.02377	1.00352	1.00365	1.00198	1.00864
Long-Term Indices as Product of Short Term	100.00	101.37	102.13	104.56	104.92	105.31	105.52	106.38
Targeted Imputation: Independent Traders								
D Independent Trader	5.49	5.49	5.49	5.65	5.75	5.80	5.80	6.00
E Independent Trader	5.99	6.50	6.50	6.90	6.90	6.90	6.90	7.00
F Independent Trader	5.99	5.99	5.99	**6.26**	**6.32**	**6.34**	6.25	6.25
Geometric Mean: A:C & D:F	5.49	5.57	5.61	5.76	5.79	5.81	5.79	5.84
Short-Term Price Relatives: A:F	1.00000	1.01371	1.00718	1.02731	1.0044	1.00377	0.99753	1.00808
Long-Term Indices as Product of Short Term	100.00	101.37	102.13	104.92	105.38	105.78	105.52	106.37
Bold: Imputed values								

Anonymous 2020 da (Table 6.3) yer alan yukarıdaki **Çizelge 5.1** tablosu, TÜFE hesaplamasını yapmak amacıyla alandan derlenen, belirlenmiş bir ürün çeşitinin işyeri satış fiyatına ait kayıp veri ve imputasyon örneğidir. Bu tabloda $X_{n\times p}$ boyutlu matris gibi değerlendirirsek; veri setindeki n satırı işyerlerini; p sütunu ilgili işyerinin belirlenmiş ürününün aylık reyon satış fiyatı değerlerini gösterir. Kalın hücreler, imputasyonla yerine değer ataması yapılmış hücrelerdir. Bu örnekte F-Bağımsız Ticaretçi (Independent Trader) işyerinden geçici olarak Mart, Nisan ve Mayıs 2020 fiyat değerleri derlenememiş kayıp veridir.

Geometrik Ortalamalar (Geometric Averages): 2004 yılında TÜFE Klavuzunun piyasaya sürülmesiyle, TÜFE temel endekslerinde bireysel fiyatlar için ağırlıklar mevcut olmadığında geometrik ortalamanın kullanılmasına büyük önem verilmiştir. Geometrik fiyat endeksi, Jevons fiyat endeksi olarak bilinir ve geometrik ortalama fiyatların oranı veya göreceli (bir malın fiyatının başka mal cinsinden verilmesi) fiyatların geometrik ortalaması olarak hesaplanır (Anonymous 2020).

$X_{n\times p}$ boyutlu belirlenmiş ürün çeşitinin işyeri satış fiyatlarının geometrik ortalaması, TÜFE hesaplamasında adı geçen ilgili ürün, madde veya hizmet çeşitlerinin birim

fiyatını temsil eder. Geometrik ortalaması hesaplanıp, ağırlıklandırılması tamamlanan birim madde fiyatları; zincirleme Laspeyres formülüyle endeks değeri (Consumer price index numbers) olarak hesaplanır. TÜFE de, geometrik ortalamaların ağırlıklandırmasıyla elde edilen t zamandaki endeks değerin, bir önceki $t-1$ zamana göre değişimine; Aylık değişim (%) (Monthly rate of change (%)) denir. Klavuzda Tablo 6.3 tablosunda aylık ve endeks değişim sırasıyla "Kısa-Dönemli Fiyat Bağıntıları: A:F" ve "Kısa-Dönemli Ürün Olarak Uzun-Dönemli Endeksler" olarak adlandırılmıştır.

Geçici Olarak Kayıp Fiyat Gözlemlerinin Klavuzdaki Tedavi Uygulamaları

Genel Ortalama İmputasyon: F Mart-Mayıs (overall mean imputation) hesaplama yönteminde: öncelikle kayıp verinin olduğu F-Bağımsız Ticaretçi verileri hariç; Şubat, Mart, Nisan ve Mayıs 2020 fiyatlarının geometrik ortalamalar hesaplanır. Sonra kayıp verileri tahmini için tam kısmın kısa dönem yani aylık değişim oranları hesaplanır: 5.67/5.54=1.02377 değeri ile F-Mart-20 tahmini hesaplanır: 5.99x1.023765=6.13. Benzer adımlar 6.13x1.003515=6.15 ve 6.15x1.003652=6.17 şeklinde kayıp veriler tamamlanır. Genel Ortalamaların değişim oranı kadar, aylık değişim ile son gözlem çarpımı şeklinde kayıp hücrelere atama yapılır. Kısa-Dönemli Fiyat Bağıntıları: A:F aylık değişim örneğin Haziran 2020 için 5.7934/5.781966=1.00198 ve Temmuz 2020 için 5.84/5.79=1.00864 olduğu açıktır. Uzun dönem endeksler yani endeks değer, Baz=100 alınarak aylık değişim oranında değişimle 100x1.01371=101.37 şeklinde hesaplanır.

Hedeflenen İmputasyon (sınıf ortalama imputasyon, class mean imputation): Bağımsız Ticaretçiler (Targeted Imputation: Independent Traders) hesaplama yöntemi benzer şekildedir. Burada genel ortalama yerine hedeflenen kendi sınıfı içerisindeki değişim ile ilgilenilir. F hariç Bağımsız Ticaretçi D ve E ye ait Şubat, Mart, Nisan ve Mayıs 2020 fiyatlarının geometrik ortalamalar hesaplanır. Kısa dönem yani aylık değişim oranları hesaplanır: 6.24/5.97=1.045216 değeri ile F-Mart-20 tahmini hesaplanır: 5.99x1.045216=6.26. Benzer adımlar 6.26x1.008811=6.32 ve 6.32x1.004338=6.34 şeklinde kayıp veriler tamamlanır. Bağımsız Ticaretçi Hedeflenen Ortalamaların değişim oranı kadar, aylık değişim ile son gözlem çarpımı yöntemiyle kayıp hücrelere

atama yapılır. Kısa-Dönemli Fiyat Bağıntıları: A:F aylık değişim örneğin Haziran 2020 için 5.794/5.808=0.9975 ve Temmuz 2020 için 5.84122/5.79439=1.008081 geometrik ortalama hesabındaki ufak dijit değişminin aylık değişimi etkilediği açıkça görülmektedir. Uzun dönem endeksler yani endeks değer, Baz=100 alınarak aylık değişim oranında değişimle benzer şeklinde hesaplanır. Mart, Nisan ve Mayıs 2020 için Kısa-Dönemli Fiyat Bağıntıları: A:F değerlerinin Genel Ortalama ve Hedeflenen İmputasyon yöntemde farklı olması, kayıp değer atamalarının değişiminden kaynaklandığı açıkça görünmektedir.

Genel ortalama ve hedeflenen imputasyon yöntemlerinindeki imputasyonlar ve imputasyon sonrası döneme geçişteki aylık değişimleri inceleyelim. Mart Nisan Mayıs kayıp veri imputasyonları ele alınırsa; hedef imputasyon değerleri, genel ortalama imputasyonlarına göre yüksektir. Haziran aylık değişimleri, genel ortalamaya göre hedeflenen imputasyon yönteminde aylık değişim düşüktür.

İleriye Taşıma İmputasyonu (Carryforward Imputation): Son gözlenen fiyatı ileriye taşımaktan kaçınılmalıdır ve bu, yalnızca sabit veya düzenlenmiş fiyatlar durumunda kabul edilebilir. Enflasyonun yüksek olduğu dönemlerde veya yüksek yenilik ve ürün devir hızı nedeniyle pazarların hızla değiştiği dönemlerde özel dikkat gösterilmelidir. Uygulaması basit olsa da, son gözlemlenen fiyatı ileriye taşımak, sonuçta ortaya çıkan endeksi sıfır değişime doğru yönlendirir. Ek olarak, kayıp çeşidin fiyatı tekrar kaydedildiğinde, endekste uygun değerine dönmek için büyük bir telafi edici adım değişikliği olması muhtemeldir. Çeşit bir süre fiyatlandırılmadan kalırsa, endeks üzerindeki olumsuz etki giderek daha şiddetli olacaktır. Tablo 6.3' teki örnekte, ileriye taşıma imputasyonu kullanılmışsa, Mart 2020' deki kayıp fiyatlar, Nisan ve Mayıs 2020' deki impute fiyatlar gibi, Şubat 2020 fiyatı olan 5,99 ile impute edilir. Haziran ayında, çeşidin dönüşünde, Mayıs' tan Haziran' a kadar fiyatta 5,99' dan 6,25' e basamaklı bir artış olacaktır. Genel olarak ileriye taşımak, bu soruna yönelik kabul edilebilir bir prosedür veya çözüm değildir (Anonymous 2020).

Çizelge 5.2 Tüketici Fiyat Endeksi Kılavuzu 2020 de Tablo 6.3 deki Geçici Olarak Kayıp Fiyat Gözlemlerinin Carryforward ile tamamlanması

Geçici Olarak Kayıp Fiyat Gözlemleri ve İmpute Edilen Fiyatlar									
Satış Noktaları		Fiyat Referans Dönemi							
		12.-19	01.-20	02.-20	03.-20	04.-20	05.-20	06.-20	07.-20
1.	İşyeri	5.25	5.25	5.49	5.49	5.49	5.49	5.49	5.49
2.	İşyeri	5.10	5.10	5.10	5.25	5.25	5.25	5.25	5.25
3.	İşyeri	5.20	5.20	5.20	5.20	5.20	5.25	5.25	5.25
4.	İşyeri	5.49	5.49	5.49	5.65	5.75	5.80	5.80	6.00
5.	İşyeri	5.99	6.50	6.50	6.90	6.90	6.90	6.90	7.00
6.	İşyeri	5.99	5.99	5.99	**5.99**	**5.99**	**5.99**	6.25	6.25
İleri Taşıma İmputasyon: 6.İşyeri 03-04-05 ay									
Geometrik Ortalama: 1.:6.		5.49	5.57	5.61	5.72	5.74	5.75	5.79	5.84
Kısa-Dönemli Fiyat Bağıntıları: 1.:6.		1.00000	1.01371	1.00748	1.01977	1.00293	1.00304	1.00711	1.00808
Kısa-Dönemli Urün Olarak Uzun-Dönemli Endeksler		100.00	101.37	102.13	104.15	104.45	104.77	105.52	106.37

Çizelge 5.2' deki ileri taşıma imputasyonu ile doldurulan veriyi, hesaplanan geometrik ortalamayı ve endekse etkisi yorumlayabilmek için aylık değişim oranlarını inceleyelim. A:E arasında herhangi geometrik ortalama hesaplaması yapılmadan yani tahminlerde herhangi bir ağırlıklandırma kullanmadan; son gözlenen değer 5.99 ile veri seti tamamlanır. Mart, Nisan ve Mayıs 2020 için artış eğiliminde olan veri setinde, ileri taşıma ile değişim olmayan atamanın; genel ortalamaya ve hedeflenen imputasyon yöntemine göre Mart, Nisan ve Mayıs 2020 aylık değişim değerini aşağı ve Haziran 2020 aylık değişim değerini yukarı çektiği açıkça görünmektedir. Anonymous 2020 de değişimin çok olduğu durumlarda önerilmeyen bir yöntem olan son gözlemi ileri taşımanın veri tamamlama öncesi ve sonrası etkileri açıkça gözlemlenmiştir. Ancak hesaplama kolaylı ve anında veri tamamlama imkanı göz ardı edilmemelidir.

Bu örneklerde sadece F Bağımsız Ticaretçi verisi kayıptı, ya kayıp daha çok olsaydı! Verilerin her zaman tam olacağı garantisi yoktur. Tablo 6.3 deki her iki yöntemde ağırlıklara göre atama yapıldığı düşünülürse; daha çok kayıp olması yani rasgele kayıp hücrelerinin daha çok olması durumunda mevcut kısımdan üretilen tahmin değerleri, mevcut verinin yapısına göre davranacaktır. Yani daha fazla kayıp hücre olması durumunda n örneklem hacmi azalması; yanlılığa sebep olabilir, standart hatayı artırabilir bu da kitlenin parametresini tahmin etme ya da kitleyi temsil etme becerisini azaltmaktadır. Peki ya anında sıcağı sıcağına atama yapan, hem de mevcut verinin

tamamlanıp imputasyonunun hesaplanmasını beklemeye gerek kalmayan bir yöntem olsaydı!

TÜFE uygulamasında kullanılan, gözlenen son fiyatı ileriye taşıma imputasyonu (carryforward) metedolojide karşılaşılan Hot Deck yönteminin bir genellemesi olduğu önceki bölümlerde işlenmişti. Anında veri derleme amacıyla hızlı bir imputasyon için alternatif olarak önerilen birinci yöntem; bir önceki yılın aylık değişim oranına bağlı ağırlıklandırılmış imputasyon yöntemi olsun. Burada amaç; alandan veri derlenip, endeks hesaplayıcılarına anında, sıcağı sıcağına veriyi göndermektir. Bu sayede veri kalitesi için analizlerin anında yapılıp geri dönüşünün hızlıca sağlanması, anlık karşılaştırma imkanı sayesinde işyerinde reyon etiket güncellemesi ihmal edilen kayıt tespiti, anlık tahminler alınması vb. günümüz koşullarına ayak uyduran hızlı ihtiyaçları karşılanması amaçlanmaktadır. Veri toplama aracı mobil veri üzerinden anında veri akışı olacak şekilde günümüzde teknolojik ekipmanlar ve imkanlar istististik ofislerinde mevcuttur. Geçici kayıp fiyat durumunda gözlenen son fiyatı ileriye taşıma imputasyon yöntemi hızlı veri akışı için zaman kısıtı altında, analiz gerektirmeyen, kayıp hücreyi anında dolduran bir yöntem olduğu aşikardır. Ancak bu yöntem değişimin çok olduğu dönemlerde geometrik ortalamayı ya aşağı çekecek ya da yukarı çıkaracaktır. Enflasyon değişimlerinin belirgin şekilde arttığı Covid-19 sonrası günümüz dünyasında, özellikle Coicop gıda sınıflamasına özel olarak bir önceki yılın, aynı ayında aylık değişim oranına bağlı ağırlıklandırılmış ve işaretlenmiş, anında veri derleme aracında hesaplama ve atama yapılabilecek yöntem aşağıdaki gibi tanımlamıştır. Burada $x_{t,miss}$: t zamandaki kayıp hücre, $x_{t-1,last}$: $(t-1)$ zamandaki son gözlem, mr_{t-12}: bir önceki yılın aylık oran değişimini (monthly rate of change) gösterir ve **Çizelge 5.3**‘ den elde edilir. Yöntem adı bir önceki yılın aynı ayına ait bilgiyi kullandığından aylık değişim imputasyonu (imputation monthly) anlamına gelen i_mon olarak isimlendirilmiş olup, analizlerde bazen kolaylık sağladığı için $yöntem1$ olarak da isimlendirilmiştir.

$i\_mon: yöntem1$: $x_{t,miss} = x_{t-1,last} + (x_{t-1,last} * mr_{t-12})$

Buradaki temel düşünce TÜFE nin zaman serisi özelliğini kullanmaktır. Aynı ürün takibi varsayımı altında, ürünlerin yılın belli dönemlerinde aynı fiyat değişim davranışında bulunduğunu düşünebiliriz. Örneğin, geçen sene kasım ayında domates ürününün fiyatındaki değişimin bu sene kasım ayında da aynı yönlü olması beklenir. Artış ya da azalış eğilimini bir önceki senenin aylık değişimine bakabileceğimiz, elimizde oransal değişim oranları mevcuttur. Aynı ürün ve aynı zaman periyodunda aynı değişim eğilimi varsayımı altında; bu değişimlerin işaretleri, bize tahmini artış ya da düşüş yönünde işaretleme; bir önceki senenin aylık değişimin oransal değeri de ağırlıklandırma verilmesi imkanı sağlar. İstatistik ofislerinin alandan veri derleme araçlarında yazılım güncellemesine eklenecek bir komut ile "son gözlemi ileri taşıma" imputasyonuna bu çarpımsal ağırlıklandırmayı tanımlaması yeterli olacaktır. Anında sıcağı sıcağına geçici olan kayıp verinin imputasyonu ağırlığı ile birlikte tamamlanıp, mobil veri aracılığıyla endeks hesaplayıcısına hızlıca ulaşacaktır.

Çizelge 5.3 G20 TÜFE -Yirmili Grup- (Dataset Monthly rate of change) (https://ec.europa.eu/eurostat, 2023)

Data extracted on 31/01/2023 13:34:36 from [ESTAT]													
Dataset:	G20 CPI all-items - Group of Twenty - Consumer price index [PRC_IPC_G20__custom_4733163]												
Last updated:	18/01/2023 11:00												
Time frequency		Monthly											
Unit of measure		Monthly rate of change											
Til	YIL /	Oca.	Şub.	Mar.	Nis.	May.	Haz.	Tem.	Ağu.	Eyl.	Eki.	Kas.	Ara.
European Union - 27 countries (from	2019	-0.8	0.3	0.9	0.7	0.2	0.1	-0.3	0.1	0.2	0.2	-0.2	0.3
European Union - 27 countries (from	2020	-0.7	0.2	0.5	0.2	0.	0.3	-0.2	-0.3	0.	0.2	-0.3	0.3
Argentina	2019	2.9	3.8	4.7	3.4	3.1	2.7	2.2	4.	5.9	3.3	4.3	3.7
Argentina	2020	2.	3.3	1.5	1.5	2.2	1.9	2.7	2.8	3.8	3.2	4.	
Brazil	2019	0.3	0.4	0.7	0.6	0.1	0.	0.2	0.1	0.	0.1	0.5	1.2
Brazil	2020	0.2	0.3	0.1	0.	-0.1	0.3	0.4	0.2	0.6	0.9	0.9	1.4
Canada	2019	0.1	0.7	0.7	0.4	0.4	0.	0.5	0.	0.	0.3	0.	0.
Canada	2020	0.3	0.4	-0.1	-0.1	0.3	0.8	0.	0.	-0.1	0.4	0.1	0.
China including Hong Kong	2019	0.5	1.	-0.4	0.1	0.	-0.1	0.4	0.7	0.9	0.9	0.4	0.
China including Hong Kong	2020	1.4	0.8	-1.2	-0.9	-0.8	-0.1	0.6	0.4	0.2	-0.3	-0.6	0.7
France	2019	-0.6	0.1	0.9	0.4	0.1	0.3	-0.2	0.5	-0.4	-0.1	0.1	0.5
France	2020	-0.5	0.	0.1	0.	0.2	0.1	0.4	-0.1	-0.6	0.	0.2	0.2
Germany (until 1990 former territory	2019	-1.	0.5	0.5	1.	0.3	0.3	0.4	-0.1	-0.1	0.1	-0.8	0.6
Germany (until 1990 former territory	2020	-0.8	0.6	0.1	0.4	0.	0.7	-0.5	-0.2	-0.4	0.	-1.	0.6
Group of Twenty	2019	0.3	0.4	0.5	0.6	0.3	0.1	0.3	0.3	0.4	0.5	0.2	0.3
Group of Twenty	2020	0.3	0.3	-0.1	0.	0.	0.4	0.4	0.4	0.2	0.3	0.	0.3
India	2019	2.	0.	0.7	1.	0.6	0.6	0.9	0.3	0.6	0.9	0.9	0.6
India	2020	0.	0.	0.	0.9	0.3	0.6	1.2	0.6	0.6	1.2	0.3	0.
Indonesia	2019	0.3	-0.1	0.1	0.4	0.7	0.6	0.3	0.1	-0.1	0.	0.1	0.3
Indonesia	2020	0.3	0.3	0.1	0.1	0.1	0.2	0.	0.	0.	0.1	0.3	0.4
Italy	2019	-1.7	-0.3	2.3	0.5	0.1	0.1	-1.8	0.	1.4	0.2	-0.3	0.2
Italy	2020	-1.8	-0.5	2.2	0.5	-0.3	0.	-0.7	-1.3	0.9	0.6	0.	0.2
Japan	2019	0.1	0.	0.	0.3	0.	-0.1	-0.1	0.3	0.1	0.3	0.1	0.
Japan	2020	-0.1	-0.2	0.	-0.1	0.	-0.2	0.1	0.1	-0.2	-0.1	-0.3	-0.2
Mexico	2019	0.1	0.	0.4	0.1	-0.1	0.1	0.4	0.	0.3	0.5	0.8	0.6
Mexico	2020	0.5	0.4	0.	-0.1	0.4	0.5	0.7	0.4	0.2	0.6	0.1	0.4
Russia	2019	1.	0.4	0.3	0.3	0.3	0.	0.2	0.	-0.1	0.1	0.3	0.4
Russia	2020	0.4	0.3	0.6	0.8	0.3	0.2	0.4	0.	-0.1	0.4	0.7	0.8
Saudi Arabia	2019	-0.2	-0.1	0.	0.1	0.	0.2	0.3	0.2	0.2	0.1	0.	0.2
Saudi Arabia	2020	-0.1	0.3	0.1	0.	0.	0.	5.9	0.2	-0.1	0.1	-0.1	0.
South Africa	2019	0.	0.8	0.8	0.6	0.3	0.3	0.3	0.2	0.3	0.	0.1	0.2
South Africa	2020	0.2	1.	0.4	0.	0.	0.4	1.4	0.1	0.1	0.4	0.	0.1
South Korea	2019	0.	0.4	-0.1	0.4	0.2	-0.1	0.	0.2	0.4	0.2	-0.1	0.2
South Korea	2020	0.4	0.1	0.	0.	-0.1	0.3	-0.1	0.6	0.5	-0.1	-0.1	0.2
Türkiye	2019	1.1	0.2	1.	1.7	0.9	0.	1.4	0.9	1.	2.	0.4	0.7
Türkiye	2020	1.4	0.4	0.6	0.9	1.4	1.1	0.6	0.9	1.	2.1	2.3	1.3
United Kingdom	2019	-0.8	0.5	0.2	0.6	0.3	0.	0.	0.4	0.1	-0.2	0.2	0.
United Kingdom	2020	-0.3	0.4	0.	-0.2	0.	0.1	0.4	-0.4	0.4	0.	-0.1	0.3
United States	2019	0.2	0.4	0.6	0.5	0.2	0.	0.2	0.	0.1	0.2	-0.1	-0.1
United States	2020	0.4	0.3	0.	-0.1	0.	0.5	0.5	0.3	0.1	0.	-0.1	0.1

$yöntem1(i\_mon)$ uygulamasını incelemek için Eurostat sayfasında yer alan G20 TÜFE -Yirmili Grup- Tüketici fiyat endeksi 2019 ve 2020 için oluşturulup dışarı aktarılan (export) veri setindeki aylık değişim oranı verilerinden Meksika örneğini ele alalım. Mart, Nisan ve Mayıs 2019 daki aylık değişim oranları sırasıyla 0.4, 0.1 ve -0.1 değerlerini, ağırlıklandırma ölçütü olarak kullanıp, klavuzdaki imputasyon hesaplamaları gibi aşağıdaki **Çizelge 5.4** de ağırlıklandırılarak, Tablo 6.3' deki F-Bağımsız ticaretçisinin Mart, Nisan ve Mayıs 2020 için kayıp veri tahmini tamamlanabilir.

Çizelge 5.4 Tüketici Fiyat Endeksi Kılavuzu 2020 de Tablo 6.3 deki Geçici Olarak Kayıp Fiyat Gözlemlerinin $yöntem1$: i_mon ile tamamlanması

Geçici Olarak Kayıp Fiyat Gözlemleri ve İmpute Edilen Fiyatlar									
Satış Noktaları		Fiyat Referans Dönemi							
		12.-19	01.-20	02.-20	03.-20	04.-20	05.-20	06.-20	07.-20
1.	İşyeri	5.25	5.25	5.49	5.49	5.49	5.49	5.49	5.49
2.	İşyeri	5.10	5.10	5.10	5.25	5.25	5.25	5.25	5.25
3.	İşyeri	5.20	5.20	5.20	5.20	5.20	5.25	5.25	5.25
4.	İşyeri	5.49	5.49	5.49	5.65	5.75	5.80	5.80	6.00
5.	İşyeri	5.99	6.50	6.50	6.90	6.90	6.90	6.90	7.00
6.	İşyeri	5.99	5.99	5.99	**8.39**	**9.22**	**8.30**	6.25	6.25
Yöntem1: *i_mon* Ağırlıklı İleri Taşıma İmputasyon: 6.İşyeri 03-04-05 ay									
Mexico 2019 Monthly rate of change					0.4	0.1	-0.1		
Geometrik Ortalama: 1.:6.		5.49	5.57	5.61	6.05	6.16	6.08	5.79	5.84
Kısa-Dönemli Fiyat Bağıntıları: 1.:6.		1.00000	1.01371	1.00748	1.07859	1.01899	0.98558	0.95378	1.00808
Kısa-Dönemli Ürün Olarak Uzun-Dönemli Endeksler		100.00	101.37	102.13	110.16	112.25	110.63	105.52	106.37

$yöntem1(i\_mon)$ imputasyon sonrasında, F- Bağımsız ticaretçisi için Mart, Nisan ve Mayıs 2020 kayıp veri imputasyonları; genel ortalamaya ve hedeflenen imputasyon yöntemlere göre yüksek, Haziran aylık değişimi; genel ortalamaya ve hedeflenen imputasyon yöntemine göre düşük çıktığı açıkça görünmektedir.

Benzer şekilde; alternatif olarak önerilen ikinci yöntem; bir önceki yılın aylık değişim oranına bağlı işaretlenmiş ve altın oran ile ağırlıklandırılmış imputasyon yöntemi olsun. Coicop gıda sınıflamasına özel olarak altın oran ile ağırlıklandırılmış ve işaretlenmiş, anında veri derleme aracında hesaplama ve atama yapılabilecek yöntem aşağıdaki gibi tanımlamıştır. Burada $x_{t,miss}$: t zamandaki kayıp hücre, $x_{t-1,last}$: $(t-1)$ zamandaki son gözlem, işaret fonksiyonu sgn olmak üzere $sgn(mr_{t-12})$: bir önceki yılın aylık

oran değişiminin işaretini gösterir ve **Çizelge 5.3**' den elde edilir. Altın oran $\varphi = 1.618033$ olarak alalım. Yöntem adı altın oran değerine ait bilgiyi kullandığından altın oran imputasyonu (imputation φ) anlamına gelen i_φ olarak isimlendirilmiş olup, analizlerde bazen kolaylık sağladığı için $yöntem2$ olarak da isimlendirilmiştir

$$i\_\varphi: yöntem2: x_{t,miss} = x_{t-1,last} + (sgn(mr_{t-12}) * \varphi)$$

Buradaki temel düşünce ise evrendeki altın oran kusursuzluğunun, kayıp veri durumunda etkisini araştırmaktır. Benzer şekilde artış ya da azalış eğilimini bir önceki senenin aylık değişimine bakabileceğimiz elimizde oransal değişimler mevcuttur. Bu değişimlerin işaretleri, bize tahmini artış ya da düşüş yönünde işaretleme; altın oran değeri de ağırlıklandırma verilmesi imkanı sağlar.

$yöntem2(i\_\varphi)$ uygulamasını incelemek için Eurostat sayfasında yer alan G20 TÜFE - Yirmili Grup- Tüketici fiyat endeksi 2019 ve 2020 için oluşturulup dışarı aktarılan veri setindeki aylık değişim oranı verilerinden Meksika örneğini ele alalım. Burada Mart, Nisan ve Mayıs 2019 daki aylık değişim oranları sırasıyla 0.4, 0.1 ve -0.1 değerlerinin pozitif/negatif işaretlerini, φ ağırlıklandırma ölçütü ile kullanıp, klavuzdaki imputasyon hesaplamaları gibi aşağıdaki **Çizelge 5.5** de ağırlıklandırılarak, Tablo 6.3 deki F-Bağımsız ticaretçisinin Mart, Nisan ve Mayıs 2020 kayıp veri tahmini tamamlanabilir.

Çizelge 5.5 Tüketici Fiyat Endeksi Kılavuzu 2020 de Tablo 6.3 deki Geçici Olarak Kayıp Fiyat Gözlemlerinin $yöntem2$: i_φ ile tamamlanması

Geçici Olarak Kayıp Fiyat Gözlemleri ve İmpute Edilen Fiyatlar									
Satış Noktaları		Fiyat Referans Dönemi							
		12.-19	01.-20	02.-20	03.-20	04.-20	05.-20	06.-20	07.-20
1.	İşyeri	5.25	5.25	5.49	5.49	5.49	5.49	5.49	5.49
2.	İşyeri	5.10	5.10	5.10	5.25	5.25	5.25	5.25	5.25
3.	İşyeri	5.20	5.20	5.20	5.20	5.20	5.25	5.25	5.25
4.	İşyeri	5.49	5.49	5.49	5.65	5.75	5.80	5.80	6.00
5.	İşyeri	5.99	6.50	6.50	6.90	6.90	6.90	6.90	7.00
6.	İşyeri	5.99	5.99	5.99	**7.61**	**9.23**	**7.61**	6.25	6.25
Yöntem2: i_φ Ağırlıklı İleri Taşıma İmputasyon: 6.İşyeri 03-04-05 ay									
Mexico 2019 Monthly rate of change					0.4	0.1	-0.1		
Geometrik Ortalama: 1.:6.		5.49	5.57	5.61	5.95	6.16	5.99	5.79	5.84
Kısa-Dönemli Fiyat Bağıntıları: 1.:6.		1.00000	1.01371	1.00748	1.06123	1.03568	0.97132	0.96776	1.00808
Kısa-Dönemli Ürün Olarak Uzun-Dönemli Endeksler		100.00	101.37	102.13	108.38	112.25	109.03	105.52	106.37

$y\ddot{o}ntem2(i\_\varphi)$ imputasyon sonrasında, F- Bağımsız ticaretçisi için Mart, Nisan ve Mayıs 2020 kayıp veri imputasyonları; genel ortalamaya ve hedeflenen imputasyon yöntemlere göre yüksek, Haziran aylık değişimi; genel ortalamaya ve hedeflenen imputasyon yöntemine göre düşük çıktığı açıkça görünmektedir.

Üçüncü olarak önerilen $y\ddot{o}ntem3(i\_m\ddot{u}d19)$; bir önceki yılın aylık değişim oranına bağlı işaretlenmiş ve $m\ddot{u}d19$ ile ağırlıklandırılmış imputasyon yöntemi olsun. Coicop gıda sınıflamasına özel olarak ağırlıklandırılmış ve işaretlenmiş, anında veri derleme aracında hesaplama ve atama yapılabilecek yöntem aşağıdaki gibi tanımlamıştır. Burada $x_{t,miss}$: t zamandaki kayıp hücre, $x_{t-1,last}$: $(t-1)$ zamandaki son gözlem, işaret fonksiyonu sgn olmak üzere $sgn(mr_{t-12})$: bir önceki yılın aylık oran değişiminin işaretini gösterir ve **Çizelge 5.3**'den elde edilir. Tanımlanan ağırlık $m\ddot{u}d19 = 0.193074$ olarak belirlenmiştir. Yöntem adı fiyat küsuratını temsil eden belirlenmiş dijit değere ait bilgiyi kullandığından 0.193074 luk değişim imputasyonu (imputation $m\ddot{u}d19 = 0.193074$) anlamına gelen $i\_m\ddot{u}d19$ olarak isimlendirilmiş olup, analizlerde bazen kolaylık sağladığı için $y\ddot{o}ntem3$ olarak da isimlendirilmiştir.

$i\_m\ddot{u}d19$: $y\ddot{o}ntem3: x_{t,miss} = x_{t-1,last} + (sgn(mr_{t-12}) * m\ddot{u}d19)$

Buradaki temel düşünce yüksek altın oran ağırlığı yerine para birimi küsuratına çok etken olmayacak şekilde ve ileriye taşıma imputasyondaki durağanlığı da biraz kıpırtıtadacak kadar bir ağırlıklandırmadır. Benzer şekilde artış ya da azalış eğilimi için bir önceki senenin aylık değişimine bakabileceğimiz elimizde oransal değişimler mevcuttur. Bu değişimlerin işaretleri, bize tahmini artış ya da düşüş yönünde işaretleme; $m\ddot{u}d19$ değeri de altın orana göre daha az bir ağırlıklandırma imkanı sağlar. Ayrıca başka ağırlıklandırma ile imputasyon çalışmaları için de ağırlıklandırma oranı olarak $m\ddot{u}d19$ değeri tavsiye edilebilir.

$y\ddot{o}ntem3(i\_m\ddot{u}d19)$ uygulamasını incelemek için benzer şekilde Eurostat sayfasında yer alan G20 TÜFE -Yirmili Grup- Tüketici fiyat endeksi 2019 ve 2020 için oluşturulup dışarı aktarılan veri setindeki aylık değişim oranı verilerinden Meksika örneğini ele

alalım. Burada Mart, Nisan ve Mayıs 2019 daki aylık değişim oranları sırasıyla 0.4, 0.1 ve -0.1 değerlerinin pozitif/negatif işaretlerini, $müd19$ ağırlıklandırma ölçütü ile kullanıp, klavuzdaki imputasyon hesaplamaları gibi aşağıdaki **Çizelge 5.6** da ağırlıklandırılarak, Tablo 6.3' deki F-Bağımsız ticaretçisinin Mart, Nisan ve Mayıs 2020 kayıp veri tahmini tamamlanabilir.

Çizelge 5.6 Tüketici Fiyat Endeksi Kılavuzu 2020 de Tablo 6.3 deki Geçici Olarak Kayıp Fiyat Gözlemlerinin $yöntem3(i\_müd19)$ ile tamamlanması

Geçici Olarak Kayıp Fiyat Gözlemleri ve İmpute Edilen Fiyatlar									
Satış Noktaları		Fiyat Referans Dönemi							
		12.-19	01.-20	02.-20	03.-20	04.-20	05.-20	06.-20	07.-20
1.	İşyeri	5.25	5.25	5.49	5.49	5.49	5.49	5.49	5.49
2.	İşyeri	5.10	5.10	5.10	5.25	5.25	5.25	5.25	5.25
3.	İşyeri	5.20	5.20	5.20	5.20	5.20	5.25	5.25	5.25
4.	İşyeri	5.49	5.49	5.49	5.65	5.75	5.80	5.80	6.00
5.	İşyeri	5.99	6.50	6.50	6.90	6.90	6.90	6.90	7.00
6.	İşyeri	5.99	5.99	5.99	**6.183074**	**6.376148**	**6.183074**	6.25	6.25
Yöntem3: *i_müd19* Ağırlıklı İleri Taşıma İmputasyon: 6.İşyeri 03-04-05 ay									
Mexico 2019 Monthly rate of change					0.4	0.1	-0.1		
Geometrik Ortalama: 1.:6.		5.49	5.57	5.61	5.75	5.80	5.78	5.79	5.84
Kısa-Dönemli Fiyat Bağıntıları: 1.:6.		1.00000	1.01371	1.00748	1.02517	1.00808	0.99792	1.00180	1.00808
Kısa-Dönemli Ürün Olarak Uzun-Dönemli Endeksler		100.00	101.37	102.13	104.70	105.55	105.33	105.52	106.37

$yöntem3(i\_müd19)$ imputasyon sonrasında, F- Bağımsız ticaretçisi için Mart, Nisan ve Mayıs 2020 kayıp veri imputasyonları; genel ortalama imputasyon yönteme göre yüksek, hedeflenen imputasyon yöntem civarlarında değerler, Haziran aylık değişimi; genel ortalamaya ve hedeflenen imputasyon yöntemine göre düşük çıktığı açıkça görünmektedir.

Son olarak istatistiksel programlama dilinde sağlam aykırı değer atama ve kayıp veri atama yöntemi ile bu tablodaki kayıp hücreleri tamamlayarak değişim oranlarını ele alalım. İstatistiksel programlamalardan R ile yapılacak işlemlerde panel veri yapısı sebebiyle Tablo 6.3 deki bilgiler, $isyerleri \times aylar$ veri yapıları şeklinde yapılacak işlemler için aşağıdaki **Çizelge 5.7** gibi dönüştürülmüştür (R Core Team, 2021). F-Bağımsız ticaretçisi için Mart Nisan Mayıs 2020' deki kayıp hücreleri, veri matrisinde sırasıyla 44., 45. ve 46. hücreye eşittir.

Çizelge 5.7 Tüketici Fiyat Endeksi Kılavuzu 2020 de Tablo 6.3 deki bilgilerin $isyerleri \times aylar$ şeklinde panel veri giriş düzenlemesi

			current methods			proposed methods					R programming language imputation				
	isyerleri	aylar	overall	target	carry	method1	method2	method3	method3_1	method3_2	R_miss	ImpReg	ImpMI	R_outlier	ImpImrob
1	1	12	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25
2	1	1	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25
3	1	2	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49
4	1	3	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49
5	1	4	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49
6	1	5	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49
7	1	6	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49
8	1	7	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49
9	2	12	5.10	5.10	5.10	5.10	5.10	5.10	5.10	5.10	5.10	5.10	5.10	5.10	5.10
10	2	1	5.10	5.10	5.10	5.10	5.10	5.10	5.10	5.10	5.10	5.10	5.10	5.10	5.10
11	2	2	5.10	5.10	5.10	5.10	5.10	5.10	5.10	5.10	5.10	5.10	5.10	5.10	5.10
12	2	3	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25
13	2	4	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25
14	2	5	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25
15	2	6	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25
16	2	7	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25
17	3	12	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20
18	3	1	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20
19	3	2	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20
20	3	3	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20
21	3	4	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20
22	3	5	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25
23	3	6	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25
24	3	7	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25
25	4	12	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49
26	4	1	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49
27	4	2	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49
28	4	3	5.65	5.65	5.65	5.65	5.65	5.65	5.65	5.65	5.65	5.65	5.65	5.65	5.65
29	4	4	5.75	5.75	5.75	5.75	5.75	5.75	5.75	5.75	5.75	5.75	5.75	5.75	5.75
30	4	5	5.80	5.80	5.80	5.80	5.80	5.80	5.80	5.80	5.80	5.80	5.80	5.80	5.80
31	4	6	5.80	5.80	5.80	5.80	5.80	5.80	5.80	5.80	5.80	5.80	5.80	5.80	5.80
32	4	7	6.00	6.00	6.00	6.00	6.00	6.00	6.00	6.00	6.00	6.00	6.00	6.00	6.00
33	5	12	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99
34	5	1	6.50	6.50	6.50	6.50	6.50	6.50	6.50	6.50	6.50	6.50	6.50	6.50	6.50
35	5	2	6.50	6.50	6.50	6.50	6.50	6.50	6.50	6.50	6.50	6.50	6.50	6.50	6.50
36	5	3	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90
37	5	4	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90
38	5	5	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90
39	5	6	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90
40	5	7	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00
41	6	12	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99
42	6	1	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99
43	6	2	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99
44	**6**	**3**	**6.13**	**6.26**	**5.99**	**8.39**	**7.61**	**6.183074**	**6.105651326**	**6.105651326**	***NA***	**6.1908**	**5.99**	**0.00**	**6.320582**
45	**6**	**4**	**6.15**	**6.32**	**5.99**	**9.22**	**9.23**	**6.376148**	**6.223535578**	**6.223535578**	***NA***	**6.2108**	**6.25**	**0.00**	**6.306284**
46	**6**	**5**	**6.17**	**6.34**	**5.99**	**8.30**	**7.61**	**6.183074**	**6.103375288**	**6.34369528**	***NA***	**6.2308**	**6.50**	**0.00**	**6.291986**
47	6	6	6.25	6.25	6.25	6.25	6.25	6.25	6.25	6.25	6.25	6.25	6.25	6.25	6.25
48	6	7	6.25	6.25	6.25	6.25	6.25	6.25	6.25	6.25	6.25	6.25	6.25	6.25	6.25

Bold: Imputed values

İstatistiksel programlamalardan R da mice komutu ile imputasyon aşağıdaki gibi yapılmıştır (**bkz. Ek 1**) (R Core Team, 2021). Dezavantajı, tüm verinin tamamlanmasının beklenmesidir.

```
> multiple_imputation[["imp"]]
$R_Mice_Impute
       1    2    3    4    5    6    7    8    9   10
44 5.99 5.99 6.25 5.99 6.25 5.99 5.99 5.99 5.99 5.99
45 5.99 6.25 5.99 5.99 5.99 6.25 5.99 5.99 5.99 5.99
```

46 6.00 6.50 6.25 5.99 5.99 6.25 6.50 6.25 6.25 5.99

İstatistiksel programlamalardan R deki mice (MI) komut uygulamasını incelemek için; Tablo 6.3 deki Mart, Nisan ve Mayıs 2020 kayıp değerlerinin veri tahminini ve aylık değişimini aşağıdaki **Çizelge 5.8** deki gibi rasgele olarak 2. iterasyon değer sonuçları ile tamamlayalım (R Core Team, 2021).

Çizelge 5.8 Tüketici Fiyat Endeksi Kılavuzu 2020 de Tablo 6.3 deki Geçici Olarak Kayıp Fiyat Gözlemlerinin R_Mice_Impute MI ile tamamlanması

Geçici Olarak Kayıp Fiyat Gözlemleri ve İmpute Edilen Fiyatlar

Satış Noktaları	**Fiyat Referans Dönemi**							
	12.-19	**01.-20**	**02.-20**	**03.-20**	**04.-20**	**05.-20**	**06.-20**	**07.-20**
1. İşyeri	5.25	5.25	5.49	5.49	5.49	5.49	5.49	5.49
2. İşyeri	5.10	5.10	5.10	5.25	5.25	5.25	5.25	5.25
3. İşyeri	5.20	5.20	5.20	5.20	5.20	5.25	5.25	5.25
4. İşyeri	5.49	5.49	5.49	5.65	5.75	5.80	5.80	6.00
5. İşyeri	5.99	6.50	6.50	6.90	6.90	6.90	6.90	7.00
6. İşyeri	5.99	5.99	5.99	**5.99**	**6.25**	**6.50**	6.25	6.25
***Paket Program* İmputasyon (R_Mice_Impute MI): 6.İşyeri 03-04-05 ay**								
Geometrik Ortalama: 1.:6.	5.49	5.57	5.61	5.90	6.07	5.93	5.79	5.84
Kısa-Dönemli Fiyat Bağıntıları: 1.:6.	1.00000	1.01371	1.00748	1.01977	1.01006	1.00962	0.99348	1.00808
Kısa-Dönemli Ürün Olarak Uzun-Dönemli Endeksler	100.00	101.37	102.13	104.15	105.20	106.21	105.52	106.37

Burada F- Bağımsız ticaretçisi için Mart, Nisan ve Mayıs 2020 kayıp veri imputasyonları; genel ortalama ve hedeflenen imputasyon yöntem civarlarında değerler, Haziran aylık değişimi; genel ortalamaya ve hedeflenen imputasyon yöntemine göre düşük çıktığı açıkça görünmektedir.

Ayrıca, istatistiksel programlamalardan R da mice (method=norm.predict) komutu ile kayıp veri regresyon imputasyon aşağıdaki gibi yapılmıştır (**bkz. Ek 2**) (R Core Team, 2021). Dezavantajı tüm verinin tamamlanmasının beklenmesidir.

> RegMiss_imputed[["imp"]]

$R_Mice_Impute

	1	2	3	4	5	6	7	8	9	10
44	6.1908	6.1908	6.1908	6.1908	6.1908	6.1908	6.1908	6.1908	6.1908	6.1908
45	6.2108	6.2108	6.2108	6.2108	6.2108	6.2108	6.2108	6.2108	6.2108	6.2108
46	6.2308	6.2308	6.2308	6.2308	6.2308	6.2308	6.2308	6.2308	6.2308	6.2308

İstatistiksel programlamalardan R deki mice (reg) komut uygulamasını incelemek için; Tablo 6.3' deki Mart, Nisan ve Mayıs 2020 kayıp değerlerinin veri tahminini ve aylık değişimini aşağıdaki **Çizelge 5.9** daki gibi hesaplanan Mice_Impute değer sonuçları ile tamamlayalım (R Core Team, 2021).

Çizelge 5.9 Tüketici Fiyat Endeksi Kılavuzu 2020 de Tablo 6.3 deki Geçici Olarak Kayıp Fiyat Gözlemlerinin R_Mice_Impute Reg ile tamamlanması

Geçici Olarak Kayıp Fiyat Gözlemleri ve İmpute Edilen Fiyatlar									
Satış Noktaları		**Fiyat Referans Dönemi**							
		12.-19	**01.-20**	**02.-20**	**03.-20**	**04.-20**	**05.-20**	**06.-20**	**07.-20**
1.	**İşyeri**	5.25	5.25	5.49	5.49	5.49	5.49	5.49	5.49
2.	**İşyeri**	5.10	5.10	5.10	5.25	5.25	5.25	5.25	5.25
3.	**İşyeri**	5.20	5.20	5.20	5.20	5.20	5.25	5.25	5.25
4.	**İşyeri**	5.49	5.49	5.49	5.65	5.75	5.80	5.80	6.00
5.	**İşyeri**	5.99	6.50	6.50	6.90	6.90	6.90	6.90	7.00
6.	**İşyeri**	5.99	5.99	5.99	**6.1908**	**6.2108**	**6.2308**	6.25	6.25
***Paket Program* İmputasyon (R_Mice_Impute Reg): 6.İşyeri 03-04-05 ay**									
Geometrik Ortalama: 1.:6.		5.49	5.57	5.61	5.90	6.07	5.93	5.79	5.84
Kısa-Dönemli Fiyat Bağıntıları: 1.:6.		1.00000	1.01371	1.00748	1.02538	1.00347	1.00358	1.00051	1.00808
Kısa-Dönemli Ürün Olarak Uzun-Dönemli Endeksler		100.00	101.37	102.13	104.72	105.08	105.46	105.52	106.37

Burada F- Bağımsız ticaretçisi için Mart, Nisan ve Mayıs 2020 kayıp veri imputasyonları; genel ortalama imputasyon yönteme göre yüksek, hedeflenen imputasyon yönteme göre düşük değerler, Haziran aylık değişimi; genel ortalamaya göre düşük ve hedeflenen imputasyon yöntemine göre yüksek çıktığı açıkça görünmektedir.

Son olarak, istatistiksel programlamalardan R da lmrob komutu ile imputasyon aşağıdaki gibi yapılmıştır (**bkz. Ek 3**) (R Core Team, 2021). Dezavantajı, tüm verinin tamamlanmasının beklenmesidir.

Çizelge 5.10 R programı augment (fit1) ile 44., 45. ve 46. hücre fitted değeri sonuçları

	R_outlier	isyerleri	aylar	.fitted	.resid
44	0.00	6	3	6.320582	-6.32058163
45	0.00	6	4	6.306284	-6.30628392
46	0.00	6	5	6.291986	-6.29198620

İstatistiksel programlamalardan R daki lmrob komut uygulamasını incelemek için; Tablo 6.3 deki Mart, Nisan ve Mayıs 2020 kayıp değerlerinin veri tahminini ve aylık değişimini aşağıdaki **Çizelge 5.11** deki gibi hesaplanan Impute_Lmrob değer sonuçları ile tamamlayalım (R Core Team, 2021). Kayıp veriyi aykırı değer gibi değerlendirip sadece hücresel aykırı değer durumunda uygulanan sağlam kestirim yöntemi "Impute_Lmrob" ile sağlam aykırı değer imputasyonu yapılmıştır.

Çizelge 5.11 Tüketici Fiyat Endeksi Kılavuzu 2020 de Tablo 6.3 deki Geçici Olarak Kayıp Fiyat Gözlemlerinin R_outlier Impute lmrob ile tamamlanması

Geçici Olarak Kayıp Fiyat Gözlemleri ve İmpute Edilen Fiyatlar								
Satış Noktaları		**Fiyat Referans Dönemi**						
		12.-19	**01.-20**	**02.-20**	**03.-20**	**04.-20**	**05.-20**	**06.-20**	**07.-20**
1.	**İşyeri**	5.25	5.25	5.49	5.49	5.49	5.49	5.49	5.49
2.	**İşyeri**	5.10	5.10	5.10	5.25	5.25	5.25	5.25	5.25
3.	**İşyeri**	5.20	5.20	5.20	5.20	5.20	5.25	5.25	5.25
4.	**İşyeri**	5.49	5.49	5.49	5.65	5.75	5.80	5.80	6.00
5.	**İşyeri**	5.99	6.50	6.50	6.90	6.90	6.90	6.90	7.00
6.	**İşyeri**	5.99	5.99	5.99	**0.00**	**0.00**	**0.00**	6.25	6.25
***Paket Program* İmputasyon ("0" ların outlier olarak ele alınması): 6.İşyeri 03-04-05 ay**									
Satış Noktaları		**Fiyat Referans Dönemi**						
		12.-19	**01.-20**	**02.-20**	**03.-20**	**04.-20**	**05.-20**	**06.-20**	**07.-20**
1.	**İşyeri**	5.25	5.25	5.49	5.49	5.49	5.49	5.49	5.49
2.	**İşyeri**	5.10	5.10	5.10	5.25	5.25	5.25	5.25	5.25
3.	**İşyeri**	5.20	5.20	5.20	5.20	5.20	5.25	5.25	5.25
4.	**İşyeri**	5.49	5.49	5.49	5.65	5.75	5.80	5.80	6.00
5.	**İşyeri**	5.99	6.50	6.50	6.90	6.90	6.90	6.90	7.00
6.	**İşyeri**	5.99	5.99	5.99	**6.320582**	**6.306284**	**6.291986**	6.25	6.25
***Paket Program* İmputasyon (R_outlier Impute lmrob): 6.İşyeri 03-04-05 ay**									
Geometrik Ortalama: 1.:6.		5.49	5.57	5.61	5.77	5.79	5.80	5.79	5.84
Kısa-Dönemli Fiyat Bağıntıları: 1.:6.		1.00000	1.01371	1.00748	1.02894	1.00255	1.00266	0.99888	1.00808
Kısa-Dönemli Ürün Olarak Uzun-Dönemli Endeksler		100.00	101.37	102.13	105.08	105.35	105.63	105.52	106.37

Burada F- Bağımsız ticaretçisi için Mart, Nisan ve Mayıs 2020 kayıp veri imputasyonları; genel ortalama göre yüksek, hedeflenen imputasyon yöntem civarlarında değerler, Haziran aylık değişimi; genel ortalamaya göre düşük ve hedeflenen imputasyon yöntemine göre yüksek çıktığı açıkça görünmektedir.

Cari uygulamada kullanılan genel ortalama, hedef ortalama, ileri taşıma imputasyonu ile alternatif olarak önerilen üç yöntemi ve bilgisayar programlama dilindeki imputasyon sonuçlarının betimsel istatistik değerlerini de karşılaştıralım:

```
> library(readxl)
> oneri_bil <- read_excel("C:/Users/oneri_bil.xlsx")
> View(oneri_bil)
> install.packages("plm")
> library(plm)
> ybkp<-pdata.frame(oneri_bil,index = c("isyerleri","aylar"))
> summary(ybkp)
```

```
> summary(ybkp)
 isyerleri     aylar    genel.ort.geo   hedef.ort.geo       carry          yontem1         yontem2         yontem3
 1:8       1      : 6   Min.   :5.100   Min.   :5.100   Min.   :5.100   Min.   :5.100   Min.   :5.100   Min.   :5.100
 2:8       2      : 6   1st Qu.:5.250   1st Qu.:5.250   1st Qu.:5.250   1st Qu.:5.250   1st Qu.:5.250   1st Qu.:5.250
 3:8       3      : 6   Median :5.490   Median :5.490   Median :5.490   Median :5.490   Median :5.490   Median :5.490
 4:8       4      : 6   Mean   :5.723   Mean   :5.733   Mean   :5.713   Mean   :5.879   Mean   :5.848   Mean   :5.729
 5:8       5      : 6   3rd Qu.:6.032   3rd Qu.:6.062   3rd Qu.:5.990   3rd Qu.:6.062   3rd Qu.:6.062   3rd Qu.:6.046
 6:8       6      : 6   Max.   :7.000   Max.   :7.000   Max.   :7.000   Max.   :9.220   Max.   :9.230   Max.   :7.000
           (Other):12
    R_Miss      R_Mice_Impute.Reg R_Mice_Impute.MI   R_outlier     R_outlier.Impute.lmrob
 Min.   :5.100   Min.   :5.100     Min.   :5.100    Min.   :0.000   Min.   :5.100
 1st Qu.:5.250   1st Qu.:5.250     1st Qu.:5.250    1st Qu.:5.250   1st Qu.:5.250
 Median :5.490   Median :5.490     Median :5.490    Median :5.490   Median :5.490
 Mean   :5.695   Mean   :5.727     Mean   :5.729    Mean   :5.339   Mean   :5.733
 3rd Qu.:5.990   3rd Qu.:6.048     3rd Qu.:5.992    3rd Qu.:5.990   3rd Qu.:6.062
 Max.   :7.000   Max.   :7.000     Max.   :7.000    Max.   :7.000   Max.   :7.000
 NA's   :3
```

```
> install.packages("psych")
> library(psych)
> describe(ybkp,trim = 0.05,type=3)
```

```
> describe(ybkp,trim = 0.05,type=3)
                       vars  n mean   sd median trimmed  mad min  max range  skew kurtosis   se
isyerleri*                1 48 3.50 1.73   3.50    3.50 2.22 1.0 6.00  5.00  0.00    -1.34 0.25
aylar*                    2 48 4.50 2.32   4.50    4.50 2.97 1.0 8.00  7.00  0.00    -1.31 0.33
genel.ort.geo             3 48 5.72 0.57   5.49    5.70 0.43 5.1 7.00  1.90  0.86    -0.45 0.08
hedef.ort.geo             4 48 5.73 0.58   5.49    5.71 0.43 5.1 7.00  1.90  0.81    -0.61 0.08
carry                     5 48 5.71 0.56   5.49    5.69 0.43 5.1 7.00  1.90  0.92    -0.28 0.08
yontem1                   6 48 5.88 0.92   5.49    5.78 0.43 5.1 9.22  4.12  1.84     3.20 0.13
yontem2                   7 48 5.85 0.84   5.49    5.77 0.43 5.1 9.23  4.13  1.80     3.68 0.12
yontem3                   8 48 5.73 0.57   5.49    5.70 0.43 5.1 7.00  1.90  0.83    -0.55 0.08
R_Miss                    9 45 5.69 0.57   5.49    5.66 0.43 5.1 7.00  1.90  1.00    -0.25 0.09
R_Mice_Impute.Reg        10 48 5.73 0.57   5.49    5.70 0.43 5.1 7.00  1.90  0.84    -0.51 0.08
R_Mice_Impute.MI         11 48 5.73 0.57   5.49    5.70 0.43 5.1 7.00  1.90  0.84    -0.55 0.08
R_outlier                12 48 5.34 1.50   5.49    5.51 0.43 0.0 7.00  7.00 -2.66     7.18 0.22
R_outlier.Impute.lmrob   13 48 5.73 0.58   5.49    5.71 0.43 5.1 7.00  1.90  0.81    -0.61 0.08
```

İstatistiksel göstergeler incelendiğinde; genel ortalama, hedef ortalama, ileri taşıma, $yöntem3\ (i\_müd19)$, "R_Mice_Impute Reg", "R_Mice_Impute MI" ve "R_outlier Impute lmrob" imputasyon sonuçları için $skew$ değeri pozitif ve $mean > median$ olduğundan dağılımda Sağa Çarpıklık vardır. Bu yöntemlerde kurtosis negatif olduğundan normal eğrisi normalden biraz basıktır. $yöntem1(i\_mon)$ ve $yöntem2(i\_\varphi)$ için bu durumlar $\text{skew} > 1$ ve $mean > \text{median}$ olup dağılım Sağa Çarpık ve kurtosis değeri pozitif olduğundan eğri normale göre daha sivri olduğu söylenebilir. Standart hata değerleri önerilen yöntemlerde kıyaslandığında $yöntem1(i\_mon)$ ve $yöntem2(i\_\varphi)$ de ortalamadan daha uzak dağılım gösterdikleri açıkça görülmektedir.

Betimsel istatistiksel sonuçlara göre önerilen yöntemlerden $yöntem1(i\_mon)$ ve $yöntem2(i\_\varphi)$ de standart sapma (sd), dağılım genişliği (range), çarpıklık (skewness), basıklık (kurtosis) ve standart hata (se) değerlerinin; programlama dilindeki sağlam imputasyon ve $yöntem3\ (i\_müd19)$ den farklılıkları açıkça gözlenmektedir. Önerilen $yöntem3\ (i\_müd19)$ için istatistiksel programlama dilindeki sağlam imputasyon sonuçlarına benzer betimsel istatiksel özellikler gösterdiği ve korelasyon analizi (correlation analysis) sonuçlarına göre sağlam yöntemlere benzerlik gösterdiği söylenebilir.

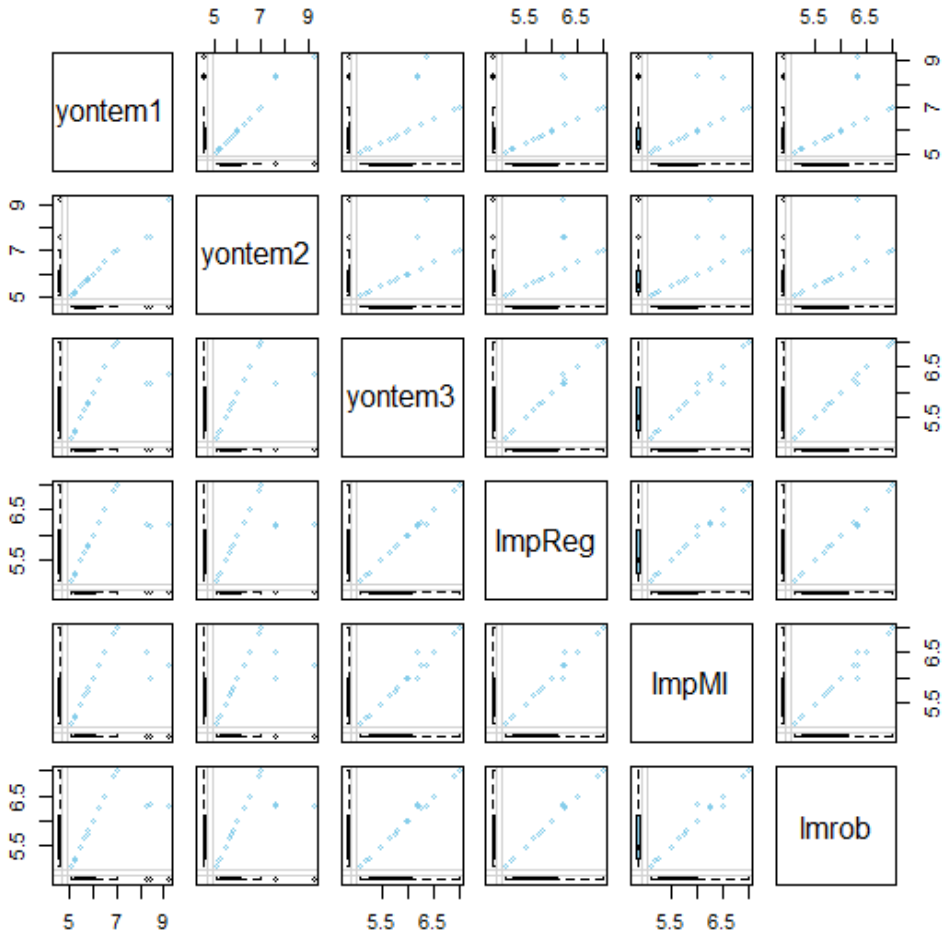
> library(VIM)

> library(DataExplorer)

> library(dataMaid)

> marginmatrix (oneri_bil [,c('yontem1', 'yontem2', 'yontem3', 'R_Mice_Impute Reg', 'R_Mice_Impute MI', 'R_outlier Impute lmrob')])

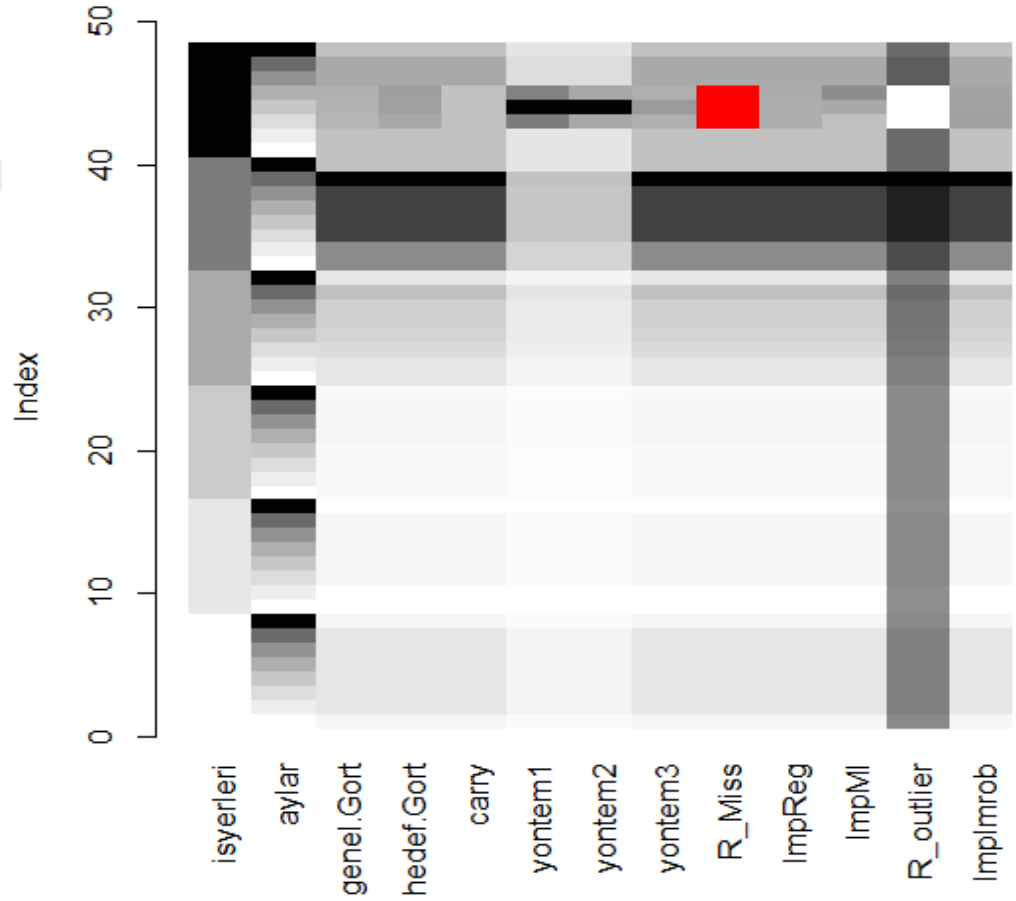
“marginmatrix {VIM}: Her panelin çizim kenar boşluklarında kayıp/atanmış (missing/imputed) değerler hakkında bilgi içeren bir dağılım grafiği matrisi oluşturur (R Core Team, 2021)”.



Şekil 5.2 Dağılım Grafiği Matrisi (scatterplot) (marginmatrix=> Plot Export)

> matrixplot(ybkp, interactive = T)

“matrixplot {VIM}: ile bir veri matrisinin tüm hücrelerinin dikdörtgenlerle görselleştirildiği bir matris grafiği oluşturulur. Mevcut veriler, sürekli bir renk şemasına göre kodlanırken, kayıp/atanmış (missing/imputed) veriler açıkça ayırt edilebilir bir renkle görselleştirilir (R Core Team, 2021)”.



Şekil 5.3 Matris Grafiği (matrixplot(ybkp)=> Plot Export)

Burada, $yöntem3(i\_müd19)$ ün; önerilen $yöntem1(i\_mon)$ ve $yöntem2(i\_\varphi)$ ye göre daha iyi sonuçlar verdiği tespit edilmiştir. Ek olarak, $yöntem3(i\_müd19)$ ün kontrolü amacıyla, alternatif hesaplama ile tanımlanan $yöntem3\_1(\ i\_müd19\_1)$ ve $yöntem3\_2(\ i\_müd19\_2)$ atama sonuçlarıyla da karşılaştırmalar yapılmıştır. Kontrol için tanımlanan $yöntem3\_1(\ i\_müd19\_1)$ folmülü aşağıdaki gibi hesaplanmıştır

$i\_müd19\_1$: $yöntem3\_1$: $x_{t,miss} = x_{t-1,last} + \left(sgn(mr_{t-12}) * x_{t-1,last} * (müd19/10)\right)$.

$i\_müd19$ adı fiyat küsuratını temsil eden belirlenmiş dijit değere ait bilgiyi kullandığından 0.193074 luk değişim imputasyonu (imputation $müd19 = 0.193074$) anlamına gelen $i\_müd19$ olarak isimlendirilmiş, analizlerde bazen kolaylık sağladığı için $yöntem3$ olarak da isimlendirilmişti. Kontrol amaçlı $i\_müd19$ den üretilen Coicop diğer sınıflamaları için düşünülen alternatif 1.imputasyon $i\_müd19\_1(yöntem3\_1)$ ve alternatif 2.imputasyon $i\_müd19\_2(yöntem3\_2)$ olarak isimlendirilmiştir.

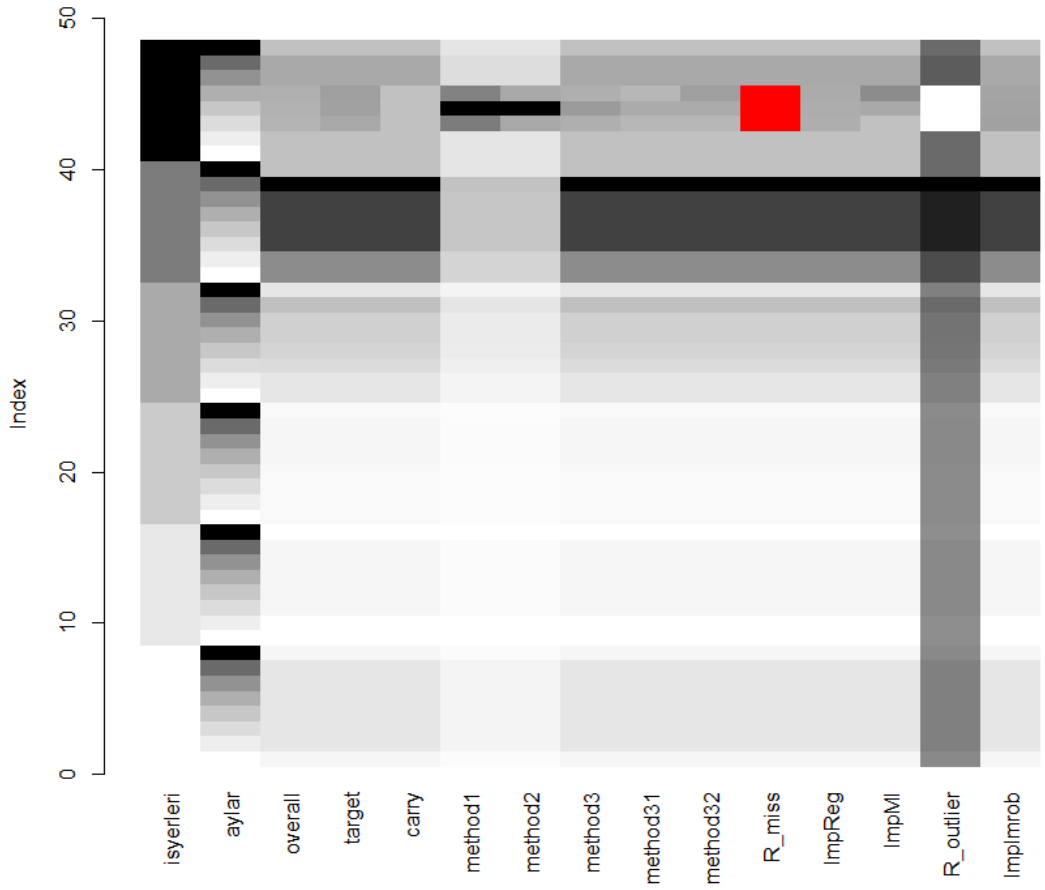
Ayrıca, önerilen yöntemin işaretlenmemiş etkisini incelemek için, kontrol amaçlı tanımlanan $yöntem3\_2(\ i\_müd19\_2)$ folmülü aşağıdaki gibi hesaplanmıştır

$i\_müd19\_2$: $yöntem3\_2$: $x_{t,miss} = x_{t-1,last} + \left(x_{t-1,last} * (müd19/10)\right)$.

TÜFE de Coicop gıda sınıflamasına özel tanımlanmış $yöntem3(\ i\_müd19)$ den üretilen kontrol amaçlı $yöntem3\_1$ ve $yöntem3\_2$ formüllerinin, veri yapısına göre Coicop daki diğer sınıflama grupları için kullanılabileceği düşünülmektedir. Aslında $yöntem3\_1(\ i\_müd19\_1)$ imputasyonu cari uygulamada kullanılan genel ortalama imputasyona (overall mean imputation) ve $yöntem3\_2(\ i\_müd19\_2)$ imputasyonu cari uygulamada kullanılan hedeflenen ortalama imputasyona (targeted mean imputation) en yakın yöntemdir. Ancak, mr_{t-12} ile işaretlemedeki amaç; TÜFE veri yapısı sebebiyle bir önceki yılın aynı ayına göre +/- işaretleme yani artış/düşüş eğilimini yansıtacak ekleme/çıkarma mantığı ile önceki bilgiye bağlı değişimi kullanmaktır. Bu çalışmada ise istatistik ofislerinin veri derleme aracında tanımlanacak yazım kuralları ile değer yüklemenin anında yapılması, tüm verinin derlenmesi sürecinin beklenmemesi amaçlanmaktadır. Veri yapısına bağlı olarak, genel geometrik ortalamanın önemli olduğu ve kullanıldığı, tüm verinin derlenmesinin beklenmesinde zaman sıkıntısı olmayan başka değer atama çalışmaları için $yöntem3\_1(\ i\_müd19\_1)$ formülü ile verilen imputasyon yöntem önerilir. Benzer şekilde hedeflenen geometrik ortalamanın yani kendi sınıfı içerisindeki değişim ile ilgilenilen geometrik ortalamanın önemli

olduğu çalışmalarda da $yöntem3\_2(i\_müd19\_2)$ formülü ile verilen imputasyon yöntem önerilir.

```
> describe(ybkp,trim=0.05,type=3)
             vars  n mean   sd median trimmed  mad min  max range  skew kurtosis   se
isyerleri*      1 48 3.50 1.73   3.50    3.50 2.22 1.0 6.00  5.00  0.00    -1.34 0.25
aylar*          2 48 4.50 2.32   4.50    4.50 2.97 1.0 8.00  7.00  0.00    -1.31 0.33
overallGmean    3 48 5.72 0.57   5.49    5.70 0.43 5.1 7.00  1.90  0.86    -0.45 0.08
targGmean       4 48 5.73 0.58   5.49    5.71 0.43 5.1 7.00  1.90  0.81    -0.61 0.08
carryGmean      5 48 5.71 0.56   5.49    5.69 0.43 5.1 7.00  1.90  0.92    -0.28 0.08
method1         6 48 5.88 0.92   5.49    5.78 0.43 5.1 9.22  4.12  1.84     3.20 0.13
method2         7 48 5.85 0.84   5.49    5.77 0.43 5.1 9.23  4.13  1.80     3.68 0.12
method3         8 48 5.73 0.57   5.49    5.70 0.43 5.1 7.00  1.90  0.83    -0.55 0.08
method3_1       9 48 5.72 0.57   5.49    5.70 0.43 5.1 7.00  1.90  0.86    -0.44 0.08
method3_2      10 48 5.73 0.57   5.49    5.70 0.43 5.1 7.00  1.90  0.84    -0.52 0.08
R_miss         11 45 5.69 0.57   5.49    5.66 0.43 5.1 7.00  1.90  1.00    -0.25 0.09
ImpReg         12 48 5.73 0.57   5.49    5.70 0.43 5.1 7.00  1.90  0.84    -0.51 0.08
ImpMI          13 48 5.73 0.57   5.49    5.70 0.43 5.1 7.00  1.90  0.84    -0.55 0.08
R_outlier      14 48 5.34 1.50   5.49    5.51 0.43 0.0 7.00  7.00 -2.66     7.18 0.22
Implmrob       15 48 5.73 0.58   5.49    5.71 0.43 5.1 7.00  1.90  0.81    -0.61 0.08
```



Şekil 5.4 Matris Grafiği ($yöntem3\_1$ ve $yöntem3\_2$ imputasyon kontrol)

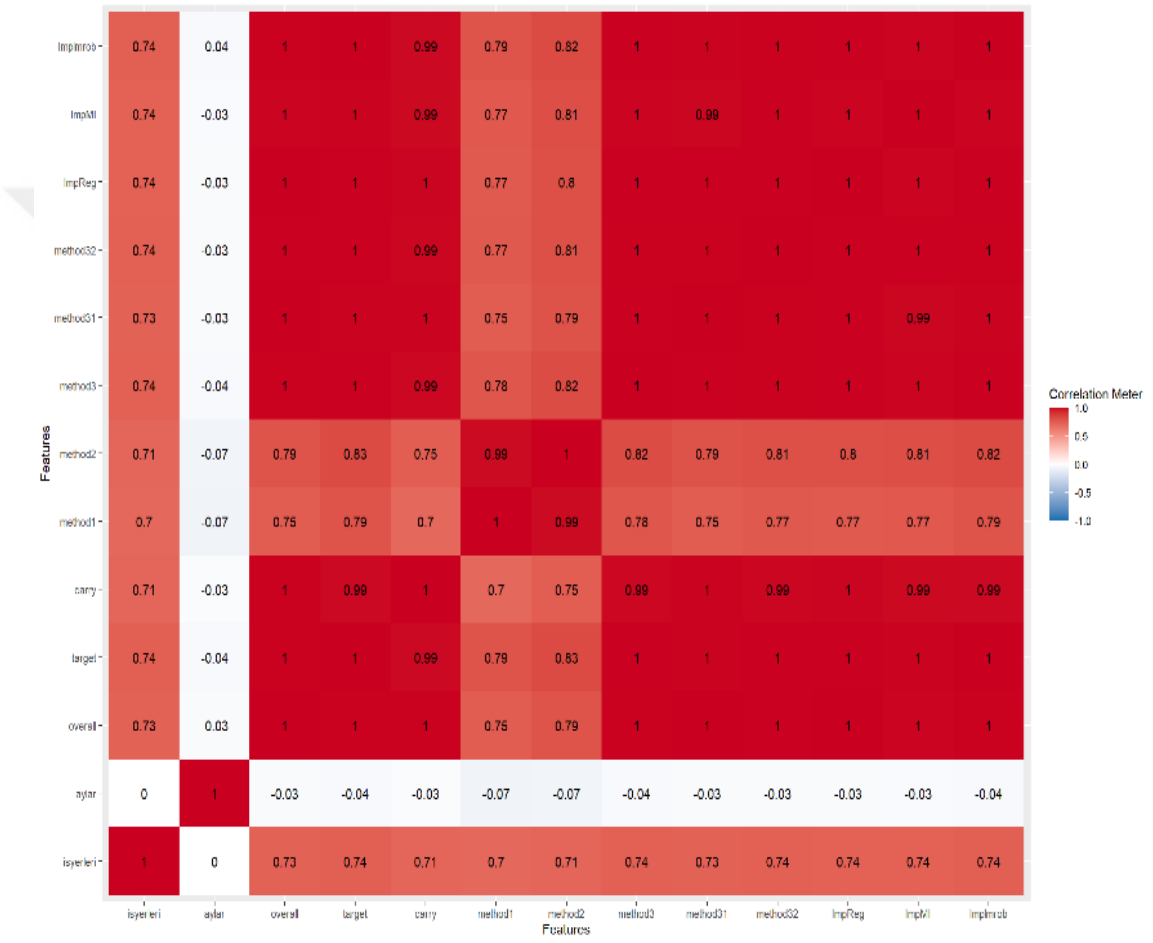
> library(pander)

> library(ggplot2)

> library(dataMaid)

> create_report(ybkp)

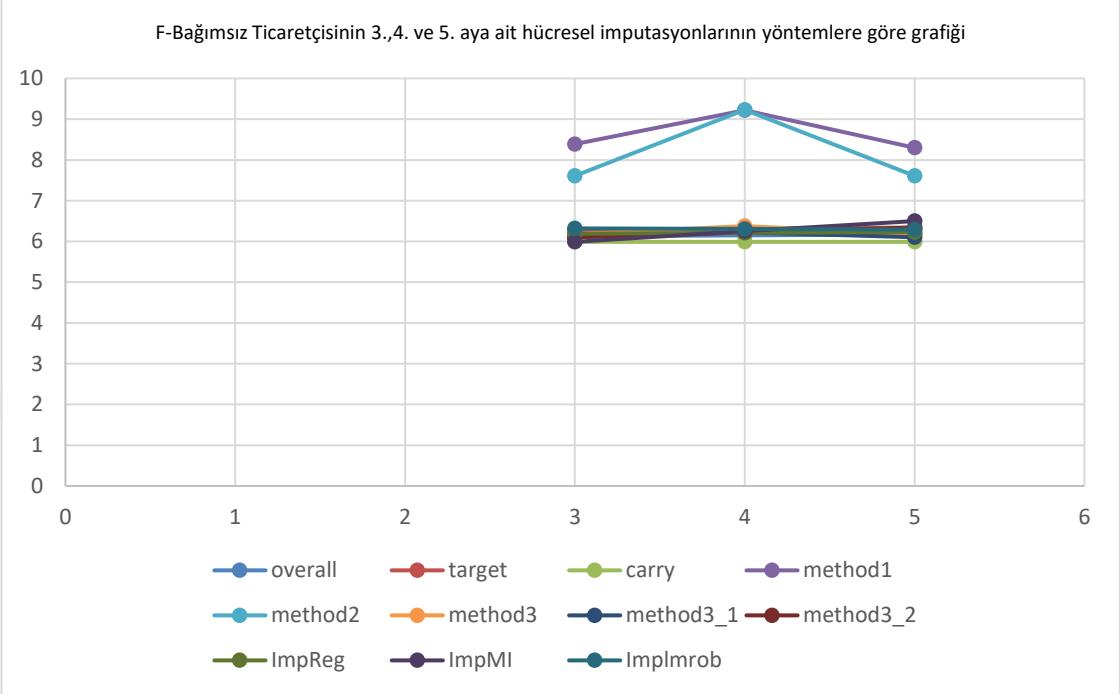
"create_report {DataExplorer}: bu fonksiyon, bir veri profili oluşturma raporu oluşturur (R Core Team, 2021)".



Şekil 5.5 Korelasyon analizi (Correlation Analysis) (create_report(ybkp)=> Data Profiling Report)

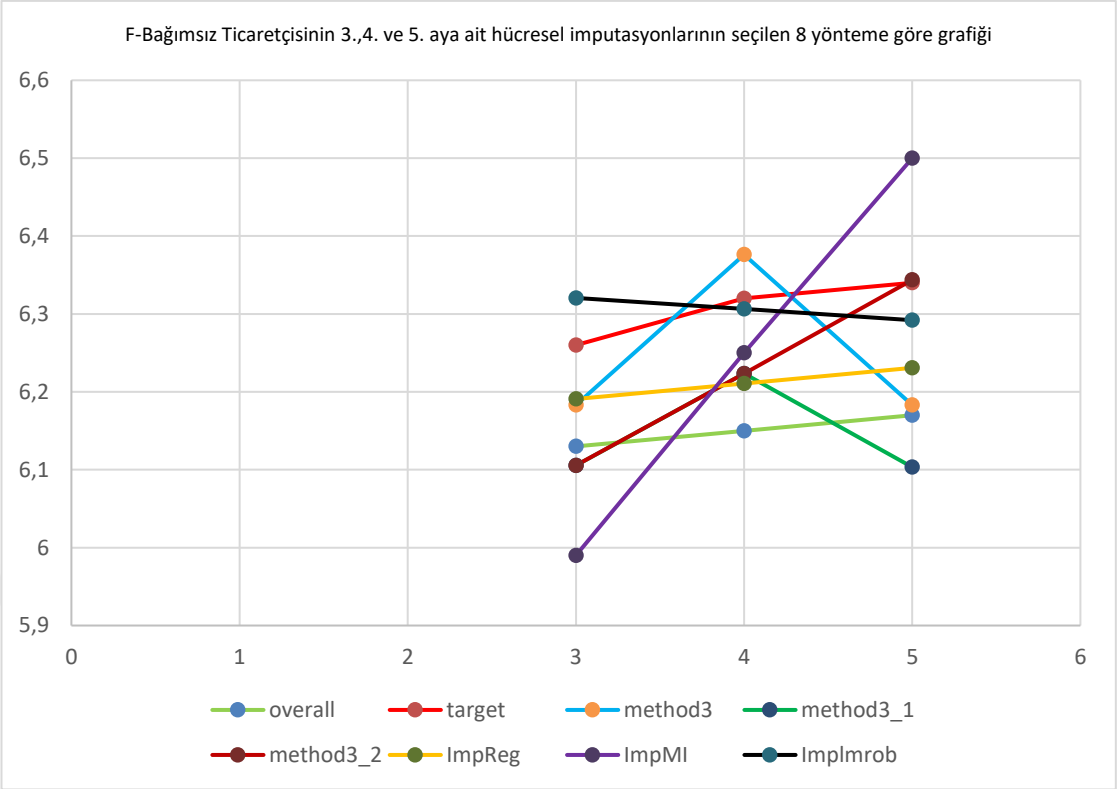
$yöntem3$ ($i\_müd19$) için korelasyon analizi sonuçlarına göre genel ortalama (overall) ve hedeflenen ortalama (targeted) imputasyon yöntemlerle aralarında direkt yönlü doğrusal ilişki olduğu söylenebilir. Standart hata değerleri önerilen yöntemlerde

kıyaslandığında $yöntem1(i\_mon)$ ve $yöntem2(i\_\varphi)$ de ortalamadan daha uzak dağılım gösterdikleri açıkça görülmektedir.



Şekil 5.6 F-Bağımsız Ticaretçisinin 3., 4. ve 5. aya ait hücresel imputasyonlarının yöntemlere göre grafiği

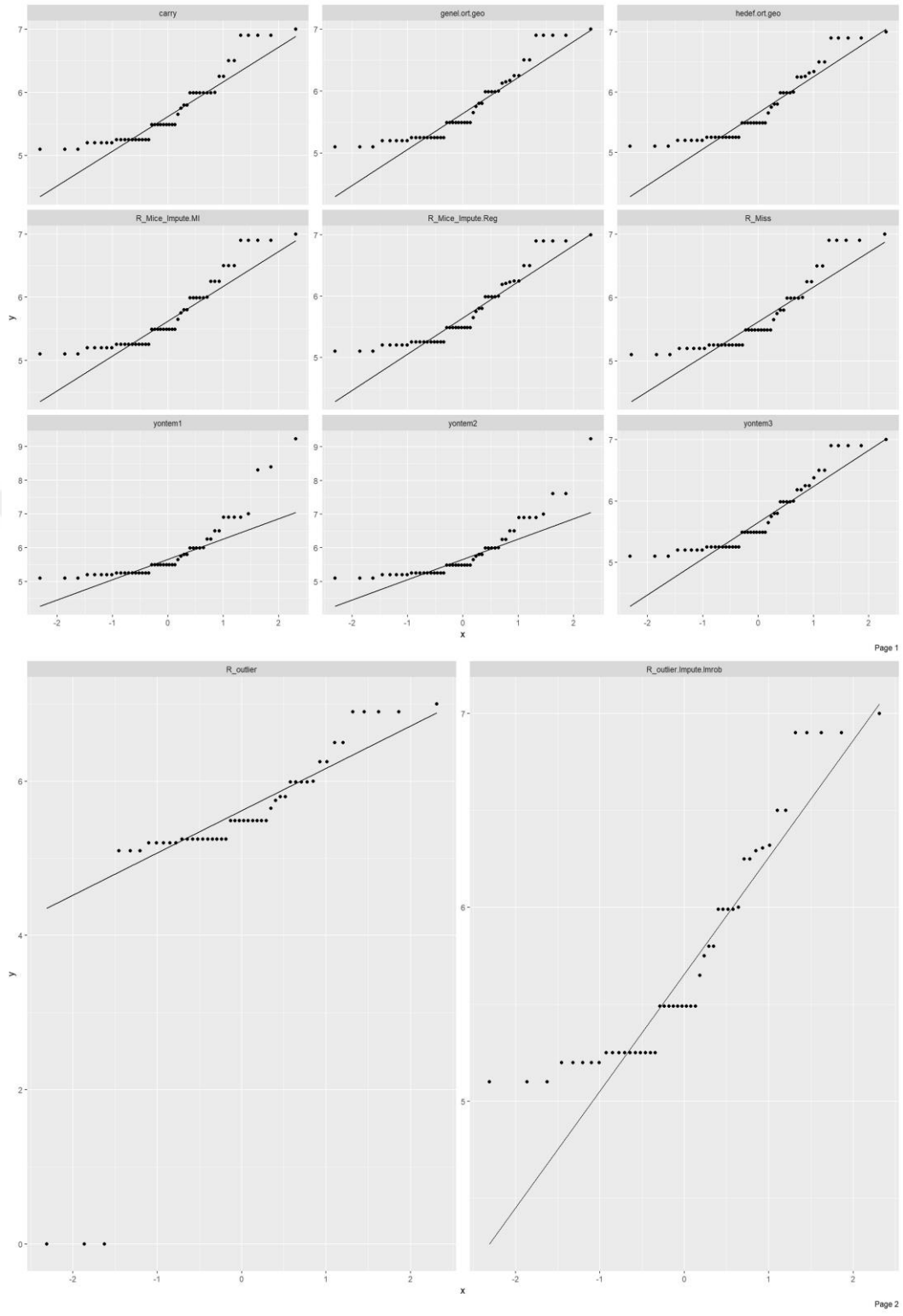
$yöntem3(i\_müd19)$ ün, önerilen $yöntem1(i\_mon)$ ve $yöntem2(i\_\varphi)$ ye göre daha iyi sonuçlar verdiği tespit edilmiştir. TÜFE de Coicop gıda sınıflamasına özel tanımlanmış $yöntem3(\,i\_müd19)$ den üretilen kontrol amaçlı $yöntem3\_1$ ve $yöntem3\_2$ formüllerinin, veri yapısına göre Coicop daki diğer sınıflama grupları için kullanılabileceği düşünülmektedir.



Şekil 5.7 F-Bağımsız Ticaretçisinin 3., 4. ve 5. aya ait hücresel imputasyonlarının seçilen 8 yönteme göre grafiği

Özetle, $yöntem3\_1(\,i\_müd19\_1)$ imputasyonu cari uygulamada kullanılan genel ortalama imputasyona (overall) ve $yöntem3\_2(\,i\_müd19\_2)$ imputasyonu cari uygulamada kullanılan hedeflenen ortalama imputasyona (targeted) yakın yöntemlerdir.

QQ Plot grafikleri incelendiğinde; genel ortalama, hedef ortalama, ileri taşıma, $yöntem3$ $(i\_müd19)$, "R_Mice_Impute Reg", "R_Mice_Impute MI" ve "R_outlier Impute lmrob" imputasyon sonuçları için $skew$ değeri pozitif ve $mean > median$ olduğundan dağılımda Sağa Çarpıklık vardır. Bu yöntemlerde kurtosis negatif olduğundan normal eğrisi normalden biraz basıktır. $yöntem1(i\_mon)$ ve $yöntem2(i\_\varphi)$ için bu durumlar skew > 1 ve $mean >$ median olup dağılım Sağa Çarpık ve kurtosis değeri pozitif olduğundan eğri normale göre daha sivri olduğu söylenebilir.



Şekil 5.8 QQ Plot (create_report(ybkp)=> Data Profiling Report)

6. SONUÇ

TÜFE hesaplamalarında anket döneminde reyonda ürün bulunamaması durumda uygulanan yöntemler ve metedoloji için IWGPS kuruluşları tarafından ortaklaşa yayınlanan 2020 TÜFE Klavuzunda yer alan kayıp veri durumundaki imputasyon yöntemlerinden; genel ortalama (overall mean), hedef (targeted) ortalama, ileri taşıma (carryforward) imputasyonları ele alınmıştır. Bu yöntemlerin uygulandığı Klavuzda kullanılan örnek veri seti ele alınarak; TÜFE hesaplamalarında kayıp veri olması durumunda uygulanan imputasyon tekniklerine alternatif üç yöntem önerilmiştir.

Benzer şekilde aynı veriyle istatistiksel programlama dilinde sağlam hücresel aykırı değer ve kayıp veri imputasyon hesaplamaları yapılmıştır. Burada kayıp veriyi ve aykırı değeri eş zamanlı aynı veri setinde değerlendirebilmek için, çok değişkenli veri setinde kayıp veri durumunda karşılaşılan problem, hücresel aykırı değer gibi değerlendirilmiştir. Bunun için uygulama çalışmasında, çok değişkenli veri setinde kayıp hücreye sıfır (0) fiyat ataması ile fiyatlar genel seviyesinde kayıp verinin aykırı değer olarak tanımlanması sağlanmıştır.

Önerilen üç yöntemin imputasyon sonuçlarıyla; TÜFE cari uygulamada kullanılan imputasyon yöntemlerinden genel ortalama, hedef ortalama, ileri taşıma imputasyon sonuçları ve istatistiksel programlama dilindeki imputasyon sonuçları karşılaştırılmıştır.

Çok değişkenli veri setinde kayıp hücreye sıfır (0) fiyat atamasıyla aykırı değer olarak tanımlayarak hesaplanan, sağlam aykırı değer imputasyon sonuçlarıyla benzerlik gösteren, $i\_m\ddot{u}d19(y\ddot{o}ntem3)$ imputasyonuyla TÜFE uygulamalarında COICOP Gıda sınıflaması için anında sıcağı sıcağına, veri derleme aracında hızlı bir şekilde hesaplamayla, tüm verinin toplanması beklenmeden, imputasyon yapılabileceği söylenebilir. Tüm dünyada TÜFE hesaplamasında kullanılan ve istatistiklerden üretilen imputasyon araçlarına yardımcı olmak için önerilen $i\_m\ddot{u}d19$ 'un, tüm kullanıcılara kolaylık sağlaması amaçlanmıştır.

TÜFE de Coicop gıda sınıflamasına özel tanımlanmış $i\_m\ddot{u}d19$ den üretilen kontrol amaçlı $i\_m\ddot{u}d19\_1(y\ddot{o}ntem3\_1)$ ve $i\_m\ddot{u}d19\_2(y\ddot{o}ntem3\_2)$ formüllerinin, veri sınıf yapısına göre Coicop daki diğer sınıflama grupları için kullanılabileceği öngörülmektedir. Kontrol amaçlı önerilen $i\_m\ddot{u}d19\_1$ imputasyonu cari uygulamada kullanılan genel ortalama imputasyona (overall mean) ve $i\_m\ddot{u}d19\_2$ imputasyonu cari uygulamada kullanılan hedeflenen ortalama imputasyona (targeted mean) en yakın yöntemdir. Veri yapısına bağlı olarak, genel geometrik ortalamanın önemli olduğu ve kullanıldığı başka yerine değer atama çalışmaları için $i\_m\ddot{u}d19\_1$ formülü ile verilen imputasyon yöntemi önerilmektedir. Benzer şekilde hedeflenen geometrik ortalamanın yani kendi sınıfı içerisindeki değişim ile ilgilenilen geometrik ortalamanın önemli olduğu çalışmalarda da $i\_m\ddot{u}d19\_2$ formülü ile verilen imputasyon yöntem önerilmektedir.

Ayrıca, hücresel aykırı gözlem olması durumunda sağlam tahmin yöntemleri ile istatistiksel veri analizi için önerilen $i\_m\ddot{u}d19$; TÜFE de Coicop gıda sınıflamasına göre ayarlanabilmesi ve bir önceki yılın aylık değişim oranına bağlı işaretleme yapılması, başka panel veri türü çalışmalarda, veri trendi yapısına göre uyarlanabilirdir. Genel olarak, çok değişkenli veri setinde hücresel aykırı gözlem olması durumunda aykırı değerleri reddetmek yerine, son gözlemi kullanarak, hiçbir zaman sıfır olmayacak şekilde ayarlanabilir $m\ddot{u}d19 = 0.193074$ değeri ile $i\_m\ddot{u}d19$, $i\_m\ddot{u}d19\_1$ ve $i\_m\ddot{u}d19\_2$ deki gibi ağırlıklandırılma ayarlanabilir ve belirlenmiş işaretlenebilir imputasyon yöntemi, TÜFE de Coicop sınıflamasına bağlı olarak ve başka panel veri çalışmalarında da kullanılması önerilmektedir. Veri yapısına bağlı olarak, $i\_m\ddot{u}d19$, $i\_m\ddot{u}d19\_1$ ve $i\_m\ddot{u}d19\_2$ de ihtiyaca bağlı revize ayarlama işlemleriyle, çok değişkenli veri setinde, hem hücresel aykırı değer hem de kayıp veri durumu için ortak bir "$m\ddot{u}d19$ ağırlıklı ayarlanabilir ileri taşıma" imputasyon yöntemi olarak da önerilmektedir.

KAYNAKLAR

Acuna, E. and Rodrıguez, C.. 2004. The treatment of missing values and its effect in the classifier accuracy. In D. Banks, L. House, F.R. Mc Morris, P. Arabie, W. Gaul (Eds). Classification, Clustering and Data Mining Applications. Springer-Verlag, 639-648, Berlin Heidelberg. DOI:10.1007/978-3-642-17103-1_60

Agostinelli, C., Leung, A., Yohai, V. J. and Zamar, R. H. 2015. Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. Springer Nature, 441-461. DOI 10.1007/s11749-015-0450-6

Allison, P. D.. 2001. Missing Data. Sage Publications, Inc, (Quantitative Applications in the Social Sciences), 104, Pennsylvania, USA.

Alqallaf, F. A., Konis, K. P., Martin, R. D. and Zamar, R. H. 2002. Scalable robust covariance and correlation estimates for data mining. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, 23 July 2002, Association for Computing Machinery, New York. 14-23. https://doi.org/10.1145/775047.775050

Alqallaf, F., Van Aelst, S., Yohai, V., and Zamar, R. H. 2009. Propagation of outliers in multivariate data. The Annals of Statistics, 37, 311-331.

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., and Tukey, J. W.. 1972. Robust estimates of locution: survey and advances. Princeton University Press, 373, Princeton, N.J.

Andy Chun Yin Leung 2016. Robust estimation and inference under cellwise and casewise contamination. PhD thesis, University of British, 125, Columbia.

Anonymous. 2020. Consumer Price Index Manual, Concepts and Methods. Identifiers: ISBN 978-1-51354-298-0 (PDF). Web Sitesi: https://www.ilo.org › publication › wcms_761444. Erişim Tarihi: 15.01.2023.

Anonymous. 2023. Unit of measure: Monthly rate of change. Web sitesi: https://ec.europa.eu/eurostat/databrowser/view/PRC_IPC_G20__custom_4733163/default/table?lang=en Data extracted on 31/01/2023 13:34:36 from [ESTAT] G20 CPI all-items - Group of Twenty - Consumer price index [PRC_IPC_G20__custom_4733163]. Erişim Tarihi: 31.01.2023.

Barnett, V. and Lewis, T. 1978. Outliers in Statistical Data. John Wiley and Sons, 376, New York.

Beckman, R. J. and Cook, R. D. 1983. Outlier…….s. Technometrics, 25 (2), 119-149.

Belsley, D. A., Kuh, E. and Welsch, R. E. 1980. Regression Diagnostics : Identifying Influential Data and Sources of Collinearity. John Wiley & Sons, 320. ISBN: 9780471691174; 0471691178

Bernoulli, D. 1777. The Most Probable Choice Between Several Discrepant Observations and The Formation There from of the Most Likely Induction. Reprinted in Biometrika, 48, 1-18 (1961, translated by C. G. Allen), London.

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

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

Buuren, Stef van. 2012. Flexible Imputation of Missing Data. CRC Press, 326, Taylor & Francis Group, NW. ISBN: 978-1-4398-6825-6.

Chauvenet, W. 1863. Method of least squares, appendix to manual of spherical and practical astronomy. Vol. 2. Lippincott, 469-566, 539-599, Philadelphia. Reprinted (1960) -5th edn. Dover, New York.

Cook, R. D.. 1977. Detection of Influential Observation in Linear Regression. Technometrics, Vol. 19, No. 1, pp. 15-18. Published by: American Statistical Association and American Society for Quality. http://www.jstor.org/stable/1268249.

Cook, R. D.. 1979. Influential Observations in Linear Regression. Journal of the American Statistical Association, Vol. 74, No. 365, pp. 169-174. Published by: American Statistical Association. http://www.jstor.org/stable/2286747.

Cool, A. L. 2000. A review of methods for dealing with missing data. Annual Meeting of the Southwest Educational Research Association, 34. Dallas. http://files.eric.ed.gov/fulltext/ED438311.pdf

Croux, C. and Öllerer, V. 2015. Comments on: Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. Springer, TEST (2015) 24,462–466. DOI 10.1007/s11749-015-0455-1

Çil, B. 1990. Regresyon analizinde tek bir sapan değerin "outlier'ın" belirlenmesine ilişkin metodların mukayesesi. Doktora Tezi, Ankara Üniversitesi Fen Bilimleri Enstitüsü, Ankara, Turkey.

Danilov, M., Yohai, V. J. and Zamar, R. H. 2012. Robust Estimation of Multivariate Location and Scatter in the Presence of Missing Data. Journal of the American Statistical Association, 107-499, 1178-1186. DOI: 10.1080/01621459.2012.699792

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

Enders, C. K. 2010. Applied Missing Data Analysis. The Guilford Publications, Inc. 401, New York.

Fisher, R. A. 1922. On the Mathematical Foundations of Theoretical Statistics. Philosophical Transactions of the Royal Society of London, 222, 309-368. https://doi.org/10.1098/rsta.1922.0009

Glaisher, J. W. L. 1872-73. On the Rejection of Discordant Observations, Monthly Notices of the Royal Astronomical Society. 33, 391-402. https://doi.org/10.1093/mnras/33.6.391

Grubbs, F. E. 1950. Sample Criteria for Testing Outlying Observations. Annals of Mathematical Statistics, 21, 27-58. https://doi.org/10.1214/aoms/1177729885

Hadi, A. 1992. Identifiying Multiple Outliers in Multivariate Data, Journal of The Royal Statistical Society. Series B. Methodological, 54, 3, Blackwell Publishing for The Royal Society.

Hampel, F. R. 1968. Contributions to the Theory of Robust Estimation. Unpublished Ph.D. thesis. University of California, 206, Berkeley (1976 published).

Hampel, F. R. 1971. General Qualitative Definition of Robustness. Annals of Mathematical Statistics, 42, 1887-96.

Hampel, F. R. 1974. The Influence Curve and its Role in Robust Estimation. Journal of the American Statistical Association, 69:346, 383-393.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. 1986. Robust Statistics: The Approach Based on Influence Functions. John Wiley & Sons, Inc., 526, New York.

Hardin, J. and Rocke, D. M. 2005. The Distribution of Robust Distances. Journal of Computational and Graphical Statistics, 14, 928-946. DOI: 10.1198/106186005X77685

Heckman, J. J. 1976. The common structure of statistical models of truncated, sample selection and limited dependent variables, and a simple estimator of such models. Annals of Economic and Social Measurement, 5, 475-492.

Hodges, J. L. Jr. 1967. Efficiency in normal samples and tolerance of extreme values for some estimates of location. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, 163-168.

Hodges, J. L. Jr. and Lehmann, E. L. 1963. Estimates of location based on rank tests, Ann. Math. Stat., 34, 598-611.

Hu, M. and Salvucci, S. 2001. A Study of Imputation Algorithms, 122. Working Paper No. 2001-17, Washington, DC. [http://nces.ed.gov/pubs2001/200117.pdf]

Huber, P. J. 1964. Robust estimation of a location parameter. Ann. Math. Statist. 35(1), 73-101.

Huber, P. J. 1972. Robust statistics: a review (The 1972 Wald Lecture). Ann. Math. Statist., 43, 1041-1067.

Huber, P. J. 1981. Robust Statistics. John Wiley and Sons, 320, New York. 0-471-41805-6.

Hubert, M. and Engelen, S. 2004. Robust PCA and classification in biosciences. Bioinformatics 20, 1728–1736.

Hubert, M., Rousseeuw, P. J. and Vanden Branden, K. 2005. ROBPCA: a new approach to robust principal components analysis. Technometrics 47, 64–79.

Hubert, M., Rousseeuw P. J. and Bossche W. V. 2019. MacroPCA: An All-in-One PCA Method Allowing for Missing Values as Well as Cellwise and Rowwise Outliers. Technometrics, 61:4, 459-473. https://doi.org/10.1080/00401706.2018. 1562989

İnal, C. ve Günay, S. 1993. Olasılık ve Matematiksel İstatistik. Hacettepe Üniversitesi Fen Fakültesi Beytepe Basımevi, 339- 349, Ankara.

İşçi Güneri, Ö., İncekırık, A. ve Durmuş, B.. 2021. Aykırı Değer Durumunda Bazı Sağlam Regresyon Yöntemlerinin Karşılaştırılması. New Era Internatıonal Journal of Interdıscıplınary Socıal Researches, Volume:6 Issue:11, ISSN 2757-5608. http://dx.doi.org/10.51296/newera.133

Irwin, J. 0. 1925. On a Criterion for the Rejection of Outlying Observations. Biometrika, 17, 238-250. DOI: 10.2307/2332079

Kalton, G., and Kasprzyk, D. 1982. Imputing for Missing Survey Responses, American Statistical Association, Proceedings of the Section on Survey Research Methods, 22-31.

Leung, A. C. Y.. 2016. Robust estimation and inference under cellwise and casewise contamination. A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy, 112, The University of British Columbia.

Leung, A., Zhang, H. and Zamar, R. H. 2016. Robust regression estimation and inference in the presence of cellwise and casewise contamination.

Computational Statistics & Data Analysis, Volume 99, July 2016, 1-11. http://dx.doi.org/10.1016/j.csda.2016.01.004

Leung, A., Yohai, V. and Zamar, R. H. 2017. Multivariate location and scatter matrix estimation under cellwise and casewise contamination. Computational Statistics & Data Analysis, Volume 111, July 2017, 59-76. http://dx.doi.org/10.1016/j.csda.2017.02.007

Liu P., Lei L. and Zhang X. F. 2004. A Comparison Study of Missing Value Processing Methods. Computer Science, 31(10): 155-156.

Little, R. J. A. and Rubin, D. B. 1987. Statistical Analysis with Missing Data. (1st Ed.) John Wiley&Sons, 291, New York.

Little, R. J. A. and Rubin, D. B. 2002. Statistical Analysis with Missing Data. (2nd Ed.) John Wiley&Sons, 409, New Jersey.

Little, R. J. A. 2011. Calibrated Bayes, for Statistics in General, and Missing Data in Particular. Statistical Science, Vol. 26, No. 2, 162–174. DOI: 10.1214/10-STS318

Lopuha, H. P. and Rousseeuw, P. J. 1991. Breakdown points of a_ne equivariant estimators of multivariate location and covariance matrices. The Annals of Statistics, 19, 229-248.

Machkour, J., Alt B., Muma M. and Zoubir A. M. 2017. Regression Shrinkage and Selection via the Lasso. J. Roy. Stat. Soc. B. Met., 267–288, 1996.

Machkour, J., Muma, M., Alt, B. and Zoubir, A. M. 2017a. A new sparse and robust adaptive Lasso estimator for the independent contamination Model. arXiv:1705.02162v1 [math.ST] 5 May 2017. Computer Science, Mathematics, 12, Cornell University.

Machkour, J., Alt, B., Muma, M. and Zoubir, A. M. 2017b. The Outlier-Corrected-Data-Adaptive Lasso: A new robust estimator for the independent contamination model. 2017 25th European Signal Processing Conference (EUSIPCO), ISBN 978-0-9928626-7-1

Maronna, R. A., Martin, R. D., and Yohai, V. J. 2006. Robust Statistics: Theory and Methods. Wiley, 417, New York.

Maronnaa, R. A. and Yohai, V. J. 2017. Robust and efficient estimation of multivariate scatter and location. Science Direct, Volume 109, May 2017, 64-75. http://dx.doi.org/10.1016/j.csda.2016.11.006

Molnar, F. J., Hutton, B. and Fergusson, D. 2008. Does analysis using "last observation carried forward" introduce bias in dementia research?. Canadian Medical

Association Journal, 179 (8): 751–753. doi: 10.1503/cmaj.080820. ISSN 0820-3946. PMC 2553855. PMID 18838445.

Mosteller, F. and Tukey, J. W. 1977. Data Analysis and Regression. Addison-Wesley, 588, Reading, Mass.

Newcomb, S. 1886. A Generalized Theory of the Combination of Observations so as to Obtain the Best Result. American Journal of Mathematics, 8 (4), 343-366.

Osborne, J. W. 2013. Best practices in data cleaning. Sage Publication, Inc, 275, California.

Pearson, E. S., and Chandra, S., C. 1936. The Efficiency of Statistical Tools and a Criterion for the Rejection of Outlying Observations. Biometrika, 28, 308-320, Oxford University Press. DOI: 10.2307/2333954

Peng, Liu and Lei, L.A. 2005. A Review of Missing Data Treatment Methods. Shanghai University of Finance and Economics, 8, Shanghai, P. R. China. ID: 15934251.

Peng, Liu and Lei, L. A. 2006. Missing Data Treatment Methods and NBI Model. Sixth International Conference on Intelligent Systems Design and Applications, 633-638, Jinan, China.

Penny, K. I. 1996. Appropriate Critical Values when Testing for a Single Multivariate Outlier by Using the Mahalanobis Distance. Applied Statistics, Vol. 45, No. 1, 73-81.

Pierce, B. 1852. Criterion for the rejection of doubtful observations. Astron. J., No. 45,2, No. 21,161-163.

R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Rawlings, J. O. 1988. Applied Regression Analysis: A Research Tool. Wadsworth & Brooks/Cole, 553, 0-534-09246-2, California.

Ricardo, A. M., Douglas, R. M. and Victor J. Y. 2006. Robust Statistics Theory and Methods. Wiley, 436. ISBN: 978-0-470-01092-1.

Rocke, D. M. 1996. Robustness properties of S-estimators of multivariate location and shape in high dimension. The Annals of Statistic, Vol. 24, No.3, 1327-1345.

Rousseeuw, P.J.. 1984. Least median of squares regression. Journal of the American Statistical Association, 79 (388), 871–880.

Rousseeuw, P. J., and Yohai, V. 1984. Robust regression by means of S-estimators, in Robust and Nonlinear Time Series Analysis. Lecture Notes in Statistics No. 26, Springer Verlag, New York, pp. 256-272.

Rousseeuw, P. J. and Leroy, A. M. 1987. Robust regression and outlier detection. John Wiley & Sons, Inc, 341, Canada. ISBN 0471-85233-3.

Rousseeuw, P. J. and Bossche, W. V. 2015. Comments on: Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. DOI 10.1007/s11749-015-0454-2

Rousseeuw, P. J. and Bossche, W. V. 2016. Detecting anomalous data cells. arXiv:1601.07251v1 [stat.ME], 19, 27 Jan 2016. https://wis.kuleuven.be/stat/robust/papers/TR/rousseeuwvandenbossche-dadc-arxiv-2016.pdf

Rousseeuw, P. J. and Bossche, W. V. 2017. Detecting deviating data cells. Technometrics, Vol 60, 2, 135-145. DOI: 10.1080/00401706.2017.1340909

Rosner, B. 1975. On the Detection of Many Outliers. Technometrics, 17, 221-227. DOI: 10.2307/1268354

Rubin, D. B. 1976. Inference and missing data (with discussion). Biometrika 63, 581–592. https://doi.org/10.1093/biomet/63.3.581

Schafer, J. L. 1997. Analysis of Incomplete Multivariate Data. Boca Raton, Chapman & Hall, 448, New York.

Schafer, J. L. 1999. Multiple imputation: a primer. Statistical Methods on Medical Research, 8(1), 3-15. https://doi.org/10.1177/096228029900800102

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

Stone, E. J. 1868. On the Rejection of Discordant Observations. Monthly Notices of the Royal Astronomical Society, 28, 165- 168.

Stone, E. J. 1873. On the Rejection of Discordant Observations. Monthly Notices of the Royal Astronomical Society, 34, 9-15. https://doi.org/10.1093/mnras/34.1.9
Student, 1927, Errors of routine analysis, Biometrika, 19, 151–164. https://doi.org/10.1093/biomet/19.1-2.151

“Student”. 1927. Errors of Routine Analysis. Biometrika, Vol. 19, No. 1/2, pp. 151-164.

Tabachnick, B. and Fidell, L. 1996. Using multivariate statistics (8th ed.). Pearson Education, 1018, USA.

Tarr, G., Müller, S. and Weber, N. C. 2016. Robust estimation of precision matrices under cellwise contamination. Computational Statistics & Data Analysis, Volume 93, 404-420. http://dx.doi.org/10.1016/j.csda.2015.02.005

Thompson, W. R. 1935. On a Criterion for the Rejection of Observations and the Distribution of the Ratio of Deviation to Sample Standard Deviation. The Annals of Mathematical Statistics, 6, 214-219. https://projecteuclid.org/download/pdf_1/euclid.aoms/1177732567

Tibshirani, R. 1996. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological). Vol. 58, No. 1, pp. 267-288.

Toka, O., Çetin, M. and Arslan, O. 2021. Robust regression estimation and variable selection when cellwise and casewise outliers are present. Hacettepe journal of mathematics and statistics, cilt.50, sa.1, 289-303.

Tukey, J. W. 1962. The future of data analysis. Ann. Math. Statist. 33 1–67.

Tukey, J. W. (1970-71) 1977. Exploratory Data Analysis (1970-71: preliminary edition). Addison-Wesley, xvi, 688, Reading, Mass.

Verboven, S. and Hubert, M. 2004. LIBRA: a MATLAB Library for Robust Analysis. Chemometrics and Intelligent Laboratory Systems, 75, 127-136.

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

Yazıcı, F. 2005. EM Algoritması ve Uzantıları. Yüksek Lisans, Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, 85, Ankara.

Yijun Zuo. 2001. Some quantitative relationships between two types of finite sample breakdown point. Statistics & Probability Letters, Volume 51 (4), 369-375.

Yohai, V. J. 1985. High breakdown-point and high efficiency robust estimates for regression. The Annals of Statistics, Vol.15 No.2, 642-656.

Wright, T. W. 1884. A Treatise on the Adjustment of Observations by the Melhod of Least Squares. Van Nostrand, 298, New York.

EKLER

EK 1 İstatistiksel Programlamalardan R da Mice Komutu ile İmputasyon

EK2 İstatistiksel Programlamalardan R da Mice, Method ="Norm.Predict" Komutu İle İmputasyon

EK 3 İstatistiksel Programlamalardan R da Lmrob Komutu İle İmputasyon

EK 1 İstatistiksel Programlamalardan R da Mice Komutu İle İmputasyon

“mice”: Generates Multivariate Imputations by Chained Equations (MICE), Impute the missing data *m* times

```
> RStudioMiceMiss <- read_excel("C:/Users/RStudioMiceMiss.xlsx")
> View(RStudioMiceMiss)
> library(plm)
> library(mice)
> pKayip<-pdata.frame(RStudioMiceMiss)
```

“RStudioMiceMiss' imizdeki değişkenler için mice ile kayıp değerler MI atamak ve 10 yeni atanmış veri kümesi oluşturur (pmm: imputation by predictive mean matching) (R Core Team, 2021).

```
> multiple_imputation = mice(pKayip, seed = 1, m = 10, print = FALSE)
> View(multiple_imputation)

> multiple_imputation[["imp"]]
$isyerleri
 [1] 1  2  3  4  5  6  7  8  9  10
<0 rows> (or 0-length row.names)
$aylar
 [1] 1  2  3  4  5  6  7  8  9  10
<0 rows> (or 0-length row.names)
```

$R_Mice_Impute

	1	2	3	4	5	6	7	8	9	10
44	5.99	5.99	6.25	5.99	6.25	5.99	5.99	5.99	5.99	5.99
45	5.99	6.25	5.99	5.99	5.99	6.25	5.99	5.99	5.99	5.99
46	6.00	6.50	6.25	5.99	5.99	6.25	6.50	6.25	6.25	5.99

EK 2 İstatistiksel programlamalardan R da mice, method ="norm.predict" komutu ile imputasyon

“method ="norm.predict", seed=1, m=10, print= F””

mice: Generates Multivariate Imputations by Chained Equations (MICE), Impute the missing data *m* times.

```
> RStudioMiceMiss <- read_excel("C:/Users/RStudioMiceMiss.xlsx")
> View(RStudioMiceMiss)
> library(plm)
> library(mice)
> pKayip<-pdata.frame(RStudioMiceMiss)
```

“RStudioMiceMiss' imizdeki emsal değerler oluşturmak üzere MICE komutu aracılığıyla norm.predict modelleme yaklaşımı kullanılır (**norm.predict_numeric_Linear regression**, predicted values) (R Core Team, 2021)”.

```
> RegMiss_imputed <- mice(pKayip, method="norm.predict” ,seed=1, m=10, print= F)
> View(RegMiss_imputed)
> RegMiss_imputed[["imp"]]
$isyerleri
 [1] 1  2  3  4  5  6  7  8  9  10
<0 rows> (or 0-length row.names)
$aylar
 [1] 1  2  3  4  5  6  7  8  9  10
<0 rows> (or 0-length row.names)
```

$R_Mice_Impute

	1	2	3	4	5	6	7	8	9	10
44	6.1908	6.1908	6.1908	6.1908	6.1908	6.1908	6.1908	6.1908	6.1908	6.1908
45	6.2108	6.2108	6.2108	6.2108	6.2108	6.2108	6.2108	6.2108	6.2108	6.2108
46	6.2308	6.2308	6.2308	6.2308	6.2308	6.2308	6.2308	6.2308	6.2308	6.2308

EK 3 İstatistiksel programlamalardan R da lmrob komutu ile imputasyon

“lmrob”: “Computes fast MM-type estimators for linear (regression) models (R Core Team, 2021)”

```
> library(readxl)
> library(DataExplorer)
> RStudioLmrobOutlier <- read_excel("C:/Users/ RStudioLmrobOutlier.xlsx")
> View(RStudioLmrobOutlier)
> library(ggplot2)
> library(plm)
> library(psych)
> paykr<-pdata.frame(RStudioLmrobOutlier)
> library(dbplyr)
> library(robustbase)

> control <- lmrob.control(method = "SMDM",compute.outlier.stats = c("S", "MM", "SMD", "SMDM"))

> fit1 <- lmrob(R_outlier ~ isyerleri*aylar, RStudioLmrobOutlier, control = control)

# “lmrob {robustbase}: The method argument takes a string that specifies the estimates to be calculated as a chain. Setting method='SMDM' will result in an intial S-estimate, followed by an M-estimate, a Design Adaptive Scale estimate and a final M-step (R Core Team, 2021)”.

> summary(fit1)
```

```
> summary(fit1)

Call:
lmrob(formula = R_outlier ~ isyerleri * aylar, data = RStudioLmrobOutlier,
    control = control)
 \--> method = "SMDM"
Residuals:
    Min      1Q  Median      3Q     Max
-6.4453 -0.2673 -0.1438  0.3042  0.8519

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.83431    0.24131  20.034  < 2e-16 ***
isyerleri      0.20150    0.06288   3.205  0.00252 **
aylar          0.01058    0.05764   0.184  0.85518
isyerleri:aylar  0.01164    0.01500   0.776  0.44216
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Robust residual standard error: 0.3966
Multiple R-squared:  0.5372,    Adjusted R-squared:  0.5057
Convergence in 8 IRWLS iterations

Robustness weights:
 3 observations c(44,45,46) are outliers with |weight| = 0 ( < 0.0021);
 33 weights are ~= 1. The remaining 12 ones are summarized as
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.6779  0.8352  0.9353  0.8968  0.9884  0.9987
Algorithmic parameters:
     tuning.chi1      tuning.chi2      tuning.chi3      tuning.chi4               bb      tuning.psi1
      -5.000e-01        1.500e+00               NA        5.000e-01        5.000e-01       -5.000e-01
     tuning.psi2      tuning.psi3      tuning.psi4       refine.tol          rel.tol        scale.tol
       1.500e+00        9.500e-01               NA        1.000e-07        1.000e-07        1.000e-10
       solve.tol      eps.outlier            eps.x warn.limit.reject warn.limit.meanrw
       1.000e-07        2.083e-03        7.640e-11        5.000e-01        5.000e-01
    nResample         max.it       best.r.s       k.fast.s          k.max    maxit.scale      trace.lev            mts
          500             50              2              1            200            200              0           1000
   compute.rd      numpoints fast.s.large.n
            0             10           2000
                psi           subsampling                   cov compute.outlier.stats1 compute.outlier.stats2
              "lqq"         "nonsingular"             ".vcov.w"                   "S"                  "SM"
compute.outlier.stats3 compute.outlier.stats4
               "SMD"                "SMDM"
seed : int(0)
```

> augment(fit1)

R programı augment (fit1) ile 44., 45. ve 46. hücre fitted değeri sonuçları

	R_outlier	isyerleri	aylar	.fitted	.resid
44	0.00	6	3	6.320582	-6.32058163
45	0.00	6	4	6.306284	-6.30628392
46	0.00	6	5	6.291986	-6.29198620