T.C.
ABANT İZZET BAYSAL ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ
ÖLÇME VE DEĞERLENDİRME ANA BİLİM DALI
EĞİTİMDE ÖLÇME VE DEĞERLENDİRME PROGRAMI

KAYIP VERİ YÖNTEMLERİNİN FARKLI DEĞİŞKENLER ALTINDA VARYANS ANALİZİ (t-TESTİ, ANOVA) PARAMETRELERİ ÜZERİNE ETKİSİNİN İNCELENMESİ

Yüksek Lisans Tezi

Hazırlayan
Begüm ÖZTEMÜR

Danışman
Yrd. Doç. Dr. İ. Alper KÖSE

BOLU-2014

EĞİTİM BİLİMLERİ ENSTİTÜSÜ MÜDÜRLÜĞÜ'NE,

Begüm ÖZTEMÜR'e ait **" Kayıp Veri Yöntemlerinin Farklı Değişkenler Altında Varyans Analizi (T-Testi, Anova) Parametreleri Üzerine Etkisinin İncelenmesi"** adlı çalışma, jürimiz tarafından Eğitim Bilimleri Anabilim Dalı, Eğitimde Ölçme ve Değerlendirme Bilim Dalında YÜKSEK LİSANS TEZİ olarak kabul edilmiştir. 24./01/2014

Akademik Unvan ve Adı Soyadı		İmza
Üye (Tez Danışmanı)	: Yrd. Doç. Dr. İbrahim Alper KÖSE	.....
Üye	: Doç. Dr. Zekeriya NARTGÜN	.....
Üye	: Yrd. Doç. Dr. Eralp BAHÇİVAN	.....

Eğitim Bilimleri Enstitüsünün Onayı

Prof. Dr. Soner DURMUŞ

Enstitü Müdürü

ABSTRACT

EXAMINING THE EFFECT OF MISSING DATA METHODS ON VARIANCE ANALYSIS (T-TEST, ANOVA) PARAMETERS UNDER DIFFERENT VARIABLES

Begüm ÖZTEMÜR
Measurement and Assessment in Education Programme
Thesis Advisor: Assist. Prof. İbrahim Alper KÖSE
January, 2014 - xvi + 81 Pages

In the process of scientific researches there could potentially be missing data. It is well known that these missing data can have negative effects to the results of the researches. It is a big impediment for the researchers to have exact results. Process of missing value analysis applications include approachies aiming to find solution to these problems experienced by the researchers. Firstly researchers should find out the factors resulting missing data and define severity of the missing.

In this research, missing value applications were compared using value sets produced in different numbers of unit. These set of values were produced in a way that they would have normal distributions in high and low correlation groups having respectively 50, 100, 200, 400 units. Under random conditions, reduced set of values having respectively %5, %10, %20 losses are in the form of Missing Completely at Random (MCAR). In the value sets produced, mean substitution, regression method, expectation-maximization (EM) method were applied, instead of deletion methods. Difference between the results of methods differentiated dramatically in value sets of different volume and different correlations. It is observed that in value sets with low numbers of units (50 unit, 100 units) regression and EM application are usefull on the other hand in value sets with high numbers of units (200 units, 400 units) mean substitution instead of regression method have more consistent results. Also It is

observed that in value sets with low correlation deletion method was not an efficent application.

Key Words: Missing Value Analysis, Imputation Method, EM, Mean Method, Regression.

ÖZET

KAYIP VERİ YÖNTEMLERİNİN FARKLI DEĞİŞKENLER ALTINDA VARYANS ANALİZİ (T-TESTİ, ANOVA) PARAMETRELERİ ÜZERİNE ETKİSİNİN İNCELENMESİ

Begüm ÖZTEMÜR
Eğitimde Ölçme ve Değerlendirme Anabilim Dalı
Tez Danışmanı: Yrd. Doç Dr. İbrahim Alper KÖSE
Ocak, 2014 - xvi + 81 Sayfa

Bilimsel araştırma sürecinde yapılan birçok çalışmalarda kayıp veriler söz konusu olabilmektedir. Veri gruplarındaki kayıp değerlerin araştırma sonuçlarını etkilediği bilinmektedir. Bu durum araştırmacıların daha net bilgiler elde etmesinde büyük bir engel oluşturmaktadır. Kayıp veri analizi yöntemleri araştırmacıların çok sık karşılaştıkları bu soruna çözüm getirmeyi amaçlayan yaklaşımlar içerir. Araştırmacılar öncelikle kayıp veriyi ortaya çıkaran nedenleri ve bu kayıp durumunun önem derecesini belirlemelilerdir.

Bu araştırmada farklı birim sayılarına sahip türetilmiş veri setleri yardımıyla, kayıp veri yöntemleri karşılaştırılmıştır. Türetilen veri setleri sırasıyla 50, 100, 200, 400 birime sahip düşük ve yüksek korelasyon gruplarındaki, normal dağılıma sahip olacak şekilde oluşturulmuştur. Rastgele koşullar altında eksiltilen %5, %10, %20 kayıp içeren veri setleri TROK (MCAR) yapıya sahiptir. Türetilen veri setlerine silme yöntemi, yerine ortalama koyma yöntemi, regresyon yöntemi ve beklenti maksimizasyonu (BM) yöntemi uygulanmıştır. Çalışma sonucunda kullanılan yöntemler arasındaki farklar, farklı korelasyona sahip, farklı büyüklükteki veri setlerinde oldukça farklılaşmıştır. Düşük birimli veri setlerinde (50 birim, 100 birim) regresyon ve BM yöntemleri daha etkin sonuçlar verirken, yüksek birimli veri setlerinde (200 birim, 400 birim) regresyon ve yerine ortalama koyma yöntemi tam verilerle daha tutarlı sonuçlar vermiştir. Ayrıca

düşük korelasyonlu veri setlerinde silme yönteminin etkin bir yöntem olmadığı da görülmüştür.

Anahtar Kelimeler: Kayıp Veri Analizi, Atama Yöntemleri, BM, Yerine Ortalamayı Koyma, Regresyon.

Etik İlkelere Uyulduğuna İlişkin Metin

Yüksek lisans tezi olarak sunduğum, **"Kayıp Veri Yöntemlerinin Farklı Değişkenler Altında Varyans Analizi (T-Testi, Anova) Parametreleri Üzerine Etkisinin İncelenmesi"** başlıklı çalışmanın yazılmasında, bilimsel ve etik kurallara uyulduğunu, başkalarının eserlerinden yararlanılması durumunda atıfta bulunulduğunu, kullanılan verilerde herhangi bir tahrifat yapılmadığını, tezin tamamının ya da bir kısmının bu üniversite veya başka bir üniversitede bir tez çalışması olarak sunulmadığını beyan ederim. 24/01/2014

Begüm ÖZTEMÜR

İÇİNDEKİLER

TABLOLAR DİZİNİ

ŞEKİLLER DİZİNİ

SEMBOLLER / KISALTMALAR DİZİNİ

TROK **:** Tamamiyle Rassal Olarak Kayıp

ROK **:** Rassal Olarak Kayıp

İE **:** İhmal Edilemez Kayıp

MCAR **:** Missing Completely at Random

MAR **:** Missing at Random

NI **:** Noignorable

BM **:** Beklenti Maksimizasyonu

EM **:** Expectation Maximization

X **:** Değişken adı

A **:** Veri Matrisi

BÖLÜM I

1. Giriş

Çalışmanın bu bölümünde araştırmanın problem durumu, araştırmanın amacı, araştırma soruları, araştırmanın önemi, araştırmanın varsayımları ve sınırlılıkları ele alınmıştır.

1.1. Problem Durumu

Yapılan birçok araştırmada verilerin kayıp olması söz konusu olabilmektedir. Veri gruplarındaki kayıp değerlerin araştırma sonuçlarını etkilediği bilinmektedir (Özdamar, 2002). Kayıp veri, veri analizinin en yaygın problemlerinden biridir. Kayıp veri içeren analizlerin doğru ve güvenilir olmadığı bilinen bir gerçektir. Mümkün olduğunca yansız ve güvenilir sonuçlara ulaşabilmek için kayıpsız ve büyük veri setlerine ihtiyaç vardır. Stekhoven & Bühlmann (2011)'a göre istatistiksel analizler için veri setlerinin kayıpsız (tam) olması büyük önem taşımaktadır; çünkü bütün istatistiksel veri analizi paket programları verilerin kayıpsız olduğu varsayımına dayanmaktadır. İstatistiksel kestirimlerde kayıp veriler olası bir yanlılık durumunu doğurmaktadır. Veri toplamadaki başarısızlık analiz için geçerli durum sayısının azalmasına, bu azalma değerli bilgi kaybının oluşmasına sebep olmaktadır. Kayıp veriler, ihmal edilmiş, kaybedilmiş, çelişkili, hatalı bilgiler topluluğudur. Bu durum tam ve kayıp kayıtlar arasında büyük farklar ortaya çıkarır ve yordama sürecinde farklı sonuçlar doğmasına yol açabilmektedir.

Kayıp veri probleminin ciddiyeti verilerin ne kadar ve neden kayıp olduğuna bağlıdır (Tabachnick ve Fidell, 2001). Günümüzde veriye ulaşmada kullanılan

gözlemlerde sık sık bilgi/veri kaybıyla karşılaşılmaktadır. Araştırmacının süreç boyunca kayıpsız veri setleri elde etmeye çalışmasına rağmen çeşitli hatalardan dolayı veri kayıpları yaşanabilmektedir. Örneğin, hane halkı araştırmalarında aile bireyleri gelirlerini belirtmekten kaçınabilir, bir endüstriyel denemede bazı sonuçlar deney süreci ile ilgisiz mekanik sorunlar nedeniyle kayıp olabilir. Bir kamuoyu araştırmasında bazı bireyler, birkaç aday içerisinden tercihlerini açıkça ifade etmemiş olabilirler (Yazıcı, 2005). Yazıcıoğlu ve Erdoğan (2007)'a göre ise bu hatalar deneklerden kaynaklanmaktadır. Özellikle psikoloji, tıp gibi veri toplamanın zor olduğu ve az sayıda gözlenen veri ile çalışılan alanlarda, oluşan kayıp veriler araştırma sonucunu gerçeği yansıtmaya engeldir. Kişinin yaş, cinsiyet, zeka, bilgi düzeyi, sosyal yapısı gibi özellikleri; ölçümleri dolayısıyla veri sonuçlarını etkilemektedir. Bireyin o zamandaki durumu, yorgunluk, zindelik, sorulan sorulara yeterli zamanı ayıramama, fazla düşünmeden seçenekleri işaretleme gibi sonuçlara götürebilmektedir. Kalaycı (2008)'ya göre bazı veri hataları da veri toplayan kişiden kaynaklanmaktadır. Hatalı gözlem, yanlılık, kayıp-yanlış kayıt yapma gibi hatalar veri sonuçlarında gerçek olmayan farklılıklar yaratmaktadır. Ancak yanıtlayıcı kaynaklı ortaya çıkan kayıp veri sürecinin önceden kestirilmesi olanaksız olabilir. Bu durumda araştırmacının yapması gereken kayıp veri sürecini ortaya çıkaran bir yapının varlığının olup olmadığının araştırılmasıdır. Bu incelemeyi yaparken araştırmacının; kayıp verilerin gözlemlere rastgele mi saçıldığını yoksa belirgin bir yapı mı oluşturduğunu ayrıca da kayıp verilerin ne kadar sıklıkla karşımıza çıktığını belirlemesi gerekmektedir.

1.2. Problem Cümlesi

Kayıp veri içeren farklı oran ve büyüklükteki veri setlerinde kullanılan silme ve atama yöntemleriyle oluşan yeni veri setlerindeki t testi, ANOVA sonuçları tam veri setleri sonuçlarıyla ne kadar benzerdir ve kullanılan bu yöntemler sonucu istatistiksel hata tipleri oluşur mu?

1.2.1. Alt problemler

Yukarıda belirtilen problem cümlesi doğrultusunda aşağıdaki alt problemlere

yanıt aranacaktır:

1. Normal dağılıma sahip düşük ve yüksek korelasyonlu tam veri setleri (50-100-200-400 birim) ve belirli oranlardaki (%5, %10, %20) veri setlerine silme ve atama yöntemi kullanılarak oluşan yeni veri setlerindeki t-testi sonuçları arasındaki benzerlik durumu nedir?
2. Normal dağılıma sahip düşük ve yüksek korelasyonlu tam veri setleri (50-100-200-400 birim) ve belirli oranlardaki (%5, %10, %20) veri setlerine silme ve atama yöntemi kullanılarak oluşan yeni veri setlerindeki ANOVA sonuçları arasındaki benzerlik durumu nedir?
3. Normal dağılıma sahip düşük ve yüksek korelasyonlu tam veri setleri ile belirli oranlardaki kayıp veri setlerine çeşitli atama yöntemleri kullanılarak oluşan yeni veri setlerinde yapılan analizler sonucu hata tiplerine rastlanır mı?

1.3. Araştırmanın amacı

Bu çalışmada kayıp gözlem içeren çok değişkenli veri setlerinde kayıp veri probleminin giderilmesine yönelik çeşitli istatistiksel yöntemler incelenecektir. Koşullu türetilecek olan farklı korelasyona sahip, farklı oranlarda kayıplar içeren, farklı büyüklükteki veri setlerinde oluşan kayıp veri probleminin çözümü için çeşitli atama yöntemleri kullanılacaktır. Normal dağılım gösteren düşük ve yüksek korelasyonlu farklı veri setleri ve kayıp veri tamamlama yöntemleri ile her biri ayrı ayrı tamamlanan yeni veri setleri arasındaki t-testi ve ANOVA sonuçlarının parametreleri hesaplanacaktır. Bu parametrelerin değerlerine göre birbirleri ile benzerlikleri bakımından karşılaştırma yapılacaktır. Ayrıca t-testi ve ANOVA sonuçlarından kullanılan yöntemlerin benzerliği ve etkinliğinin belirlenmesi amaçlanmaktadır.

Bu araştırmada koşullu türetilmiş veri setleri;

- Çok değişkenli normal dağılıma sahip veri setleri,
- Farklı büyüklükteki veri setleri (n=50, n=100, n=200, n=400) ,
- Değişkenlerdeki kayıp veri sayısı farklı olan veri setleri (%5, %10, %20) ,

- Değişkenler arasındaki korelasyona göre yüksek ya da düşük korelasyonlu veri setleri olarak ayrılacaktır.

1.4. Araştırmanın önemi

Kayıp veri içeren analizlerin doğru ve güvenilir olmadığı oldukça açıktır. Mümkün olduğunca yansız ve güvenilir sonuçlara ulaşabilmek için kayıpsız ve büyük veri setlerine ihtiyaç vardır. İstatistiksel analizler için veri setlerinin kayıpsız (tam) olması büyük önem taşımaktadır; çünkü bütün istatistiksel veri analizi paket programları verilerin kayıpsız olduğu varsayımına dayanmaktadır (Özdamar, 2002). Uzun yıllar araştırmacılar kayıp verileri göz ardı edilmesi gereken bir durummuş gibi düşünmüş ya da basit çözüm yolları ile sorun giderilmeye çalışılmıştır. Araştırmayı yapan bilim insanlarının çalışmalarında bu durumdan bahsetmemeleri ise böyle bir sorunun giderilmesi için kapsamlı çalışmaların yapılmasını geciktirmiştir. Eğitim araştırmacılarının kendi ihtiyaçlarına en uygun yöntemi belirlemede eksiklikleri bulunmaktadır. Genellikle de daha önce kullanılmış ve popülerlik kazanmış yazılım yöntemlerini kullanırlar. Ancak bu yolu izlerken yanlış yöntem kullanımının araştırma sonuçları üzerinde önemli bir önyargı doğuracağını fark etmezler. Bu eğilimin de araştırmacılar içinde farklı sebepleri vardır. Murtonen ve Lethtinen (2003)'e göre bu sebepler arasında alınan eğitimin yüzeysel olması, teori ve pratik arasındaki ilişkiyi kurmanın zorluğu, araştırmanın görsel resmini çizememe, içerik eksikliği ve eğitim / sosyoloji alanlarındaki bazı kavram ve yöntemlere yönelik tutumlar gösterilebilir (Cheema, 2012).

Büyüklüğü ve kayıp miktarlarının farklı olduğu veri setlerinde kayıplılık sorununu gidermede, alternatif olarak kullanılan atama ve silme yöntemleri ile yapılmış varyans analizleri sonucu çıkan parametrelerde, anlamlı bir farklılık oluşturup oluşturmadığı oldukça önemlidir. Araştırmacılar silme yönteminin daha kolay uygulanmasından dolayı sıkça tercih etmektedirler. Bu durum diğer yöntemlerin az kullanılmasına neden olmaktadır. Kayıp veri sorununu gidermede diğer yöntemlerin de etkiliği araştırmacılara yol gösterici olacaktır. Araştırmacıların önem verdiği bir diğer

nokta ise hipotez ve araştırma sonucunda karar vermeyi etkileyen hata tipleridir. Bu durumda aslında gerçekte manidar farklılık olmayan ($p>0.05$) durumlara yanlış olarak fark var ($p<0.05$) denilmesiyle oluşan 1. tip hata ya da aslında manidar farklılık olan ($p<0.05$) durumlara yanlış olarak fark yoktur ($p>0.05$) denilmesiyle oluşan 2. tip hatalar araştırmacıyı yanlış yönlendirmektedir.

Yukarıdaki durumlar göz önüne alındığında ve ulaşılabilen literatür tarandığında, kayıp veri problemine yönelik çalışmaların özellikle Türk bilim dünyasında oldukça az ve yetersiz olduğu görülmektedir. Yapılan bu araştırma sonuçlarının araştırmacı ve uygulamacılara yol gösterici nitelikte olacağı düşünülmektedir.

1.5. Sınırlılıklar

1. Bu araştırma normal dağılıma sahip veri setleri ile sınırlıdır.
2. Araştırmada kullanılan veri setleri hacimce n=50, n=100, n=200, n=400 birimlik verilerle sınırlıdır.
3. Bu araştırmadaki değişkenlerimiz X_1, X_2, X_3 olacak şekilde sadece üç ile sınırlıdır.
4. Bu araştırmadaki veri setleri %5, %10, %20 kayıp olma durumu ile sınırlıdır.
5. Bu araştırmadaki veri setleri TROK yapısı ile sınırlıdır.
6. Bu araştırma silme yöntemi, yerine ortalamayı koyma yöntemi, BM yöntemi ve regresyon yöntemi ile sınırlıdır.
7. Bu araştırma varyans analizlerinden t testi ve ANOVA testi ile sınırlıdır.

BÖLÜM II

2. Kuramsal Temeller ve Literatür Taraması

Bu bölümde kayıp veri kavramı ile ilgili tanımlar, kayıp veri süreci, kayıp veri mekanizmaları, kayıp veri yapıları, ele alınan silme ve atama yöntemleri ile ilgili kuramsal açıklamalara yer verilmiştir.

2.1. Genel Bilgiler

Günümüzde, kuramsal ve uygulamalı bilim alanlarında bilimsel araştırmalara büyük önem verilmektedir. Bilimsel araştırmalarda; planlama, araştırma alanlarının ve örneklemin seçilmesi, veri toplama, verilerin tablo ve grafiklerle özetlenmesi, hipotezlerin oluşturulması, hipotezlerin test edilmesi, sonuçların yorumlanması ve genellenmesi, bilimsel raporların hazırlanması ve sunum teknikleri ile ilgili bilimsel bilgi üretiminde istatistiksel yaklaşımlar kullanmak büyük önem taşımaktadır. Bilimsel araştırmalarda, bilgi iletişimini kolaylaştıracak ortak yaklaşımlar kullanmak ve etkin çalışmalar yapabilmek için istatistiksel yöntemlerden yararlanmak gerekmektedir (Özdamar, 2002).

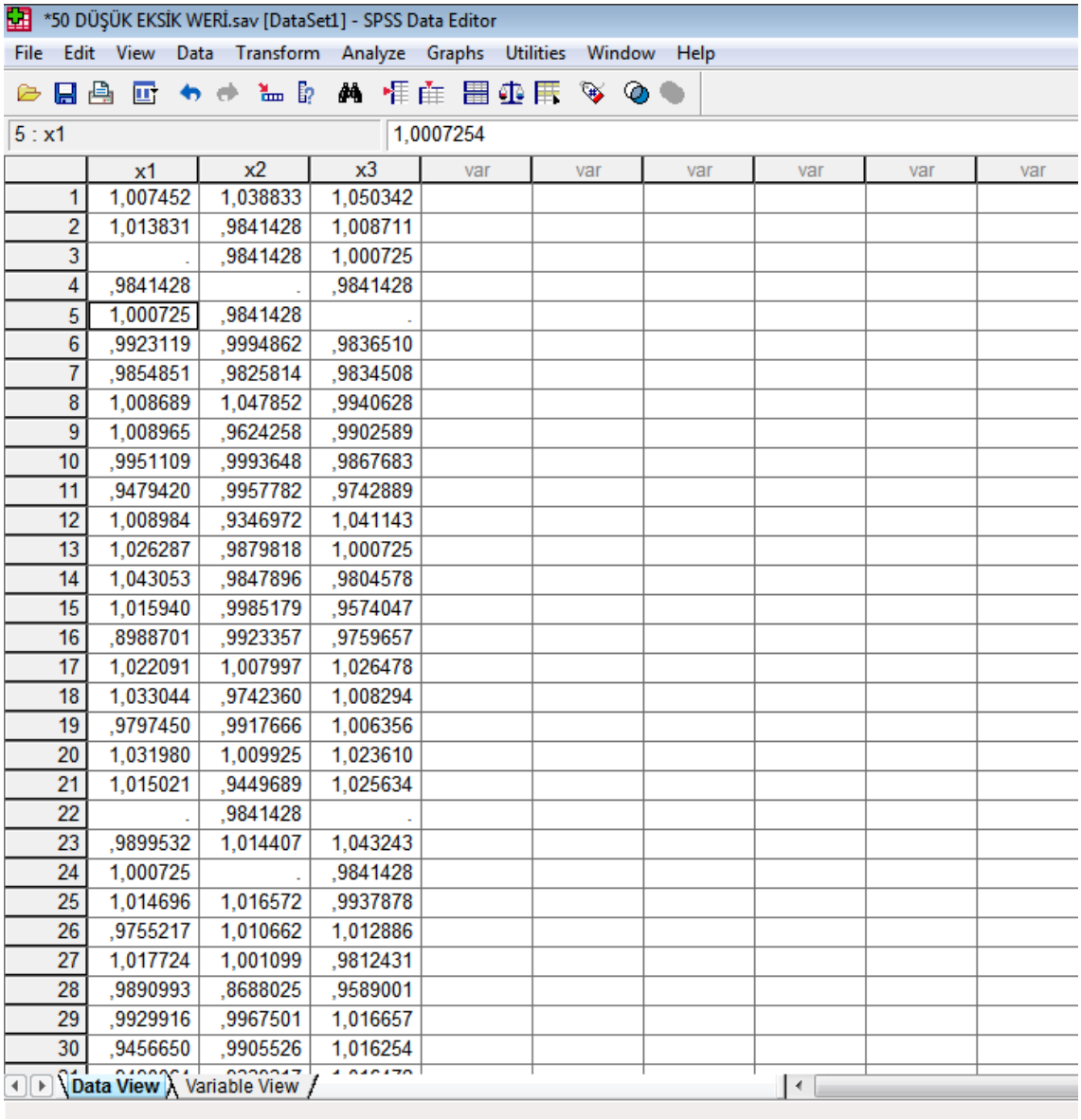
İstatistik, uygulamalı matematik araştırmacılarına tanımlamalar, ilişkilendirmeler ve bu ilişkileri anlamlandırma imkanı sağlayan bir bilim dalıdır (Shukla, Sınghaı ve Thakur, 2011). İstatistiksel yöntemler ise, elde edilen verilerin incelenmesi sonucunda, hangi çözümlemenin ya da modelin, probleme uygun olduğunun belirlenmesi için geliştirilmişlerdir. Verileri incelemek için oluşturulan veri matrisinin ‘satır’ları birimleri, gözlemleri ya da konu ile ilişkili verileri temsil ederken, ‘sütun’ları da ölçülen her bir birim için değişkenleri temsil etmektedir (Yazıcı, 2005).

Bir dizi sayının düzenli sıra ve sütunlar halinde dikdörtgen biçiminde gösterilmesine matris adı verilir. Matrisler genellikle büyük harflerle gösterilir ve matris isminin altına alt yazım olarak matrisin boyutunu (kaç sıra, kaç sütun içerdiğini) belirtmek mümkündür. Bir matriste n tane sıra ve p tane sütun bulunur. n sıra ve p sütun içeren A matrisi ve bu matrisin elemanları a_{ij} Şekil 2-1.'deki gibi gösterilir. Genelde n birimden elde edilen ve p değişken değerini içeren matrisler veri matrisi olarak adlandırılır (Özdamar, 2010).

$$A_{nxp} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{11} & \dots & a_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{np} \end{bmatrix}$$

Şekil 2-1. Verilerin matris ile gösterimi

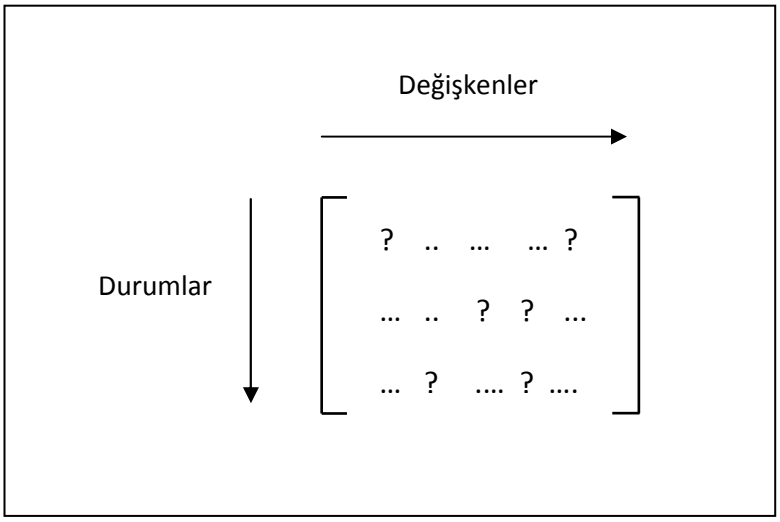
Bir araştırma, konusu ne olursa olsun iki ya da daha fazla değerlendirme durumu ve bu durumlarda düzeyi ölçümlenen iki ya da daha fazla değişken özellik içerir. Böylece bir araştırma süreci, değişkenler ve durumlar ölçeğinde iki boyutlu bir matris olarak düşünülebilir (Baygül, 2007). Analiz sonuçlarında verilerden geçerli sonuç ve çıkarımlara ulaşabilmek için, nitelikli verilerle çalışılmış olması önemlidir. Özdamar (2002)'a göre elde edilen veri setinin analize hazır hale getirilmesi aşamasında dikkat edilmesi gereken birtakım unsurlar bulunmaktadır. Bu unsurlardan biri de verilerdeki hata kontrolüdür. Özellikle veri setindeki hatalı giriş, minimum-maximum, aşırı değer veya kayıp veri gibi, araştırmanın sonucunu direkt etkileyecek ve hatalara sebebiyet vermeyecek kontrollerin yapılması gerekmektedir.



	x1	x2	x3	var	var	var	var	var	var
1	1,007452	1,038833	1,050342						
2	1,013831	,9841428	1,008711						
3	.	,9841428	1,000725						
4	,9841428	.	,9841428						
5	1,000725	,9841428	.						
6	,9923119	,9994862	,9836510						
7	,9854851	,9825814	,9834508						
8	1,008689	1,047852	,9940628						
9	1,008965	,9624258	,9902589						
10	,9951109	,9993648	,9867683						
11	,9479420	,9957782	,9742889						
12	1,008984	,9346972	1,041143						
13	1,026287	,9879818	1,000725						
14	1,043053	,9847896	,9804578						
15	1,015940	,9985179	,9574047						
16	,8988701	,9923357	,9759657						
17	1,022091	1,007997	1,026478						
18	1,033044	,9742360	1,008294						
19	,9797450	,9917666	1,006356						
20	1,031980	1,009925	1,023610						
21	1,015021	,9449689	1,025634						
22	.	,9841428	.						
23	,9899532	1,014407	1,043243						
24	1,000725	.	,9841428						
25	1,014696	1,016572	,9937878						
26	,9755217	1,010662	1,012886						
27	1,017724	1,001099	,9812431						
28	,9890993	,8688025	,9589001						
29	,9929916	,9967501	1,016657						
30	,9456650	,9905526	1,016254						

Şekil 2-2. Verilerin SPSS programına yerleştirilmesi

Veri matrisindeki girdiler gerçek sayılardır. Bunlar sürekli olan değişkenleri ya da cevap kategorilerini temsil ederler. Birçok araştırmada veri matrisindeki girdilerden bazılarının gözlenemediği durumlar olabilmektedir (Yazıcı, 2005). Şekil 2-2.'de girdileri gerçek sayılardan oluşan 3 değişkenli bir SPSS veri değerleri görülmektedir. Şekilde görüldüğü gibi bazı verilerde kayıp olma durumu söz konusudur.



Şekil 2-3. Kayıp Veri Matrisi Gösterimi

Veri matrisinde kayıplar söz konusu ve bu kayıplar önlem almayı gerektirecek düzeyde ise, araştırmamızda kayıp veri sürecinin varlığından söz edilebilir. Bu durumda, kayıp veriler yardımıyla elde edilen istatistiksel sonuçlar, çalışmada kullanılan değişkenlerin kayıp veri sürecinden etkilendiği derecede yanlı olabilecektir. Ancak, kayıp verinin etkisi, sonuçlar üzerindeki yanlılığı ile sınırlı değildir. Örneğin, kayıp veriye ilişkin sorunların giderilemediği çalışmalarda, kayıp verili gözlemlerin veriden çıkartılması söz konusu olabilmektedir. Bu ise gözlem sayısında ciddi bir azalmayı beraberinde getirebilecek ve yeterli olarak düşünülmüş bir örneklem yetersiz sayıdaki bir örnekleme dönüşecektir (Alpar, 2003).

Sonuç olarak herhangi bir araştırmada karşımıza nedeni bilinmeyen kayıp veri veya veriler çıkabilir. Bu durumda araştırmacılar iki farklı yol izlemektedirler. Araştırmacılar ilk olarak kayıp verileri gözardı etmektedir. İkinci yol olarak da çeşitli teknikleri kullanarak kayıp verileri tamamlama yoluna gitmektedirler (Yazıcı, 2005). Kalaycı (2008)'ya göre bazı araştırmacılar kayıp veriye yol açan gözlemleri veri grubundan çıkartmaktadırlar. Bu durumda gözlem sayısında önemli bir ölçüde azalma olmakta bu da yeterli olan örneklem büyüklüğünü olumsuz yönde etkilemektedir. Ayrıca araştırmanın güvenirliğini ve sonuçlarını önemli ölçüde etkilemektedir. O halde kayıp veriyle karşılaşıldığında şunlar yapılabilir:

- Veriye yeni gözlem değerleri eklenebilir.
- Verideki kayıp değerler çeşitli istatistiksel yaklaşımlarla giderilmeye çalışılır.

2.1.1. Kayıp veri süreci

Verideki kayıp değerlerin giderilmesi öncesinde kayıp veri süreçlerinin belirlenmesi gerekir. Bazen kayıp veri süreci araştırmacının denetimi altındadır ve tanımlanabilir. Bu gibi durumlarda kayıp veri ‘’*ihmal edilebilir kayıp veri’’* olarak adlandırılır. Bu tür kayıp veri sorununu çözmek için kayıp veri problemini çözecek belirli yöntemler kullanılmasına gerek duyulmaz. Örneğin, evrenden bir örneklem çektiğimizde, örnekleme girmeyen gözlemler kayıp veri çeşidi olarak algılanabilir. Araştırmalarda iyi bir örnekleme planı yapılarak (olasılıklı örnekleme yöntemleri ile) örneklemde olmayan gözlemler nedeniyle ortaya çıkan kayıp veri sorunu ihmal edilebilir konuma getirilebilir ve örnekleme alınmama nedeniyle ortaya çıkan kayıp veri, istatistiksel işlemlerde örnekleme hatasına eşdeğer tutulabilir. Yani, kayıp verinin ihmal edilebilir olarak değerlendirilmesi için kayıp veri sürecinin rastgele olması, dolayısıyla eldeki verinin tam ve kayıp değerler setinin rastgele örneklemi olmasıdır (Alpar, 2003).

2.2. Kayıp Veri ile İlgili Çalışmalar

Kayıp veri çalışmaları 1966’da Afiffi ve Elashoff’un araştırma verilerinin sapma eksikliklerinin algoritmik ve hesaba dayalı çözümlerinin yanı sıra gözlenemeyen birimlerin atılması ve tekli atama işlemleri üzerinde yapılan araştırmalarla başlamıştır. 1977’de Dempster, Laird ve Rubin beklenti maksimizasyonu (BM) üzerinde yoğunlaşmışlardır. Çalışmada BM algoritmasının olabilirliğinin ve yakınsamasının veriler üzerindeki monoton davranışı izlenmiştir. 1987’de Rubin’in veri ataması gibi algoritmik hesaplarla kayıp veri sorununa çözüm aranmıştır. 1990’lı yıllarda çoklu atama (multiple imputation) ve Monte Carlo yöntemleri ile Bayes teknikleri üzerine çalışmalar artmıştır.

2000’li yıllara gelindiğinde farklı alanlarda özellikle BM üzerindeki çalışmalar artmıştır. Gildea ve Hofmann (1999) bilgisayar ses tanıma programları için BM algoritması kullanmış olup bir istatistiksel dil modellemesi oluşturmuşlardır. Iturria ve

Blangero (2000) insanların genlerinin genelleşmesine yönelik varyans bileşenleri yöntemini geliştirmek için BM algoritmasına bir yaklaşım önermişlerdir. Schneider (2001) BM algoritmasını iklim verilerine uygulayarak parametre tahmini yapmıştır. Krishner, Cadez, Smyth ve Kamath (2003) BM algoritmasını uzay şekillerini sınıflandırmada kullanmışlardır. Evens (2003) sualtı kamerasında balıkların sayısı, çeşidi ve gürültülerini sınıflandırmada parametre tahminini BM ile yapmıştır. Cheema (2012) farklı kayıp içeren farklı büyüklükteki veri setlerinde tek grup için t testi, bağımsız gruplar için t testi, tek yönlü ANOVA ve lineer regresyon yöntemlerini karşılaştırmıştır.

Türkiye'de ise Oğuzlar (2001), alan araştırmalarında kayıp değer sorununa çözüm önerileri getirmek için çalışma yapmıştır. Çalışmasında Dünya Bankası' nın web sayfasından aldığı verilerle liste bazında silme, çiftler bazında silme, BM ve regresyon yöntemlerini incelemiştir. Bal (2003) sağlık alanında çok gruplu veri setlerinde eksik gözlem sorununun çözümlemesi için çiftler bazında silme, yerine ortalama koyma, BM ve regresyon yöntemlerini karşılaştırmıştır ayrıca verilerin standart sapma, mod, medyan değerlerindeki değişmeleri incelemiştir. Yazıcı (2005) kayıp veri içeren problemler için parametre tahmini hesaplamalarında, son yıllarda yoğun olarak kullanılan BM algoritması ve uzantılarını tanıtan bir çalışma yapmıştır. Baygül (2007) sağlık alanında kayıp veri yönteminde etkin kullanılan yöntemleri karşılaştırmıştır. Satıcı ve Kadılar (2009) kayıp gözlem olması durumunda kitle ortalamasına ilişkin Sing ve Horn (2000), Sing ve Deo (2003) ve Rueda vd. (2005) tarafından sunulan tahmin ediciler incelenmiş ve bunlara alternatif iki yeni tahmin edici önerilmiştir. Yozgatlıgil vd. (2011) Türkiye'nin iki farklı iklim bölgesine ait yağış ve sıcaklık serilerinde kayıp veri tahmini için kullanılan yöntemler karşılaştırılmıştır. Çokluk ve Kayrı (2011) bir ölçme aracının yapı geçerliliğinin test edilmesinde, kayıp değerin olmadığı ve farklı oranlarda kayıp durumunda, farklı yöntemler sonucu elde edilen faktör yapılarının, düzeltilmiş madde-toplam korelasyonlarının ve Cronbach-alfa içtutarlılık katsayılarını karşılaştırmalı olarak incelemişlerdir.

2.3. Kayıp Veri Mekanizmaları

Kayıp veri mekanizmasının belirlenmesi bu sorunun giderilmesinde kullanılacak analizin seçilmesinde, dolayısıyla da doğru sonuçlara ulaşılmasında oldukça önem taşımaktadır. Kayıp veri mekanizmalarının sınıflandırılması ilk kez Rubin (1976) tarafından yapılmıştır. Kayıp veri mekanizmaların sınıflandırılması hangi atama yönteminin kullanılmasına karar verilmesi açısından son derece önemlidir. Little ve Rubin (2002) kayıp veri mekanizmalarını;

- Tamamıyla Rassal Olarak Kayıp (Missing Completely at Random, MCAR, TROK)
- Rassal Olarak Kayıp (Missing at Random, MAR, ROK)
- İhmal Edilemez Kayıp (Noignorable, NI, İEK) olacak şekilde üç kategoriye ayırmıştır.

2.3.1. Tamamıyla rassal olarak kayıp (missing comletely at random, MCAR, TROK)

X ve Y birbirinden farklı değişkenler olsun. Yanıt olasılığı X ve Y değişkenlerinden bağımsız olduğunda yani iki değişkenin de gözlenme olasılığı birbirinden etkilemediğinde; X ve Y'nin gözlemlenmiş olması tamamen rassal olarak kayıp durumunu oluşturmaktadır (Little ve Rubin, 2002; Allison, 2001). Diğer bir deyişle kayıp olma durumu analizde yer alan özel değişkenlerle ilgili değildir. TROK koşulunun sağlanıp sağlanmadığı, yanıt verenler ile yanıt vermeyenler arasındaki gözlenen verilerin karşılaştırılmasıyla sağlanabilir. Eğer veriler TROK değilse, sonuçlar yanlı olacaktır ve bu yüzden de daha güçlü tekniklerin kullanılması uygun olacaktır (Little ve Rubin, 2002; Alpar, 2003; Yazıcı, 2005).

Y= (y_{ij}) , (nxk) boyutunda, kayıp değer içermeyen dikdörtgen bir veri setini göstersin. Bu matriste; i. satır y_i=(y_{i1},…,y_{ik}) şeklinde gösterilir ve y_{ij}, Y_j değişkeninin i. biriminin değeridir. Kayıp veri yapısı içeren veri setlerinde, kayıp veri gösterge matrisi M=(m_{ij}) olarak tanımlanır. Burada eğer y_{ij} kayıp veri ise m_{ij} =1 değerini, değilse 0 değerini alır. M matrisi kayıp veri yapısını tanımlar. İlerleyen kısımlarda kayıp veri yapısı örnekleri gösterilmiştir. Bazı yöntemler herhangi bir kayıp veri yapısına

uygulanabilirken, bazı yöntemler ise özel yapılara uygulanmak için sınırlandırılmıştır (Little ve Rubin, 2002).

2.3.2. Rassal olarak kayıp (missing at random, MAR, ROK)

X ve Y gibi değişkenleri arasından X'in yanıt olasılığının Y'ye bağlı fakat Y'nin yanıt olasılığının X'e bağlı olmaması halinde, bu durumdaki kayıp veri rassal kayıp veri olarak adlandırılır (Allison, 2001). $Y_{gözlenen}$, Y veri setinin gözlenebilen kısmını, Y_{eksik} de veri setinin kayıp kısmını göstersin. Bu mekanizma TROK'tan daha az sınırlayıcı bir varsayımdır; kayıp olma durumu $Y_{gözlenen}$'e bağlıdır ve şöyle ifade edilir:

$f(M \mid Y, \phi) = f(M \mid Y_{gözlenen}, \phi)$ (tüm Y_{eksik} , ϕ ler için) (Little ve Rubin, 2002).

Rastgele seçimdeki kayıp veri mekanizması kayıp bilginin olasılığı diğer değişkenler üzerinde de geçerli olduğu zaman ortaya çıkar (Enders, 2011).

M'nin dağılımı Y veri matrisindeki kayıp değerlere bağlı ise bu mekanizma Rassal Olarak Kayıp (ROK) diye adlandırılır. Bu mekanizmada belki de en basit veri yapısı, bazı birimleri kayıp olan tek değişkenli rassal örneklemdir. $Y=(y_1,\ldots,y_n)^T$ tanımlansın. Bu matriste y_i , i. birim için rassal değişkenin aldığı değeri belirtir. $M=(M_1,\ldots,M_n)$ olarak tanımlanabilir; M_i=0 olması i. birime ait verinin varlığını, yani bu verinin gözlendiğini, M_i=1 olması ise i. birime ait verinin kayıplılığını belirtir. (y_i,M_i) ortak dağılımının birbirinden bağımsız olduğu varsayılabilir yani birinin gözlenme olasılığı diğer birimlerin Y veya M değerlerine bağlı değildir. Bu durumdan hareketle;

$$f(Y,M \mid 0,\phi) = f(Y \mid Y\phi)\, f(M \mid Y,\phi) = \prod_{i=1}^{n} f(y_i \mid \phi) \prod_{i=1}^{n} f(M_i \mid y_i, \phi)$$

olduğu gösterilir. Bu denklemde $f(y_i, |\phi)$ ifadesi y_i'nin bilinmeyen parametrelerinin ϕ ile indekslenmiş olasılık yoğunluk fonksiyonudur ve $f(M_i |y_i, \phi)$ y_i'nin kayıp

olması olasılığının Pr $(M_i=1 |y_i, \phi)$ olduğu ikili M_i göstergesi için Bernoulli

Dağılımının olasılık yoğunluk fonksiyonudur. Eğer kayıp olma Y den bağımsız ise yani

Pr$(M_i=1 |y_i, \phi)$ y_i'den bağımsız katsayı ise, kayıp veri mekanizması TROK olarak

nitelendirilir. Eğer bu mekanizma y_i' ye bağımlı ise bu mekanizma ROK olur. Çünkü bu durumda ROK kayıp olan y_i değerlerine bağımlıdır ve bu kayıp değerlerin az sayıda olduğu varsayılır (Little ve Rubin, 2002; Bal, 2003). Özetle bu varsayım altında kayıp verinin olasılık dağılımı gözlenen/gözlenebilen değişkenlerin bilgisi ile ifade edilebilir. Bu tür durumlarda kayıplar ele alınmadan yapılan analizler yanlı olacaktır (Yozgatlıgil vd., 2011).

ROK mekanizmasına uygun kayıp veri yapıları birkaç modelleme altında toplanabilir. Kayıp veriler gözlenebilen değişkenlerle birlikte bağımlılığı matematiksel olarak ifade edilebilen yapılara sahiptir. Bunlardan bazıları;

- Tek değişkende kayıp olma durumu
- Aynı zamanda meydana gelen kayıp veri durumu
- Rastgele meydana gelen kayıp veri durumu
- Dosya eşleştirme problemi
- Monoton kayıp olma durumudur (Little ve Rubin, 2002).

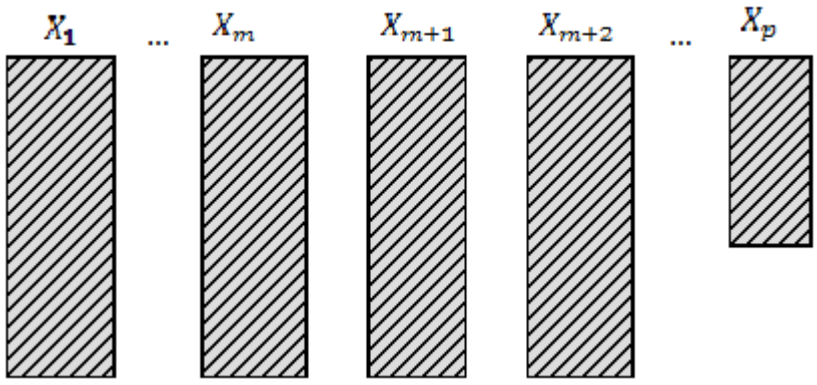
Tek değişkende kayıp olma durumu (tek değişkenli cevapsız model-univariate nonresponse)

İstatistiksel olarak sık karşılaşılan bu problemde veri matrisindeki değişkenlerden birinde kayıplılık söz konusudur. Bu problemin matris formu ise aşağıda gösterilmektedir. Şekil 2-4.'de i= 1, 2, 3 . . . , n ve j= 1, 2, 3 . . . , p için p. değişkende

gözlenen verilerin sayısı m olmak üzere x_{1j}, x_{2j}, x_{3j} , . . . , x_{mj} tamamıyla gözlenen, geri kalan $x_{(n-m)j}$ ise gözlenemeyen verileri göstermektedir. Şekil 2-5.'de verilen matriste, p. değişkenin ilk üç gözlemden sonra kayıp veri problemi olduğu görülmektedir ve (n-3) tane gözlem kayıptır.

$$A = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} & \cdots & x_{1p} \\ x_{21} & x_{22} & x_{23} & x_{24} & \cdots & x_{2p} \\ x_{31} & x_{32} & x_{33} & x_{34} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & \vdots & \cdots & ? \\ x_{n1} & x_{n2} & \cdots & \cdots & \cdots & ? \end{bmatrix}$$

Şekil 2-4. Tek değişkenli kayıp olma durumunun veri matrisi ile gösterimi



Şekil 2-5. Tek değişkenli kayıp olma durumunun veri matrisi ile gösterimi

$$A_{4x5} = \begin{bmatrix} 2 & 3 & 5 & 7 \\ 1 & 2 & 4 & 3 \\ 5 & 2 & 7 & ? \\ 1 & 3 & 9 & ? \\ 8 & 7 & 6 & ? \end{bmatrix}$$

Şekil 2-6. Tek değişkenli kayıp olma durumuna ait örnek

Şekil 2-6.da 4x5 tipindeki matris için değişkendeki gözlenen verilerin sayısı 17 ve x_{35}, x_{45}, x_{55} verilerinin gözlenemeyen değer olduğu görülmektedir.

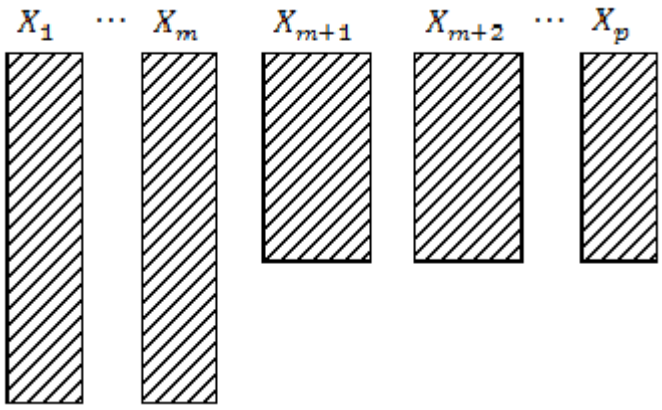
Bu tür probleme genellikle kayıp nokta problemi denir (Yazıcı, 2005). Birçok farklı etmenin ölçülmek istenildiği bir durumda, herhangi bir etmenin gözlemlenmesindeki kayıplılıkların ortaya çıkardığı kayıp veri mekanizmasıdır.

Aynı zamanda meydana gelen kayıp veri durumu (çok değişkenli cevapsız model-multivariate two patterns)

Şekil 2-7.'deki tek değişken kayıp X_p'nin yerine m tane gözlenmiş ve geri kalan X_{m+1}, X_{m+2} , . . . , X_p değişkenlerde kayıp olduğuna aşağıdaki şekilde gösterilebilecek yaygın tipte bir örnek elde edilir.

$$A = \begin{bmatrix} x_{11} & \cdots & x_{1m} & x_{1(m+1)} & x_{1(m+1)} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2m} & x_{2(m+1)} & x_{2(m+2)} & \cdots & x_{2p} \\ x_{31} & \cdots & x_{3m} & x_{3(m+1)} & x_{3(m+2)} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & \vdots & ? & ? & ? \\ \vdots & \vdots & \vdots & \vdots & ? & ? & ? \\ x_{n1} & \cdots & x_{nm} & \cdots & ? & ? & ? \end{bmatrix}$$

Şekil 2-7. Aynı zamanda meydana gelen kayıp olma durumunun veri matrisindeki gösterimi



Şekil 2-8. Aynı zamanda meydana gelen kayıp olma durumunun modellenmesi

$$A_{5x6} = \begin{bmatrix} 4 & 2 & 5 & 4 & 5 \\ 6 & 7 & 3 & 9 & 7 \\ 3 & 1 & ? & ? & ? \\ 4 & 5 & ? & ? & ? \\ 3 & 2 & ? & ? & ? \\ 2 & 1 & ? & ? & ? \end{bmatrix}$$

Şekil 2-9. Aynı zamanda kayıp olma durumuna ait örnek

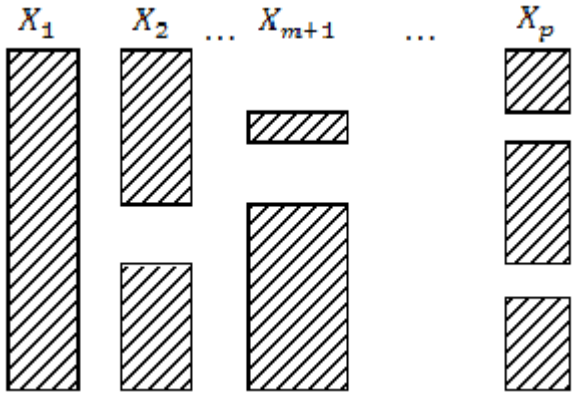
Şekil 2-9.'daki veri matrisinde

$x_{33}, x_{34}, x_{35}, x_{43}, x_{44}, x_{45}, x_{53}, x_{54}, x_{55}, x_{63}, x_{64}, x_{65}$ değişkenleri gözlenemeyen değerlerdir.

Örneklem araştırmalarında birim cevapsızlığı buna örnek olarak verilebilir. Yapılan bir ankette, örneklem bireylerini oluşturan grup, anketi iletişimsizlik, reddetme ya da başka bir sebeple kayıp tamamlamış olabilir. Bu durumda araştırma birimleri kayıp değişkenlerdir ve tamamıyla gözlenmiş değişkenler, cevap verenler ile cevap vermeyenler olmak üzere gruplanmış değişkenlerden oluşmaktadır. Anket düzenleme, cevap verenler ile cevap vermeyenler arasındaki ölçümleri içermektedir (Yazıcı, 2005).

Rastgele meydana gelen kayıp veri durumu (general)

Anket maddelerinde belirli maddelerin kayıp değerleri rastgele şekilde dağılım göstermişse bu durum ise şekil 2-10.'daki gibi şematik hale getirilebilir.



Şekil 2.10. Rastgele meydana gelen kayıp veri durumunun modellenmesi

$$A_{5x6} = \begin{bmatrix} 1 & 2 & ? & ? & 5 \\ 2 & 3 & ? & ? & ? \\ 3 & 4 & ? & ? & 2 \\ 4 & 5 & 2 & ? & ? \\ 3 & ? & 5 & ? & 4 \\ 2 & 1 & 2 & ? & ? \end{bmatrix}$$

Şekil 2-11. Rastgele meydana gelen kayıp veri durumuna ait bir örnek

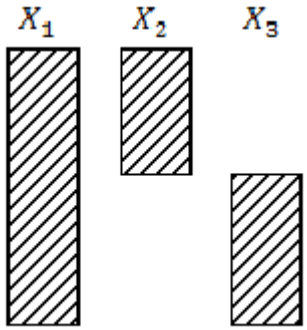
Şekil 2-11.'deki 6x5 tipi veri matrisinde kayıp verilerin rastgele dağılımı söz konusudur.

Dosya eşleştirme problemi (file matching)

Dosya eşleştirme problemi iki gruplu değişkenlerin asla birlikte gözlenemediği bir durum olarak da bilinmektedir. Kayıp verinin miktarı çok olduğunda, bazı verilerin birlikte gözlenemiyor olması ihtimali ortaya çıkar. Şekil 2-12.'deki, iki farklı kaynaktan alınan verileri birleştiren örnek için bir uç problemi göstermektedir. Bu modelde, X_1, tam gözlenmiş veri kaynaklarından tam gözlenmiş veri kaynaklarından oluşan gözlemlerin kümesini temsil etmektedir. X_2 değişkenler kümesi, birinci veri kaynağı için gözlemlenmiş ama ikinci veri kaynağı için gözlemlenememiştir. X_3 değişkenler kümesi de ikinci veri kaynağı için gözlenmiş fakat birinci veri kaynağı kümesi için gözlenememiştir (Yazıcı, 2005).

$$A = \begin{bmatrix} x_{11} & x_{12} & ? \\ x_{21} & x_{22} & ? \\ x_{31} & x_{32} & ? \\ x_{41} & ? & x_{43} \\ x_{51} & ? & x_{53} \\ \vdots & \vdots & \vdots \\ x_{n1} & ? & x_{n3} \end{bmatrix}$$

Şekil 2-12. Dosya eşleştirme probleminin veri matrisi gösterimi



Şekil 2-13. Dosya eşleştirme probleminin modellemesi

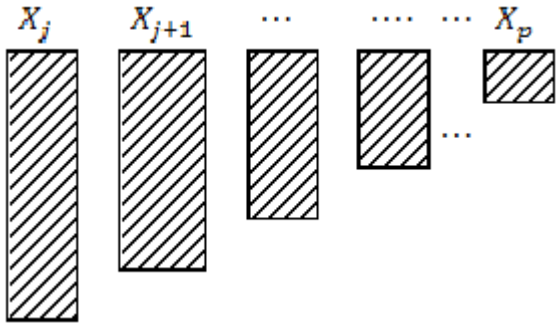
Bu tip kayıp veri modellemesinde bir kaynaktan alınamayan cevapları diğer kaynaktan alınması söz konusudur.

$$A_{6x3} = \begin{bmatrix} 2 & ? & 1 \\ 4 & ? & 3 \\ 3 & 2 & ? \\ 4 & 2 & ? \\ 5 & 2 & ? \\ 6 & 4 & ? \end{bmatrix}$$

Şekil 2-14. Dosya eşleştirme problemine ait sayısal örnek

Şekil 2-14.'te 6x3 tipindeki matriste x_{12}, x_{22}, x_{33}, x_{43}, x_{53}, x_{63} değerleri gözlenmeyen değerlerdir.

Monoton kayıp olma durumu (monotone)



Şekil 2-15. Monoton kayıp olma durumunun modellenmesi

Monoton artan kayıp veri durumunda Şekil 2-15.'te gösterildiği gibi X_{j+1},X_{j+2},..., X_p değişkenleri X_i'nin kayıp olduğu yerlerde ve her j= 1, 2, 3, . . . , m-1 olacak şekilde sıralanırlar.

$$A_{nxp} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & ? \\ x_{31} & x_{32} & ? & ? \\ x_{41} & ? & ? & ? \\ x_{51} & ? & ? & ? \end{bmatrix}$$

Şekil 2-16. Monoton kayıp olma durumunun veri matrisi gösterimi

$$A = \begin{bmatrix} 1 & 3 & 1 & 4 \\ 7 & 6 & 2 & ? \\ 4 & 1 & ? & ? \\ 5 & ? & ? & ? \\ 2 & ? & ? & ? \end{bmatrix}$$

Şekil 2-17. Monoton kayıp olma durumuna ait sayısal örnek

Şekil 2-17.'deki veri matrisinde ilk sütunda kayıp veri bulunmazken ikinci sütunda iki adet üçüncü sütunda üç adet, dördüncü sütunda dört adet bulunmaktadır. Şekil incelendiğinde sütun sayısı arttıkça kayıpların arttığı görülmektedir.

Kayıp veri mekanizması Y= (y_{ij}) veri setinde ve M=(m_{ij}) kayıp veri gösterge matrisinde y belli iken M nin şartlı dağılımıdır ve $f = (M \mid Y, \phi)$ şeklinde ifade edilir. Bu fonksiyonda φ bilinmeyen parametrelerdir. Kayıplılık durumun Y nin değerlerine bağlı değilse bu fonksiyon bütün Y, φ ler için $f = (M \mid Y, \phi) = f(M \mid \phi)$ ile gösterilir ve bu veriler TROK olarak adlandırılır. Bu mekanizmada kayıp olma durumu verilerin değerlerine bağlı değildir (Little ve Rubin, 2002). Yani rastgele seçimdeki bir kayıp (TROK) bir verideki kayıpsızlığın diğer değişken ölçütleriyle ilgisiz olduğu zaman

ortaya çıkar (Enders, 2011). Özetle, veri setlerini oluştururken tamamen istek dışı oluşan kayıplardır. Bir anket çalışmasında soruyu görmeyip cevaplamama veya verilerden bazılarının kaybolması TROK mekanizması olarak nitelendirilebilir (Sezgin & Çelik, 2013). Örneğin, anket çalışmasında anketörün soru sormayı unutması, bir koşu yarışmasında atletin kural ihlali nedeniyle bitişi görememesi, laboratuvar deneyinde numunenin yere düşmesi durumlarında oluşan kayıplar da TROK yapısını oluşturur (Yozgatlıgil vd., 2011).

Bir anket çalışmasında örneklem bireylerini oluşturan grubun sorulan soruları bilerek atlaması veya yanlış cevap vermesi ROK olarak nitelendirilebilir (Sezgin ve Çelik, 2013). Anket çalışmasında gelir sorusuna ilişkin cevaplar düşünülsün. Geliri çok yüksek kişiler gelirlerini doğru cevaplamak eğilimindedirler. Zengin kişilerin gelirlerinin gerçekte çok olması beklendiğinde böyle bir çalışma için ortalama gelir hesabı gerekenden daha küçük çıkacaktır, yani yanlı olacaktır. Bu örnekte kişinin gelir sorusu karşılığında verilen cevaplardaki kayıplar kişinin varlıklı olmasına (gözlenen/gözlenebilen) bağlıdır, gelir miktarına (kayıp verinin kendisine) bağlı değildir. Bu durum da ROK için örnek gösterilebilir (Yozgatlıgil vd , 2011).

TROK ve ROK tanımlamalarını değişik bir yaklaşımla vermek de olasıdır. Buna göre TROK tanımı, kayıplılığın değişkenlerin değerleri ile ilişkili olmaması şeklinde yapılabilir. Bu amaçla yaş ile gelir düzeyi değişkenlerini ele alalım. Veri kayıplılığın yaş ve gelir düzeyi değerleri ile ilişkili olmaması durumunda kayıp veri yapısının TROK olduğu söylenebilir. Örneğin, yaş kayıp veri içermez iken gelir düzeyi kayıp veri içersin. Araştırmaya giren tüm kişiler için gelir düzeyinin veride yer alma olasılığı; kişilerin yaş ve gelir düzeylerini dikkate almaksızın eşit ise verinin TROK olduğu söylenir. Gelir düzeyinin veride yer alma olasılığı, yaş ile ilişkili ancak gelir düzeyi ile ilişkisiz ise veri ROK'tur (Alpar, 2003).

2.3.3. İhmal edilemez kayıp/rastlantısal olmayan kayıp (noignorable, NI, IE)

ROK ve TROK olmayan süreçler rastlantısal olmayan kayıp olarak tanımlanır. Yani kayıp olma durumu rassal değildir ve veri tabanındaki bir diğer değişkenden

tahmin edilemez (Allison, 2001). IE mekanizması kayıp veri olasılığının diğer değişkenlerle olan ilişkisi doğrultusunda ortaya çıkmıştır (Enders, 2011). Değişken üzerindeki kayıp ihtimalini ortaya çıkaran veri, yine o değişkenin kendi değerinin fonksiyonu olduğu zaman IE grubuna girer. IE değişkenine örnek olarak maaş verilirse, maaşı yüksek olan kişiler düşük maaşlı kişilere göre daha az maaşlarını açıklama eğilimindedirler. Bu sebeple maaş için kayıp veri ihtimali maaşın kendi değeri fonksiyonudur. Bu durumda kayıp veri IE'dir. Kayıp olma durumu değişkenlerimize bağlı olmadığından kayıp veriler hakkında tahminlerde bulunmak hatalara yol açacaktır, dolayısıyla bu durum yanlı sonuçlara sebebiyet verecektir (Allison, 2001). Örneğin anketteki bir sorunun yanlış sorulmasından dolayı doğru cevabın çözülemediği durumlardaki kayıp durumu IE olarak tanımlanabilir (Sezgin ve Çelik, 2013). Veri setlerinde, kayıp değerlerin rastgele koşullar altında olup olmadığı belirlendikten sonra bu kayıplılığın giderilmesine yönelik çözümlemeler yapılmasını gerektirir. Veri tabanındaki değerlerin ROK veya TROK olması durumunda kullanılacak yöntemler de farklı olacaktır.

Özetle, bir veri TROK grubuna dahil değilse ya rassal olarak kayıptır ya da rassal kayıp değildir. Bir veri üzerindeki veri kaybı ihtimalini oluşturan değişken, bu değişken ile ilgisi olmadığı ve veri tabanındaki diğer değişkenlerin değeri ile ilgili olduğu durumlarda, ROK yani rassal (tesadüfi) olarak kayıptır. Örneğin ROK yaklaşımının altında şunu söyleyebiliriz; Y (öz yeterlilik) değişkeni için kayıp veri ihtimali X (ırk) değişkenine bağlı olabilir ama X kontrol altına alındığı zaman Y'nin değeri ile ilgili değildir. Buna zıt bir örnek ise Y'nin değeri maaş değişkeni ile ırk değişkeni arasında kurulabilir. Irk değişkeni kontrol altına alındığı zaman maaş değişkeni için kayıp veri ihtimali akla gelebilir ama maaş değişkenindeki kayıp veri ihtimali yine maaşın kendisi ile ilgilidir, çünkü yüksek maaşlı kişiler maaşlarını söylemek istemeyebilirler (Cheema, 2012). ROK yaklaşımı TROK yaklaşımının esnetilmiş halidir. TROK yaklaşımına göre kayıp veri ihtimali kendi değişkeni ya da diğer değişkenlerle ilgili olmamalıdır. Veri TROK ya da ROK grubunda ise tahmin sürecinin bir parçası olarak kayıp veri mekanizmasını örneklemeye gerek yoktur. Başka bir deyişle, TROK veya ROK verisine bir kayıp veri mekanizmasıyla yaklaşıldığı zaman herhangi bir analiz metodu kullanılabilir. Ama verimiz IE ise analiz konusunda

çok da rahat davranılamaz, çünkü IE verisi ile ilgili kayıp veri ihtimali süreçten bağımsız şekilde devam etmez. Ayrıca IE verisi ile ilgili kayıp veri ihtimalinin ele alınabilmesi için gözlem ve bu gözlem sonucunda ele geçecek ön bilgiye ihtiyaç vardır. Bütün bu tartışma sonucunda söylenmesi gereken IE verisi için kayıp veri ihtimali değerlendirilirken sınırlayıcı ve genelleyici açıklamalardan kaçınmak gerekir. Bu noktada doğal olarak bu mekanizmaları yanlış bir şekilde ele alırsak ne olur sorusunu sorarız. IE verisi yanlış bir şekilde ROK veya TROK olarak ele alınırsa bu da kayıp veri sürecinin örneklendirilemediğini gösterir ve bu analiz metodunun parametreleri doğru sonuç çıkarmayacaktır. TROK verisi yanlış bir şekilde ROK veya IE olarak değerlendirilirse bu da gereksiz bir karışıklık ortaya çıkarır. ROK verisine IE muamelesi yapmak da TROK verisine IE verisi muamelesi yapmakla aynıdır. ROK verisine TROK muamelesi uygularsak kayıp veriyi aşırı şekilde basitleştiriyoruz demektir. Ama yine de parametreleri toplam örnekleme uyarlayamayız (Allison, 2001). Ama TROK yine de ROK ve IE'nin gölgesinde kalabilir. Öte yandan TROK ve ROK mekanizmaları ileri seviyede benzerlik ve çoklu atama bakımından daha doğru tahminler ortaya çıkarabilir (Enders, 2011).

2.4. Kayıp Veri Sürecinde Rastgeleliğin Sorgulanması

Baygül (2007) ve Alpar (2003)'e göre kayıp veri analizinde kullanılacak yöntemin belirlenmesi için kayıp veri sürecinin rastgeleliğinin aşağıdaki yöntemlerle irdelenmesi gerekir:

1. Veri setindeki bir değişkene ilişkin gözlemler kayıp veri içerenler ve içermeyenler olarak iki gruba ayrılmalı ve ilgilenilen diğer değişkenlerin aldığı değerler açısından bu iki grup arasında anlamlı bir fark olup olmadığı araştırılmalıdır. Bu araştırma iki ortalama arasındaki farkın anlamlılığı test eden ‘‘t testi’’kullanılarak yapılabilir. Anlamlı fark, rastgele olmayan kayıp veri sürecinin varlığını gösterir.
2. Veri setindeki değişkenler kayıp değer içeren ve içermeyenler olmak üzere iki gruba ayrılır, tam veriler 1, kayıp veriler 0 olarak kodlanır ve bu değişkenler arasındaki Pearson korelasyon katsayısı hesaplanır. Bulunan korelasyon

katsayıları her bir değişken çifti için kayıp veriler arasındaki ilişki miktarının derecesini belirtir. Küçük korelasyon katsayısı rastgeleliği işaret eder. Ancak rastgele olmayan kayıp veri sürecini tanımlamak için kesin bir kural söz konusu değildir. Diğer taraftan, korelasyon katsayısının anlamlılık testleri ise rastgeleliğin derecesini tanımlamada oldukça tutucu bir kestirim sağlamaktadır. Eğer rastgelelik tüm değişken çiftleri için söz konusu ise, araştırıcı, kayıp veriyi TROK olarak sınıflandırabilir. Ancak, bazı değişken çiftleri arasında anlamlı korelasyonlar var ise veri ROK olarak düşünülebilir.

3. Little (1998)'ın TROK testi rastgeleliğin araştırılmasında sıklıkla kullanılan bir χ^2 testidir. p< 0, 05 olması durumunda veri yapısının TROK olmadığı sonucuna varılır. Bu test aşağıda formülle açıklanmıştır:

$\chi^2_{MCAR}=\sum_{her\ bir\ tek\ yapı}$(birim sayısı * ortalamadan elde edilen Mahalanobis D^2 uzaklığı)

Little (1998)'ın TROK testi SPSS uygulamasında sık kullanılır. Buradaki kayıp veri terimi, belirli bir değişken üzerindeki veri gruplarının kayıp olan ya da olmayan değerlerine göre gruplandırılmasına karşılık gelir. Örneğin iki değişkenli bir veri grubunu ele alalım. Buradaki X, Y değişkenleri için dört tane model öngörülür. Hem X hem Y kayıptır, sadece X değişkeni kayıptır, sadece Y değişkeni kayıptır veya ne X ne de Y kayıptır. Y'nin başarı skoru ve X'in cinsiyete karşılık geldiği bir örneği düşünelim. Bir araştırmacı cinsiyet bazlı araştırma yaparken kayıp verilerin erkekler ve kadınlar arasında değişip değişmediğini analiz edebilir. Kadınlar için başarı skoru bazlı bir araştırmada kayıp veri yüzdesi %80 iken erkekler için %20 ise buradaki veri grubunu TROK içinde ele alamayız. Eğer cinsiyet değişkenini kontrol altına alırsak, başarı skorundaki kayıp veri değeri başarı skorunun kendisiyle ilgili değildir ve böylece veri grubu ROK grubunda ele alınır (Cheema, 2012). Test edilebilirlik açısından TROK varsayımı ROK varsayımına göre daha kolay test edilebilir. Bunun da sebebi ROK yaklaşımının ''bir değişken üzerindeki kayıp veri değerinin o değişkenin kendi değeriyle alakasız'' olduğu varsayımına dayanmasıdır. Bazı durumlarda bu yaklaşım ya da varsayım geçersiz olarak kabul edilir. Örneğin maaşları yüksek olan kişiler maaşlarını belirtmeme eğilimi sergilerler. Bundan dolayı maaş değişkeni üzerindeki kayıp veri değeri, verinin kendi değeriyle yakından ilgilidir ve buradaki kayıp veri mekanizması ne TROK ne de ROK grubuna girer (Groves vd, 2004; Aktaran: Cheema,

2012). TROK mekanizması verilerin örneklemi temsil etme gücünü belirlemede oldukça etkindir.

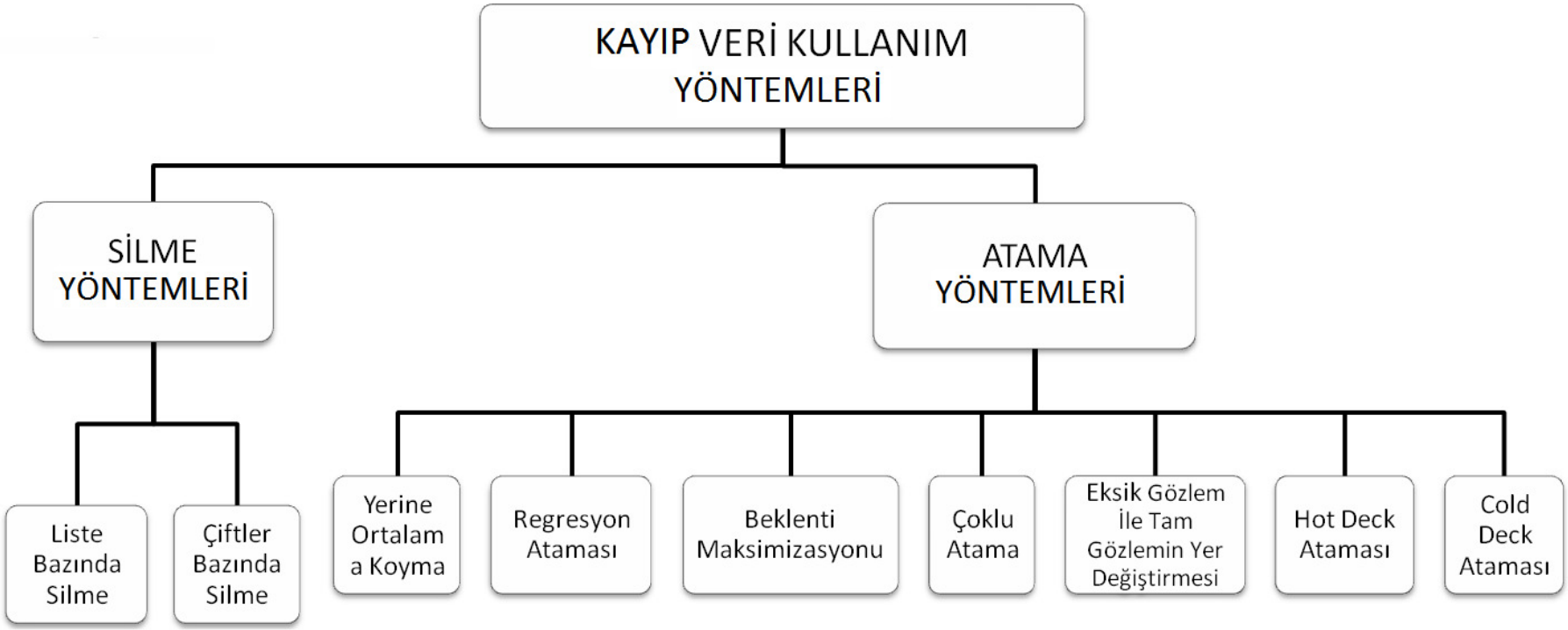
2.5. Kayıp Veri Ele Alma Yöntemleri

Kayıp veri sorununa çözüm üretmek amacıyla kullanılan yöntemlerin başlıcaları ele alınacaktır.

Tabachnick ve Fidell (2001) kayıp değerler bazı değişkenlerde yoğunlaşıyorsa ve bu değişkenler analizler açısından yüksek korelasyona sahip bir değişken ise, kayıp değer içeren bu değişkenin rahatlıkla veri setinden çıkarabileceğini belirtmektedir. Ancak kayıp değerler denek ve değişkenler boyunca dağılabilir ve bu değişkenlerin çıkarılması önemli denek kayıplarına neden olabilir.

İşte bu durumda araştırmacılar kayıp verilere ilişkin kestirimler yapma ve bunları analizlerinde kullanma gereği duymaktadırlar. Bu kestirimleri yapmanın en yaygın yöntemlerinden biri ‘‘geçmiş bilgileri kullanmak’’ yöntemidir. Bu yöntemde araştırmacılar daha önceki gözlemlenmiş bilgilerden yola çıkarak kayıp değerlere yeni değer atama yoluna giderler. Tabachnick ve Fidell (2001)’e göre bu yöntem araştırmacıların o alanda sahip olduğu bilgi birikimi ve uzmanlığına dayanarak kayıp değerler hakkında iyi tahminler yaparak değer atamasıdır. Araştırmacı uzun yıllardır o alanda çalışıyorsa, çalıştığı örneklem büyükse ve kayıp veri gözlenen değer sayısı az ise bu yola başvurulabilir.

Araştırmacılar silme tekniklerinden Liste veya Durum Bazında Silme (listwise or casewise data deletion-LD or CD) ve Çiftler Bazında Silme (pairwise data deletion-PD) gibi kayıp değerler problemini ortadan kaldıracak yöntemleri kullanabilirler. Bunların dışında en sık kullanılan atama teknikleri: Kayıp Gözlem İle Tam Gözlemin Yer Değiştirmesi (Case Substitution), Yerine Ortalamayı Koyma (Mean Substitution), Regresyon Ataması (Regression Imputation), Hot Deck Ataması (Hot Deck Imputation), Cold Deck Ataması (Cold Deck Imputation), Beklenen Maksimizasyon Yaklaşımı (Expectation Maximization-EM-approach) ve Çoklu Atama (Multiple Imputation) olarak sıralanabilir.



Şekil 2-18. Kayıp veri kullanım yöntemlerinin sınıflandırılması

2.5.1. Silme yöntemleri

Liste veya durum bazında silme (tam gözlemlerin kullanılması yöntemi/ listwise deletion-LD)

Bu yöntemde sadece, tam gözlemler dikkate alınır, kayıp gözlemler dikkate alınmaz. Bu yöntemde kayıpsız gözlemler dikkate alındığından kayıp veri sayısının az olduğu durumlarda kullanılması tavsiye olunur. Çok kullanılan bir yöntem olmasının yanında kayıp veri yapısının tamamıyla rassal kayıp (TROK) olması gerekmektedir (Kalaycı, 2008). Çünkü TROK olmayan bir yapı, sonuçları yanlılaştıracak rastgele olmayan elemanlara sahiptir. Diğer bir deyişle, TROK varsayımı altında, tam gözlemlerin orijinal gözlemlerin rastgele bir örneklemi olduğu ve dikkate alınmayan kayıp verili gözlemlerin kestirimleri ile yanlılaştırılamayacağı, ancak değişkenlere ilişkin güven aralıklarında artışların görüleceği belirtilmektedir (Alpar, 2003). Bu yöntemde sadece tam bilgi taşıyan veri gruplarıyla ilgilenildiği için "Tüm durum metodu" olarak da bilinir. Bu yöntem sayıca küçültülen örnekleminin temsil edildiği zaman iyi çalışmasına rağmen, örneklem ve güç arasındaki negatif ilişki yüzünden örneklemin küçük olmasında testler ve hipotezler üzerinde olumsuz etkiye sahiptir (Cheema, 2012). Bu yöntemin ROK veya IE durumlarında kullanılması yanlı sonuç verecektir. Tablo 2-1.'de sayısal bir örneğe yer verilmiştir:

Tablo 2-1. Liste bazında veri silmeye uygun örnek veri seti

DURUM	Değişken1	Değişken2	Değişken3
1	7	-	8
2	12	6	10
3	15	13	-
4	9	14	13
5	-	9	11
6	5	7	14

Tablo 2-1.'de 3 değişken ve 6 durum söz konusudur. 18 gözlem değerinin 3 tanesi kayıptır. Liste bazında veya durum bazında silme yöntemine göre kayıp değerler içeren 1-3-5 nolu durumların analiz dışı bırakılarak kayıp değer içermeyen 2-4-6 nolu durumlara dayalı hesaplamaların yapılması gerekmektedir.

Çiftler bazında durum silme (eldeki tüm verilerin kullanılması yöntemi / pairwise deletion-PD)

Kayıp veri içeren araştırmalarda eldeki tüm bilginin kullanılması yöntemine sıklıkla başvurulur. Bu yöntem, gerçekte kayıp verilerin yerine konulması ya da kayıp verilerin atanmış değerlerle doldurulması değil, eldeki tüm değerler yardımıyla dağılımı tanımlayıcı ölçülerin (ortalama, standart sapma vb) ve ilişkili ölçülerin (korelasyon, kovaryans) elde edilmesidir. İstatistiksel yazılımların çoğunda bu yönteme ilişkin menüler olduğu için araştırmacılar tarafından sıklıkla kullanılmış ve bu yaklaşımla elde edilen korelasyon ve kovaryans matrisleri (faktör analizi) girdi matrisi olarak kullanılmıştır (Alpar, 2003). Bu yönteme göre her değişken çifti için bütün durumları kayıpsız olanların kovaryans / korelasyon tahmini yapılacaktır. Örneğin Tablo 2-1.'de değişken 1 ve değişken 2 için kovaryans / korelasyon tahmini 2-3-4-6 nolu durumlara, benzer şekilde değişken 2 ve değişken 3 için kovaryans / korelasyon tahmini 2-4-5-6 nolu durumlardan hareketle, değişken 1 ve değişken 3 için ise kovaryans / korelasyon tahmini 1-2-4-6 numaralı durumlara dayanarak yapılacaktır.

Bu yöntemin liste bazında veri silme yönteminden farkı sadece kayıp veriye sahip şartların veya durumların yok edilmesidir. Örneğin; X, Y, Z araştırmamızdaki üç değişkenimiz ve bir gözlem durumunda da Z değişkeni üzerinde kayıp veri değeri olsun. Korelasyon hesabı gibi bir süreçte r_{xy}'nin hesabı için n sayıda gözlem kullanılırken r_{xz} ve r_{yz}'nin hesaplanmasında n-1 sayıda gözlemden yararlanılacaktır. Liste bazında veri silme yönteminde ise tüm korelasyonların hesabında sadece n-1 adet gözlem kullanılacaktı (McKnight vd, 2007; Aktaran: Cheema, 2012).

Eğer kayıp veri süreci TROK değil ise tam gözlemlerin kullanılması yönteminde olduğu gibi eldekilerin tümü yaklaşımında da elde edilen sonuçlar yanlı olabilmektedir. Bu yöntem kullanılan veriyi en fazlaya çıkarmasına rağmen bazı sorunları da beraberinde getirebilmektedir. Bu sorunların en belirgini, korelasyonların beklenen sınırların dışında ede edilmesi ve hesaplanan korelasyon matrisindeki diğer korelasyonlar ile tutarsız olmasıdır (Alpar, 2003). Korelasyon matrisindeki bu soruna bağlı olarak ortaya çıkan bir diğer sorun ise korelasyon matrisinin özdeğerlerinin negatif

değerli çıkmasıdır. Çiftler bazında durum silme metodunda her bir korelasyon için en iyi tahmin yapılmış olunur; çünkü elimizde var olan tüm veriler kullanılmaktadır. Bu açıdan liste bazında durum silmeye oranla daha etkili bir yöntem olduğu da söylenebilir. Ancak bu durum yukarıda bahsedilen korelasyonlar arası sorunlardan dolayı her zaman geçerli değildir. Yani liste bazında silme değişkenler arası korelasyon katsayısı yüksek olduğu durumlarda, çiftler bazında silme ise değişkenler arası korelasyon katsayısı düşük olduğu durumlarda daha etkin sonuçlar vermektedir.

2.5.2. Yaklaşık değer atama yöntemleri (imputation methods)

Kayıp gözlem ile tam gözlemin yer değiştirmesi (case substitution)

Bu yöntemde yedek bir denek atama durumu söz konusudur. Bu yöntemin en yaygın örneği, örnekleme giren ancak kendisine ulaşılamayan (verileri tamamen kayıp olan) ya da bilgilerinin çoğunun kayıp olduğu kişi ile örneklemde olmayan fakat örneğe göre ve kendisine ulaşılamayan kişi ile benzer özellikleri taşıyan başka bir kişi ile yer değiştirmesidir (Little ve Rubin, 2002; Alpar, 2003; Bal, 2003).

Kayıp gözlem yerine ortalamayı koyma (mean substitution)

Kayıp veri yerine sıkça kullanılan bir strateji, kayıp değer içeren değişkenin ortalamasını kayıp değerin yerine koymaktır (Little ve Rubin, 2002). Böylece kişiye ilişkin herhangi bir bilgi bulunmadan, herhangi normal dağılımlı bir değişken için değerlerin en iyi tahminini ortalamalar oluşturacağından tam veriler için elde edilen ortalama, tam verilerle elde edilen ortalamaya eşit çıkacaktır.

Bu yöntemin olumsuz yönleri ise:

- Varyansların olduğundan daha az çıkması,
- Kovaryansların negatif eğilimli olması,
- Asıl değişkene ait dağılım, ortalamaya eşit olan değerlerin eklenmesiyle oluştuğundan, atama elde edilen yeni değişkenin oluşturduğu dağılımın toplumu yeterince yansıtamayabilir (Little ve Rubin, 2002; Bal, 2003).

Araştırmacıların kullandığı bu işlem temel analizlerin gerçekleştirilmesinden

önce yapılır. Eğer araştırmacı başka bilgilere sahip değilse, ortalama değer atamak en iyi yöntemdir. Ancak bu deneklere değer atamakla ortalamanın değişmeyeceğini gözden kaçırmamak gerekir ve bu durumda varyansın bir miktar düşmesi söz konusudur; çünkü belki de gerçek değer ortalamaya tam olarak eşit değildir. Çok sayıda kayıp değer olmadığı takdirde, bu önemli bir problem değildir. Bu durumda olası bir kabul kayıp değerler için genel ortalama yerine grup ortalamasını kullanmaktır. Bu durum özellikle gruplar arası karşılaştırmalar içeren analizler yapılmak istendiğinde daha uygun bir işlemdir (Mertler ve Vannatta, 2005; Aktaran: Çokluk ve Kayrı, 2011). Eğer veriler yaklaşık olarak normal dağılım göstermekteyse ve ROK koşulu sağlanıyorsa, bu yöntem standartlaştırılmamış yanlı parametre tahmini olmayacaktır. Diğer taraftan çok sayıda cevaplayıcı bir değişkene ilişkin benzer skorlara sahipse, kayıp değerli değişkenler arasındaki kovaryans ve değişkenler arasındaki varyans daralacaktır. Varyansın daralması R^2 ve β gibi standardize olmuş katsayıların tahminini azaltacaktır. Azaltılmış varyans normal olarak standart hataları arttırmakta ve t oranlarını azaltmakla birlikte, yerine ortalama koyma yöntemi aynı zamanda örneklem büyüklüğünü arttıracaktır. Bu yapay olarak şişirilmiş örneklem büyüklüğünün kullanılması sonucunda, bu yöntem t oranlarını daha anlamlı hale getirecektir (Oğuzlar, 2001).

Değişkenlere atama yapılması sonucunda ortaya çıkan varyans ise;

n: toplam gözlem sayısı

$n^{(j)}$: tam gözlemlerin sayısı

$s_{ij}^{(j)}$: tam gözlemlerin varyansını temsil ediyorken;

Tam ve atanmış verilerin varyansı

$s_{jj}^{(j)}\left[\frac{n^{(j)}-1}{n-1}\right]$

şeklinde gösterilebilir. Burada tam ve atanmış veri setlerine ilişkin dağılımın varyansı

$\left[\frac{n^{(j)}-1}{n-1}\right]$

çarpımı derecesinde olduğundan kayıp bulunur (Little ve Rubin, 2002). Bunun nedeni ise değişkenlerdeki tüm kayıp gözlemler tek bir değer ile (ilgili değişkenin ortalaması ile) yer değiştirdiği için değişkenlerin gerçek dağılımından uzaklaşılır ve verideki gerçek varyans olduğundan küçük çıkar (Alpar, 2003).

Regresyon ataması (regression ımputation)

Tabachnick ve Fidell (2001)'e göre kayıp değerlere ilişkin kullanılabilecek bir diğer yöntem regresyon yaklaşımıdır. Kayıp değerlerin bağımlı değişken kabul edilerek işe başlanılan bu yöntemde bağımlı değişkenin yaklaşık değerini tahmin etmek için bir ya da birden fazla bağımsız değişken arasında bir bağıntı kurulmaya çalışılır. Bu bağıntı araştırmacıya basit bir ortalama atamaktansa daha fazla bilgi vermesinin yanında geçmiş bilgileri kullanma ve ortalama değer atama yöntemine göre de objektiflik sağlar.

Regresyon yönteminin amacı, bir ya da daha fazla bağımsız değişken yardımıyla bağımlı değişken değerinin test edilmesidir. Regresyon atama yönteminde, bağımlı değişken kayıp gözlemli değişken; bağımsız değişkenler ise diğer değişkenlerdir. Bu yöntemin özellikle kayıp verinin sayısının orta düzeyde olduğu ve veride yaygın bir dağılım gösterdiği durumlarda kullanılması önerilmektedir. Bu yöntemin kullanılabilmesi için bağımlı değişkenle bağımsız değişkenler arasında ilişkinin oldukça yüksek olması gereklidir (Alpar, 2003; Kalaycı, 2008). Bu atama tekniği, bağımsız değişkenlerdeki kayıp değerin ataması için kullanıldığında, bu durum çoklu doğrusal bağlılığa katkıda bulunacaktır; çünkü kayıp değerler için atamada bulunulan değerler modeldeki diğer değişkenlerle ilişkili olacaktır. Regresyona dayalı atama tekniğinde modele dahil edilmeyen diğer değişkenlerin kullanılması da mümkündür. Fakat bu durumda daha zayıf tahmin değerleri elde edilecektir (Oğuzlar, 2001).

Regresyon ataması kullanmanın bir avantajı, atamada bulunacak kayıp değer içeren değişkenin her bir kayıp değeri için farklı bir bağımsız değişkenler kümesini kullanmasıdır. Bu yaklaşımın yerine ortalamayı koyma yaklaşımından bir avantajı, kayıp değer içeren değişkenlerin varyans ve kovaryanslarını korumasıdır. Çünkü bir değişkenin kayıp her bir durumu, diğer değişkenlerin değerlerine bağlıdır ve her seferinde farklı bir tahmin değerini verecektir (Oğuzlar, 2001). Bu yöntemin olumsuz yanları ise şöyle sıralanabilir:

- Değişkenler arasındaki ilişkiden yola çıkıldığı için atama sonucunda veride zaten var olan ilişki daha da kuvvetlenmiş olur.

- Kullanımı arttıkça, elde edilen veri kendine özgü bir yapıya ulaşabilir ve genellenebilirliği azalır.
- Regresyon yöntemi sonucu elde edilen kestirimin sınırları, veride var olan sınırların ötesinde bulunabilecektir. Örneğin bağımlı değişken değerleri 0 ile 20 arasında değişiyor ise bu aralıktaki bir değer için kestirim değeri 21 olarak bulunabilir. Bu nedenle bazı düzeltme işlemlerinin yapılması gerekli olabilir (Alpar, 2003).

Hot deck ataması (hot deck ımputatıon)

Bu yöntem daha çok sosyal bilimlerde kullanılmasına rağmen diğer yöntemlere oranla daha az gelişmiş gibi görünür. Bu yöntemde kayıp veri için birçok durum havuzu seçilir ve buna donör (verici) havuzu denilir. Donör havuzundaki durumlara ait değişkenler arasında benzerlik vardır ve içinden rastgele bir durum ele alınır. Rastgele seçilen duruma ait veri, kayıp veriye ait kayıp değeri ile değiştirilir. Bir başka yöntem de en yakın donör ile değişim yapılır. Bu yöntemde doğal değişim açık bir şekilde göz ardı edilir. Hot deck imputasyon yönteminin en iyi şekilde çalışması için değişkenlerin büyük boyutlu örnekleme ait olması gerekmektedir. En uygun donörü seçmek için alıcı benzer durumlarla eşleştirilir, bu eşleştirme sırasında göz önünde bulundurulan iki kriter vardır; harici değişkenin ayarlanan değişkenle bağlantılı olup olmadığı ve harici değişkenin kayıp veri içeren ya da içermeyen ikili değişkenle alakalı olup olmadığıdır (Cheema, 2012).

Hot Deck Atama yöntemi uzun bir kullanım tarihine sahiptir. Bu yöntem liste bazında veri silme, çiftler bazında veri silme, yerine ortalama koyma yöntemlerinden üstün bir tekniktir. Hot Deck Atama yönteminin avantajları arasında kavramsal basitliği, değişkenlerin ölçüm düzeylerinin korunması (kategorik değişkenler kategorik olarak, sürekli değişkenler sürekli olarak kalır) ve tamamlanmış veri matrisi elde edilir. Tamamlanış veri matrisi sayesinde de standart istatistiksel analizler uygulanabilir. Bu yöntemin en önemli dezavantajı ‘‘benzerlik’’kavramının tanımlanmasındaki güçlüktür. Bu nedenle Hot Deck yöntemi kayıp veriler için standart bir yol sağlamamaktadır. Bu benzerliğin belirlenmesi için verici (donör) durumların seçimini başarabilecek bir

yazılım gereklidir (Oğuzlar, 2001).

Tablo 2-2. Hot deck atamasına uygun örnek veri seti

DURUM	Değişken1	Değişken2	Değişken3	Değişken4
1	5	3	2	-
2	4	1	5	2
3	2	4	2	1

Tablo 2-2. incelendiğinde Durum 1 için 4 nolu değişkenin kayıp değer içerdiği görülmektedir. Hot Deck Atama yöntemi tam gözlemli veriler için kayıp değere en çok benzediği düşünülen veri değerini atar. Tablo2-2. için tam gözlemli durumlar 2 ve 3'tür. Bu durumların değişkenler için değerleri incelendiğinde durum 1'in durum 3'e daha çok benzediği görülmektedir. Dolayısıyla durum 1'deki 4. değişkene ait kayıp veri değerine durum 3'teki 4. değişkenin değeri atanır.

Cold deck ataması (cold deck ımputation)

Bu yöntemde; önceki çalışmalardan, aynı çalışmanın bir önceki uygulamasında ya da dış kaynaklardan elde edilen sabit bir değer, ilgili değişkendeki tüm kayıp veriler yerine atanır. Bu yöntemin kayıp gözlem ile ortalamanın yer değiştirmesi yönteminden tek farkı, atanacak değerin kaynağıdır. Araştırmacı, dış kaynaklardan elde edilen değerin, eldeki veriler yardımıyla elde edilen ortalama gibi bir değerden daha geçerli olabileceğinden emin olmalıdır. Bu yöntem, ortalama atama yöntemine benzer olumsuzluklara sahiptir (Alpar, 2003).

Beklenti maksimizasyon yaklaşımı (expectation maximization-EM-approach-BM)

BM imputasyonu iki adımlı bir yöntemdir. İlk adımda başlangıç değerler kayıp veriye aktarılır; ikinci adımda da bu başlangıç değerleriyle oluşturulan beklentiler yükseltilir. Bu beklenti yükseltme döngüsü ayarlanan değerler daha önceden belirlenen farklı bir kriter üzerine yoğunlaşana kadar defalarca tekrarlanır. BM imputasyon yöntemi parametre değerleri için önyargısız standart hataları oluşturur (Cheema, 2012).

Genel olarak BM algoritması,

- Tahmin edilmiş değerleri kayıp verilerin yerine koyar,

- Bu kayıp verileri kullanarak parametre tahmin eder,
- Daha uygun parametre buluncaya dek algoritmayı yineler.

BM algoritması; B-adımı, beklenen adımı ve M-adımı (en büyükleme) adımı olmak üzere iki adımdan oluşmaktadır. B-adımı(Beklenen adımı): Gözlenemeyen verinin ya da kayıp verinin yerinin doldurulması problemidir. M-adımı (En büyükleme adımı): Tahmin edilen kayıp veri değerini kabul ederek oluşan tam-veri modeli üzerinden, bilinen en çok olabilirlik tahmini hesaplanmaktadır. M-adımı sonucunda ortaya çıkan tahminler, BM algoritmasının çıktısını oluşturmaktadır (Yazıcı, 2005).

Çoklu atama (multiple ımputation)

Bu yöntemde, kayıp veri, iki ya da daha fazla atama yönteminin birlikte kullanılması sonucu kestirilir. Dolayısıyla, çoklu atama yöntemi, karma bir kestirim değeri elde etmeyi amaçlar. Bu karma kestirim değeri, genellikle, iki ya da daha fazla yöntemle elde edilmiş kestirim değerinin ortalamasıdır. Bu son kestirimin tek bir yöntemle elde edilen kestirim değerine göre daha güvenilir olacağı düşünülmektedir (Alpar, 2003). Bu yöntem kayıp verilerdeki doğal değişimi tetikleyen ileri bir atama tekniğidir. Çeşitli veri setlerinden oluşan tahmin grupları daha sonra tekli sete dönüştürülür. Bu dönüştürme sırasında örnek olarak ortalama yöntemi kullanılır. Çoklu atama doğal değişimi tetiklediği için bu yöntemle ortaya çıkan parametre tahminlerinin standart hataları önyargısız kalır (Cheema, 2012).

Çoklu atıf tekniği, kayıp değerlerin yerine m tekrar sayısı ve m >1 olmak üzere simüle edilmiş yöntemlerin kullanıldığı bir Monte Carlo tekniğidir. Buradaki m sayısı 3-10 arasında oldukça küçük bir değerdir (Schafer, 1997).

Bu yöntem üç temel adımı gerektirir:

- Atama (imputation)
- Analiz etme (analysis)
- Bir araya getirme (pooling) (Schafer,1997; Oğuzlar, 2001).

Bu adımlar arasında başarması en zor olan atama adımıdır. Bu adımda karşılaşılabilecek tipik problemler şunlardır:

- Herhangi bir gözlemin kayıp olması, o gözlemin değerine bağlıdır. Örneğin yüksek veya düşük gelir düzeyine sahip kişiler gelir sorusunu atlama eğilimindedirler.
- Kayıp değerler veriler kümesinin herhangi bir yerinde görülebilir.
- Atama adımında kullanılan yöntem, daha sonra yapılması düşünülen tamamlanmış veriler ile analiz aşamasının öngörülmesini zorunlu kılar (Schafer,1997; Oğuzlar, 2001).

Oğuzlar (2001)'a göre atama yapılan veriler için tekrar uygulanan analiz adımı, atama yöntemi uygulanmadan önce yapılan aynı analizden daha basittir. Çünkü kayıp değerlerin sıkıntısı ortadan kalkmıştır. Bir araya getirme adımı ise, m defa tekrarlanmış analizlerden, p değerleri, güven aralıkları, varyanslar ve ortalamaların hesaplamasını içerir. Bu hesaplamalar da genel olarak basit hesaplamalardır. Ayrıca çoklu atama tekniği her şeyden önce oldukça iyi anlaşılabilecek bir teknik olması, analizde yer alan değişkenlerin normalliği ihlal ettiği durumlarda da güçlü ve üstün çözümler sunması, standart hataların daha iyi tahminler sunması gibi yönleriyle oldukça avantajlıdır. Bu yöntemin dezavantajı ise m sayısında 3'ten 10'a kadar veri kümesinde atama işlemi yaparken yoğun zaman gerektirmesi ve eğer önemli mekanizmalar analize dahil edilmediyse yanlı sonuçlar doğurmasıdır.

Tablo 2-3. Kayıp veri sürecinde kullanılan yöntemlerin karşılaştırılması

YÖNTEM	AVANTAJI	DEZAVANTAJI
Liste / durum bazında silme	• Pozitif tanımlı korelasyon matrisi üretir. • Tüm analiz tek örnekleme dayanır. • Hesaba gerek olmadığı için basit bir yöntemdir.	• Eğer veriler ROK koşulunu sağlamıyorsa, sonuçlar yanlı olabilir. • Eldeki verilere önem vermez. • Eksik veri kullanıldığından varyansı arttırır.
Çiftler bazında silme	• Eldeki verilerin tamamını kullanır. Böylece daha etkin sonuç verir. • Daha küçük varyans üretir. • Basit bir yöntemdir.	• Ürettiği korelasyon matrisi pozitif tanımlı değildir. • Her bir korelasyon farklı örnekleme dayanır. • Veriler TROK ise yanlı sonuçlar verir. • Eğer veriler ROK ise iyi tahminler ve standart hatalar üretir.
Yerine ortalamayı koyma	• Hesaplaması kolaydır.	• Varyansları, dolayısıyla da korelasyon ve kovaryansları düşürür. • Verilerin gerçek dağılımları bozulur.
Regresyon ataması	• Kayıp değerler için iyi bir tahmin sağlar. • BM' den farklı olarak atanacak her değer için tahmincilerin farklı bir kümesini kullanır.	• SPSS kullanılırsa; rassal hata eklenmezse sonuçlar tahminin üzerinde çıkar, çoklu bağlılık oluşur, varyans düşer. • Tek iterasyondan oluştuğu için BM kadar çok bilgi kullanmaz.
Beklenti maksimizasyonu	• Eldeki tüm verileri kullanır. • Atamaya rassal hata terimini ekler.	• Yanlı sonuçlar doğurabilir.
Hot/Cold Deck ataması	• Uygulaması kolaydır.	• Benzer gözlemler araştırılırken farklılıklar ortaya çıkabilir.
Çoklu atama	• Atamaya rassal hata terimini ekler. • Standart hataların daha iyi tahminini sağlar.	• Yanlı sonuçlar doğurabilir.

BÖLÜM III

3. Yöntem

Araştırmanın bu bölümünde araştırmanın modeli, örneklem, veri toplama aracı ve verilerin analizine yer verilmiştir.

3.3. Araştırmanın Modeli

Araştırma modelimiz iki ve daha çok sayıdaki değişken arasında birlikte değişim varlığını veya derecesini belirlemeyi amaçlayan ilişkisel tarama modelidir (Karasar, 2004).

3.4. Örneklem

Bu araştırmada n=50, 100, 200, 400 birim ve her biri üç değişkenden oluşan toplamda 2250 adet koşullu türetilmiş veri kullanılmıştır.

3.2.1. Veri türetimi

Bu çalışmada yapay verilerden yararlanılmıştır. Veri türetimi belirli koşullar göz önüne alınarak R-Studio paket programı yardımıyla elde edilmiştir.

Veri setleri; birim sayıları 50, 100, 200, 400 birim içerecek şekilde, çok değişkenli standart normal dağılıma uygun türetilmiştir. Her bir veri setinde oluşturulan üç değişken sırasıyla X_1, X_2, X_3 şeklinde isimlendirilmiştir. Bu değişkenlerin

aralarındaki korelasyonlar ise -0.30< r < 0.30 yani düşük korelasyon ve r<-0.70 veya r>0.70 yani yüksek korelasyon olacak şekilde iki alt gruba ayrılmıştır.

N=50, 100, 200, 400 birimlik rastgele türetilen tam veri setlerindeki değişkenler sırasıyla %5, %10, %20 oranlarında tesadüfi olarak eksiltilmiş ve kayıp veri analizleri için kullanılacak yeni veri setleri elde edilmiştir. Düşük ve yüksek korelasyona sahip 50, 100, 200, 400 birimlik toplam 8 veri seti içinden sırasıyla %5, %10, %20 gözlem eksiltilerek 24 yeni veri seti oluşturulmuştur. Yani analizlerimiz için toplamda 32 veri seti ile çalışılmıştır. Oluşturulmuş yapay veri periyotlarından yararlanılarak, uygulanılan yöntemlerin başarımlarını karşılaştırmak amacıyla kayıp veri tamamlama yöntemlerinden yerine ortalamayı koyma, beklenti maksimizasyonu (BM), regresyon yöntemi ile silme yöntemlerinden liste/durum bazında silme teknikleri seçilmiştir. Kayıp veri içermeyen verilerle çeşitli oranlardaki kayıp senaryolarına göre kayıp veri içeren veri setlerine t- testi ve ANOVA uygulanmıştır. Elde edilen gerçek parametreler ile kayıp veri yöntemleri sonucu elde edilen parametrelere karşılaştırma analizi uygulanmıştır.

3.5. Kayıp Veri Tamamlama Yöntemleri

3.5.2. Liste/durum bazında silme yöntemi

Kayıp değer bulunan veri setlerinde akla gelen ilk yöntem, kayıp değer içeren kayıtların yok sayılmasıdır. Bunun için kayıp değer içeren gözlemler veri setinden çıkarılarak analize dahil edilmez (Sezgin ve Çelik, 2013).

3.5.3. Yerine ortalamayı koyma yöntemi

Bu yöntem, veri setinde kayıp verinin olduğu alandaki diğer verilerin ortalamasını alarak kayıp olan verileri doldurmaya yarayan yöntemdir (Sezgin ve Çelik, 2013).

3.3.3. Beklenti maksimizasyonu (EM/ BM)

BM Algoritması bir objenin hangi kümeye ait olduğunu belirlemede kesin mesafe ölçütlerini kullanmak yerine tahminsel ölçütleri kullanmayı tercih eder (Sezgin ve Çelik, 2013).

BM algoritmasının genel basamakları aşağıda verilmiştir:

1. Parametrelerin başlangıç değerleri verilir;
2. B-adımında kayıp veri üzerinden log-olabilirliğin beklenen fonksiyonu alınarak tam-veri örneklemi hesaplanır;
3. M-adımında tam-veri üzerinden örneklemin en çok olabilirlik tahmini hesaplanır;
4. Bulunan parametrenin beklenen fonksiyonu ile bir önceki parametrenin beklenen fonksiyonu karşılaştırılır, aralarındaki fark azalırsa, algoritma durdurulur;
5. Eğer fark mevcutsa, hesaplanan parametre, bir önceki parametre ile yer değiştirip yeniden hesaplanır (Yazıcı, 2005).

3.3.4. Regresyon yöntemi

Regresyon yönteminde bir ya da daha fazla bağımsız değişken yardımıyla bağımlı değişken değerinin test edilmektedir. Regresyon atama yönteminde, bağımlı değişken kayıp gözlemli değişken; bağımsız değişkenler ise diğer değişkenlerdir (Alpar, 2003).

Tablo 3-1.Yüzdeliklere göre kayıp veri miktarları

Yüzdeliklere Göre Kayıp Veri Miktarları			
0%	5%	10%	20%
50	7	15	30
100	15	30	60
200	30	60	120
400	60	120	240

Örneğin, Tablo 3-1.'de 100 birimlik veri seti için X1-X2-X3 değişkenleri için toplamda 300 adet verimiz bulunmaktadır. %10 kayıp durumu için 300 veriden 30 adet verinin silinmesi gerekmektedir.

Tablo 3-2. 50 birimlik veri setinde değişkenlerin eksik olma sayı ve yüzdeleri

Değişken	Tam Veri	%5 Eksik	%10 Eksik	%20 Eksik
X1	50	2	5	10
X2	50	2	5	10
X3	50	3	5	10

Tablo 3-3. 100 birimlik veri setinde değişkenlerin eksik olma sayı ve yüzdeleri

Değişken	Tam Veri	%5 Eksik	%10 Eksik	%20 Eksik
X1	100	5	10	20
X2	100	5	10	20
X3	100	5	10	20

Tablo 3-4. 200 birimlik veri setinde değişkenlerin eksik olma sayı ve yüzdeleri

Değişken	Tam Veri	%5 Eksik	%10 Eksik	%20 Eksik
X1	200	10	20	40
X2	200	10	20	40
X3	200	10	20	40

Tablo 3-5. 400 birimlik veri setinde değişkenlerin eksik olma sayı ve yüzdeleri

Değişken	Tam Veri	%5 Eksik	%10 Eksik	%20 Eksik
X1	400	20	40	60
X2	400	20	40	60
X3	400	20	40	60

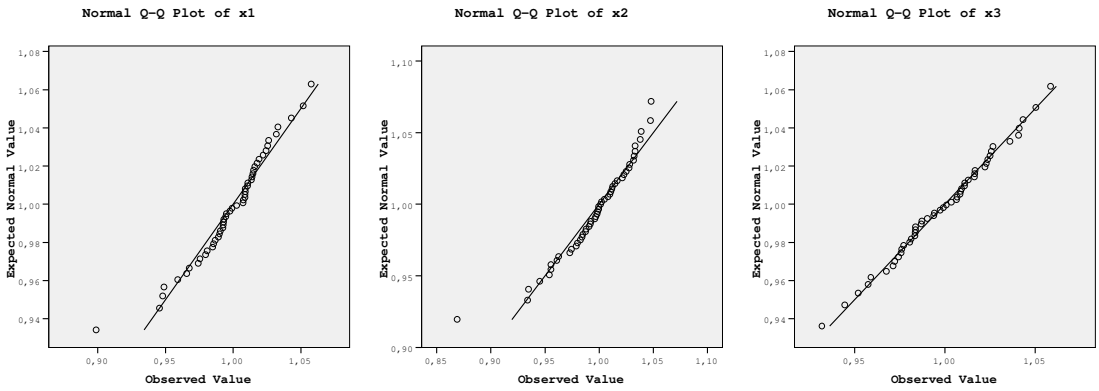
Yukarıdaki tablolar incelendiğinde eksiltmelerde verilen yüzde değerlerine uygun azaltmalar gerçekleştirilerek hedef sayılara ulaşıldığı kabul edilmiştir. Verilerin el ile silinmesinde rastgele olması dikkate alındığı görülür. Ayrıca rastgelelik şartının sağlandığı da Little'ın TROK testi sonuçlarıyla da görülür.

Tablo 3-6. Türetilen veri setlerine ait Little'ın TROK (MCAR) test sonuçları

Veri Setleri	Korelasyon Düzeyleri	Eksik Olma Durumu %	χ^2	sd	P
50 Birimlik Veri Seti	Düşük Korelasyon	5	3.029	6	0.805
		10	1.343	6	0.969
		20	3.465	6	0.749
	Yüksek Korelasyon	5	6.259	6	0.395
		10	6.561	6	0.363
		20	5.314	6	0.504
100 Birimlik Veri Seti	Düşük Korelasyon	5	9.490	6	0.148
		10	5.194	6	0.519
		20	2.361	6	0.884
	Yüksek Korelasyon	5	1.164	6	0.979
		10	1.446	6	0.963
		20	1.586	6	0.954
200 Birimlik Veri Seti	Düşük Korelasyon	5	7.193	6	0.303
		10	4.141	6	0.658
		20	3.527	6	0.740
	Yüksek Korelasyon	5	5.556	6	0.475
		10	5.002	6	0.544
		20	6.454	6	0.374
400 Birimlik Veri Seti	Düşük Korelasyon	5	3.481	6	0.747
		10	2.984	6	0.811
		20	3.600	6	0.731
	Yüksek Korelasyon	5	3.645	6	0.725
		10	1.964	6	0.923
		20	1.201	6	0.977

Tablo 3-6. incelendiğinde tüm veri setleri için p anlamlılık değerinin $p>0.05$ olduğu için veri setlerinin TROK yapısına uygun olduğu görülür.

Veri setlerine t-testi ve ANOVA uygulamalarının yapılabilmesi için ön şart normalliğin sağlanmasıdır. Veri setlerini oluşturan X1, X2, X3 değişkenlerin normalliği Q-Q plot grafiğinden yararlanarak, değişkenlerimiz arasındaki ilişkinin varlığı ise her bir grup için ayrı ayrı tabloda gösterilmiştir.

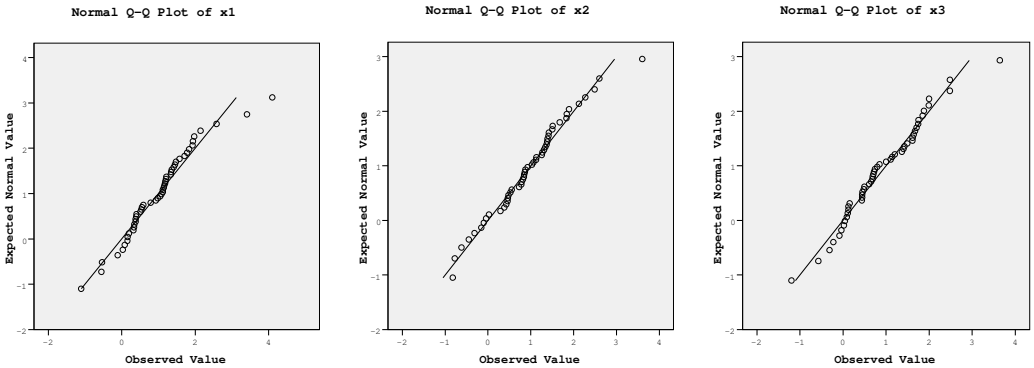


Şekil 3-1. 50 birimlik düşük korelasyonlu veri setindeki değişkenler için Q-Q plot grafikleri

Tablo 3-7. Değişkenler arasındaki korelasyon katsayıları

	N	r
x1 & x2	50	,148
x1 & x3	50	,058
x2 & x3	50	,006

Şekil 3-1. ve Tablo 3-7. incelendiğinde veri setinin düşük korelasyonlu (-0.30 < r < 0.30) ve normal dağılıma sahip olduğu görülür.

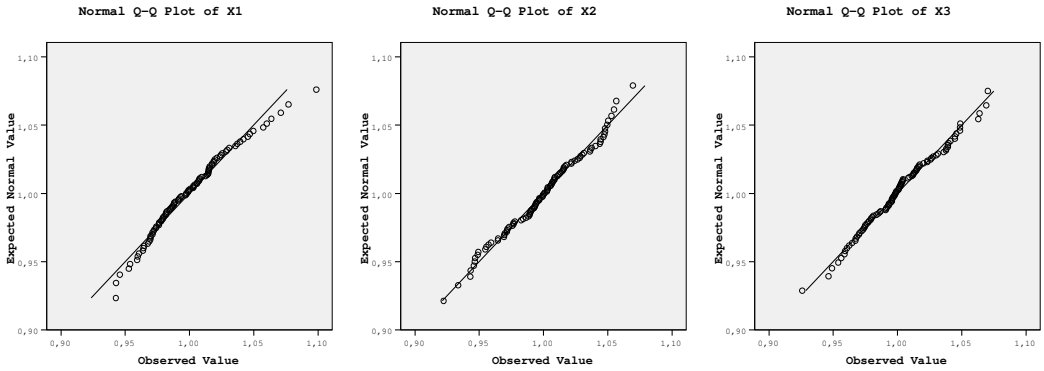


Şekil 3-2. 50 birimlik yüksek korelasyonlu veri setindeki değişkenler için Q-Q plot grafikleri

Tablo 3-8. Değişkenler arasındaki korelasyon katsayıları

	N	r
x1 & x2	50	,858
x1 & x3	50	,855
x2 & x3	50	,838

Şekil 3-2. ve Tablo 3-8. incelendiğinde veri setinin yüksek korelasyonlu ($r < -0.70$ veya $r > 0.70$) ve normal dağılıma sahip olduğu görülür.

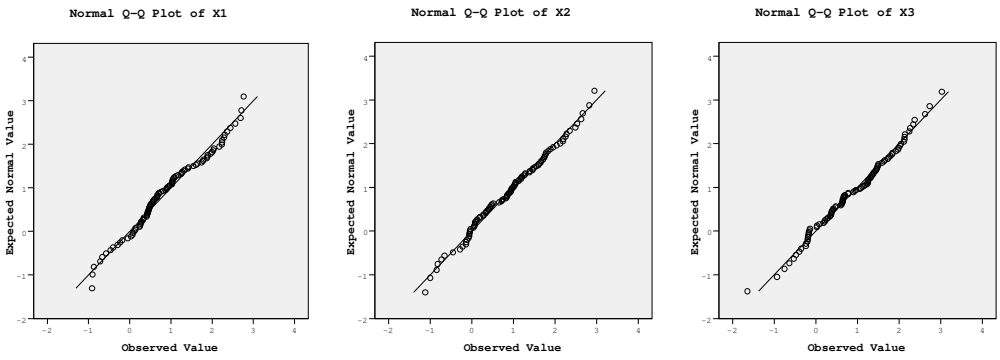


Şekil 3-3. 100 birimlik düşük korelasyonlu veri setindeki değişkenler için Q-Q plot grafikleri

Tablo 3-9. Değişkenler arasındaki korelasyon katsayıları

	N	r
x1 & x2	100	-,105
x1 & x3	100	,029
x2 & x3	100	-,039

Şekil 3-3. ve Tablo 3-9. incelendiğinde veri setinin düşük korelasyonlu ($-0.30 < r < 0.30$) ve normal dağılıma sahip olduğu görülür.

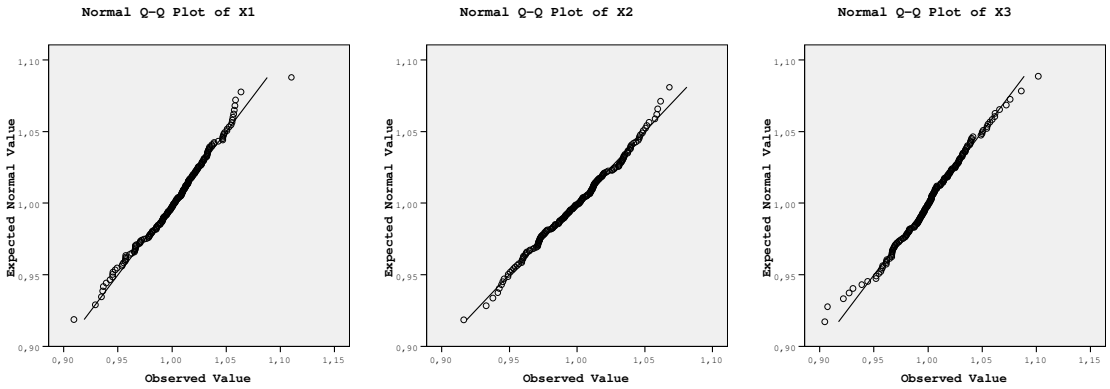


Şekil 3-4. 100 birimlik yüksek korelasyonlu veri setindeki değişkenler için Q-Q plot grafikleri

Tablo 3-10. Değişkenler arasındaki korelasyon katsayıları

	N	r
x1 & x2	100	100
x1 & x3	100	100
x2 & x3	100	100

Şekil 3-4. ve Tablo 3-10. incelendiğinde veri setinin yüksek korelasyonlu ($r < -0.70$ veya $r > 0.70$) ve normal dağılıma sahip olduğu görülür.

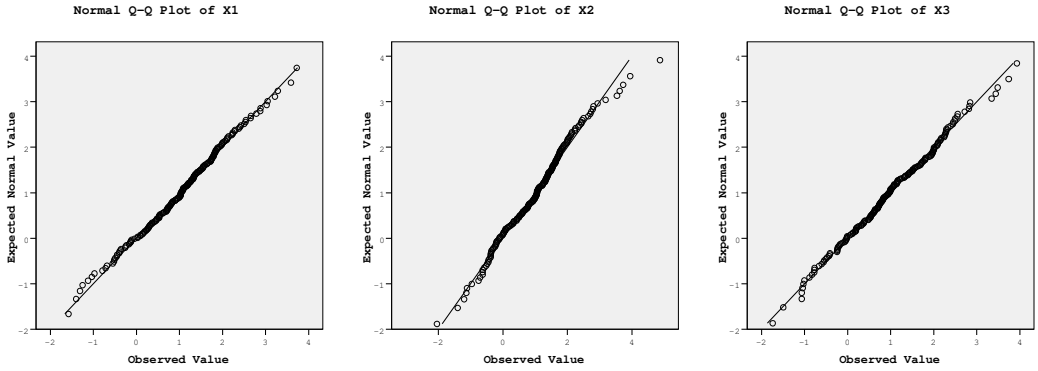


Şekil 3-5. 200 birimlik düşük korelasyonlu veri setindeki değişkenler için Q-Q plot grafikleri

Tablo 3-11. Değişkenler arasındaki korelasyon katsayıları

	N	r
x1 & x2	200	-,052
x1 & x3	200	-,132
x2 & x3	200	-,033

Şekil 3-5. ve Tablo 3-11. incelendiğinde veri setinin düşük korelasyonlu ($-0.30 < r < 0.30$) ve normal dağılıma sahip olduğu görülür.

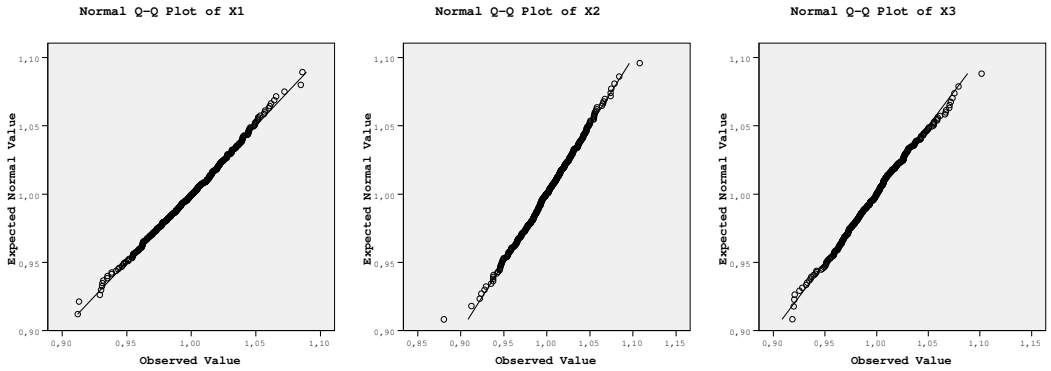


Şekil 3-6. 200 birimlik yüksek korelasyonlu veri setindeki değişkenler için Q-Q plot grafikleri

Tablo 3-12. Değişkenler arasındaki korelasyon katsayıları

	N	r
x1 & x2	200	,911
x1 & x3	200	,924
x2 & x3	200	,922

Şekil 3-6. ve Tablo 3-12. incelendiğinde veri setinin yüksek korelasyonlu ($r<0.70$ veya $r > 0.70$) ve normal dağılıma sahip olduğu görülür.

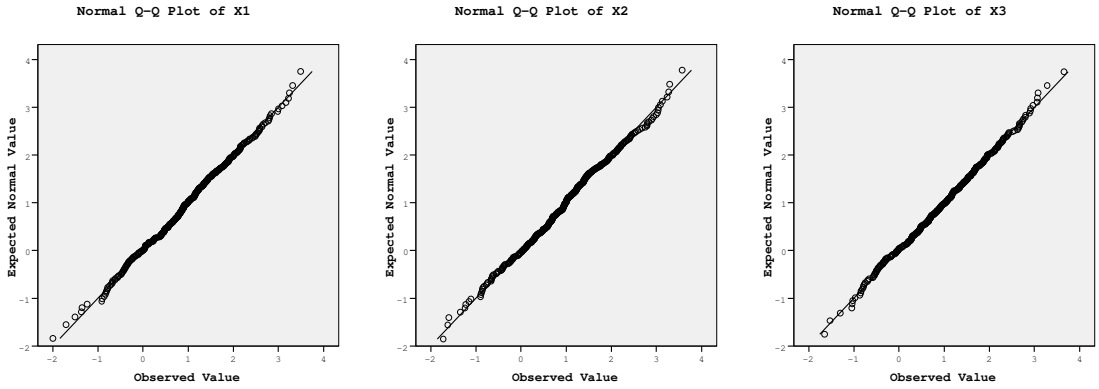


Şekil 3-7. 400 birimlik düşük korelasyonlu veri setindeki değişkenler için Q-Q plot grafikleri

Tablo 3-13. Değişkenler arasındaki korelasyon katsayıları

	N	r
x1 & x2	400	,005
x1 & x3	400	-,043
x2 & x3	400	-,023

Şekil 3-7. ve Tablo 3-13. incelendiğinde veri setinin düşük korelasyonlu ($-0.30 < r < 0.30$) ve normal dağılıma sahip olduğu görülür.



Şekil 3-8. 400 birimlik yüksek korelasyonlu veri setindeki değişkenler için Q-Q plot grafikleri

Tablo 3-14. Değişkenler arasındaki korelasyon katsayıları

	N	r
x1 & x2	400	,891
x1 & x3	400	,906
x2 & x3	400	,885

Şekil 3-8. ve Tablo 3-14. incelendiğinde veri setinin yüksek korelasyonlu ($r<-0.70$ veya $r > 0.70$) ve normal dağılıma sahip olduğu görülür.

ANOVA testinin uygulanması için sağlanması gereken ön şartlardan biri de grupların varyanslarının homojen olmasıdır. Homojenliği Levene istatistiği sonuçlarına göre p değerlerine göre değerlendiririz. Bu değerler Tablo 3-15.'de gösterilmiştir. Tabloya göre tüm p değerleri >0.05 olduğu için veri setlerinin her birine ANOVA uygulanabilir.

Tablo 3-15. Levene istatistiğine göre p değerleri

			TAM	SİLME	ORTALAMA	REGRESYON	BM
	N	%	p	P	p	p	p
	50	0	0,558				
		5		0,558	0,607	0,604	0,601
		10		0,609	0,473	0,525	0,520
		20		0,925	0,411	0,391	0,437
	100	0	0,034				
		5		0,012	0,017	0,010	0,018
		10		0,078	0,029	0,005	0,030
		20		0,371	0,280	0,220	0,180
Düşük	200	0	0,7				
		5		0,830	0,765	0,912	0,760
		10		0,808	0,684	0,836	0,651
		20		0,981	0,737	0,643	0,705
	400	0	0,128				
		5		0,206	0,263	0,141	0,272
		10		0,096	0,329	0,407	0,342
		20		0,331	0,477	0,231	0,464
	50	0	0,902				
		5		0,948	0,974	0,958	0,948
		10		0,967	0,976	0,946	0,985
		20		0,583	0,639	0,871	0,852
	100	0	0,159				
		5		0,192	0,172	0,261	0,176
		10		0,082	0,115	0,083	0,135
		20		0,242	0,191	0,084	0,142
Yüksek	200	0	0,154				
		5		0,211	0,238	0,133	0,187
		10		0,305	0,225	0,274	0,218
		20		0,608	0,423	0,149	0,198
	400	0	0,8				
		5		0,761	0,782	0,770	0,748
		10		0,922	0,790	0,492	0,712
		20		0,768	0,743	0,544	0,676

BÖLÜM IV

4. Bulgular ve Yorumlar

Bu bölümde 1.2.1'deki alt problemlere ait bulgular ve yorumlar yer almaktadır.

1. Normal dağılıma sahip düşük ve yüksek korelasyonlu tam veri setleri (50-100-200-400 birim) ve belirli oranlardaki (%5, %10, %20) veri setlerine silme ve atama yöntemi kullanılarak oluşan yeni veri setlerindeki t-testi sonuçları arasında ki benzerlik durumu nedir?
2. Normal dağılıma sahip düşük ve yüksek korelasyonlu tam veri setleri (50-100-200-400 birim) ve belirli oranlardaki (%5, %10, %20) veri setlerine silme ve atama yöntemi kullanılarak oluşan yeni veri setlerindeki ANOVA sonuçları arasında ki benzerlik durumu nedir?
3. Normal dağılıma sahip düşük ve yüksek korelasyonlu tam veri setleri ile belirli oranlardaki kayıp veri setlerine çeşitli atama yöntemleri kullanılarak oluşan yeni veri setlerinde yapılan analizler sonucu 1. tip ve 2. tip hataya rastlanır mı?

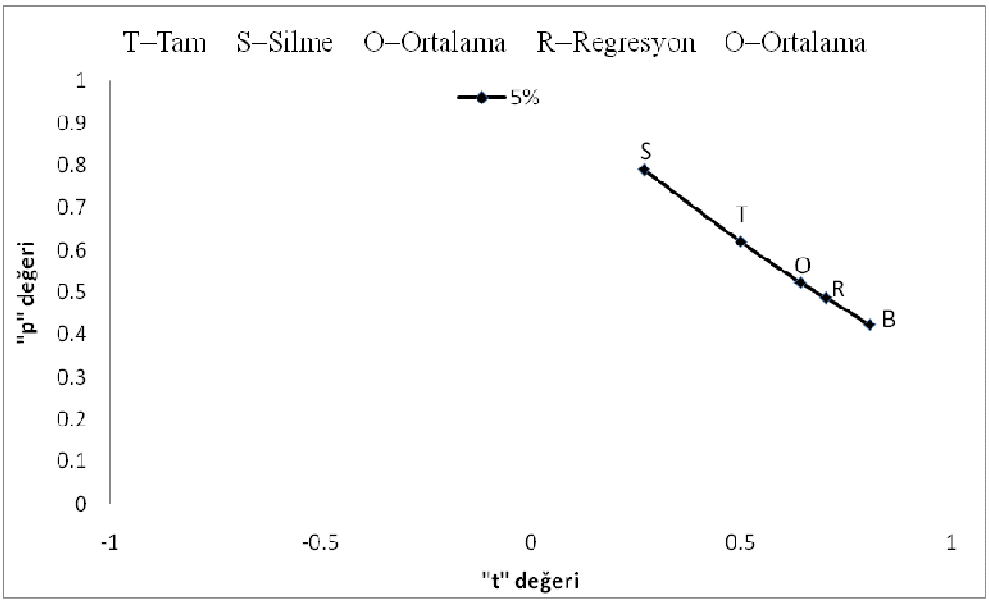
4.1. t- Testine Ait Bulgular ve Yorumlar

Tablo 4-1.'de tam (kayıp olmayan) verilerin ve belirli oranlarda kayıp söz konusu iken silme ve atama yolları ile tamamlanmış yeni veri setlerine uygulanmış t-testine ait t ve p değerleri bulunmaktadır. Veri setlerindeki farklı miktardaki kayıpların giderilmesi için uygulanacak (bilimsel-istatistiksel) uygun yöntemin belirlenmesinde, tam verilere ait sonuçlara yakınlığı/benzerliği gözlenmeye çalışılmıştır. Grafiklerde t-testine ait t değerler yatay eksende, p değerleri düşey eksenlerin karşılıklı çakıştırılması ile oluşan noktalarla belirtilmiştir. Böylece tam (Tam veri = T) verilere ait nokta

çiftlerine ne kadar benzer/yakın değerler oluşmuş ise (Silme yöntemi = S, Yerine ortalama koyma yöntemi = O, Regresyon yöntemi = R, Beklenti maksimizasyonu Yöntemi = B) seçilen yöntemin gerçeğe yakın değerler tahmin edilebileceği sonucuna varılmıştır.

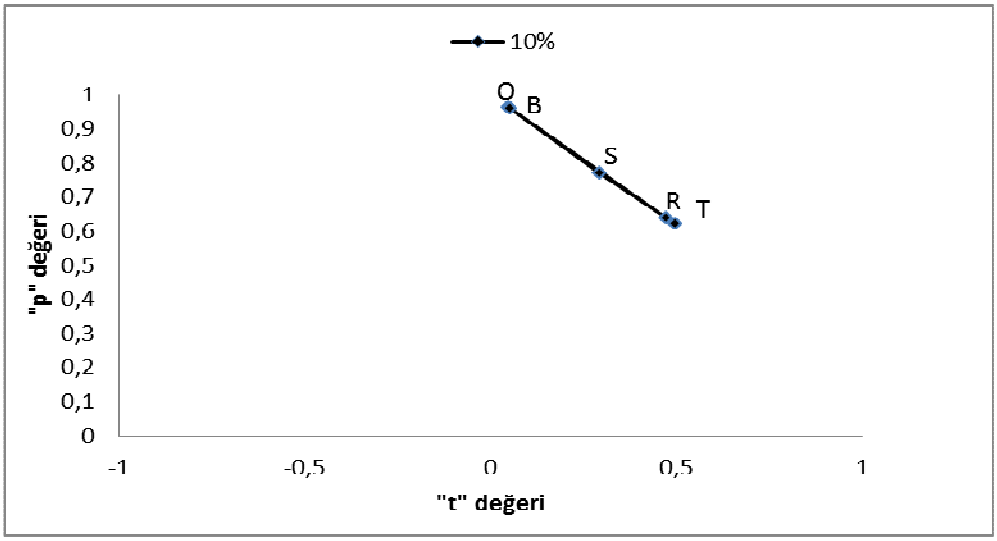
Tablo 4-1. Tüm veri setlerine ait t testinin t ve p değerleri

			TAM		SİLME		ORTALAMA		REGRESYON		BM	
	N	%	t	p	t	P	T	p	t	p	t	p
		0	0.498	0.620								
		5			0.2688	0.7893	0.702	0.486	0.806	0.424	0.642	0.524
	50	10			0.2927	0.7714	0.049	0.961	0.472	0.639	0.052	0.959
		20			-0.1823	0.8572	-0.085	0.933	0.024	0.981	-0.055	0.957
		0	-0.089	0.929								
		5			-0.6263	0.5328	-0.260	0.795	-0.528	0.598	-0.229	0.820
	100	10			0.3887	0.6987	0.192	0.848	-0.153	0.879	0.213	0.831
		20			0.2715	0.7874	0.030	0.976	-0.009	0.993	-0.033	0.974
Düşük		0	1.158	0.248								
		5			0.8082	0.4201	1.318	0.189	1.505	0.134	1.340	0.182
	200	10			1.299	0.1962	1.388	0.167	1.687	0.093	1.429	0.155
		20			0.9199	0.3604	1.276	0.203	0.941	0.348	1.376	0.170
		0	-0.652	0.515								
		5			-1.057	0.2911	-0.747	0.456	-0.927	0.354	-0.758	0.449
	400	10			-13.107	0.191	-0.901	0.368	-0.606	0.545	-0.895	0.372
		20			-13.637	0.1746	-1.160	0.247	-0.617	0.538	-1.182	0.238
		0	0.851	0.399								
		5			0.5166	0.6081	-0.167	0.868	0.410	0.684	0.720	0.475
	50	10			0.755	0.4553	0.791	0.433	0.991	0.327	1.601	0.116
		20			0.0516	0.9593	-0.210	0.835	0.541	0.591	1.391	0.170
		0	-0.238	0.813								
		5			-0.0711	0.9434	-0.120	0.905	-0.059	0.954	-0.272	0.786
	100	10			-0.3423	0.7331	-0.303	0.763	-0.270	0.788	-0.463	0.644
		20			-14.779	0.1473	-1.670	0.098	-1.780	0.078	-1.883	0.063
Yüksek		0	0.780	0.436								
		5			0.3721	0.7103	0.768	0.443	1.454	0.148	0.985	0.326
	200	10			0.6683	0.505	-0.121	0.904	0.171	0.864	0.843	0.400
		20			0.7347	0.4646	-1.433	0.154	-0.289	0.773	0.118	0.906
		0	-0.359	0.720								
		5			-0.1428	0.8865	-0.320	0.749	-0.697	0.486	-0.261	0.794
	400	10			0.0666	0.9469	1.122	0.263	0.439	0.661	0.336	0.737
		20			0.1427	0.8867	-0.275	0.783	-0.506	0.613	-0.397	0.692



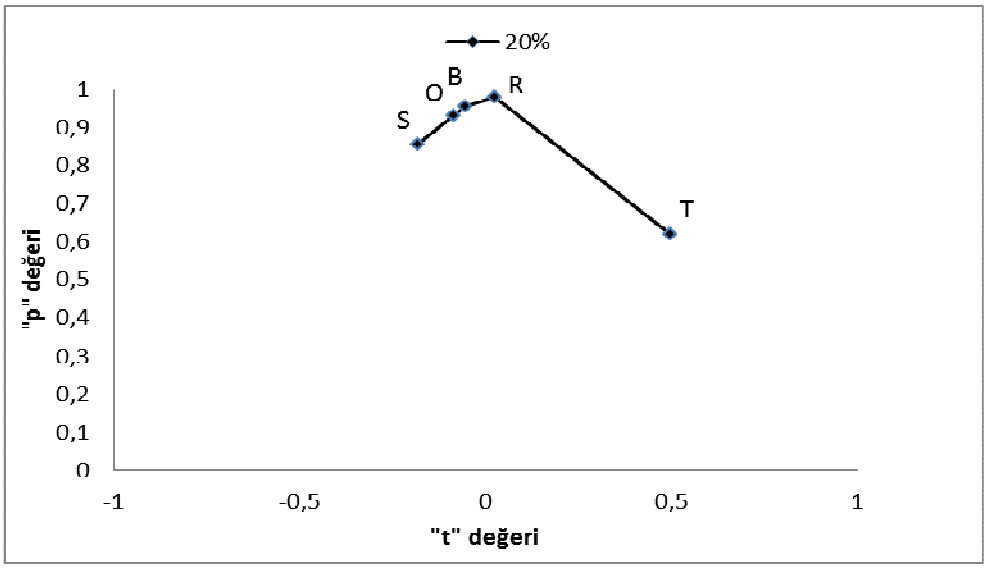
Şekil 4-1. 50 birimlik düşük korelasyonlu veri setinde %5 kayıp için t testi değerleri

Şekil 4-1. 'e göre 50 birimlik düşük korelasyona sahip %5 kayıp içeren veri setinde tam verilere en yakın değer BM yöntemi ile elde edilmiştir. Ayrıca tüm p değerleri >0.05 olduğu için analiz sonuçlarında 1. tip ve 2. tip hata gözlenmemiştir.



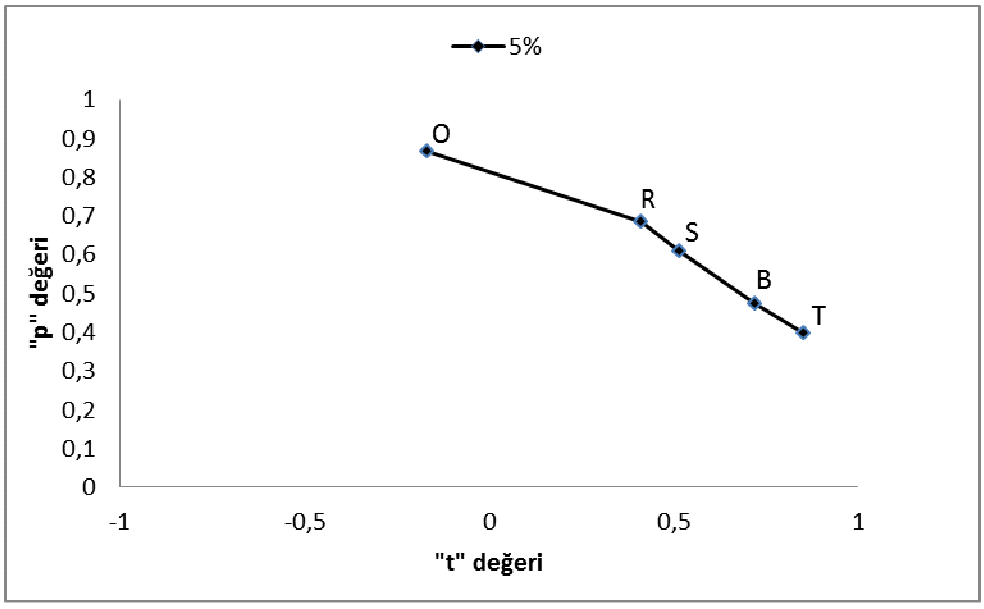
Şekil 4-2. 50 birimlik düşük korelasyonlu veri setinde %10 kayıp için t testi değerleri

Şekil 4-2.'ye göre 50 birimlik düşük korelasyona sahip, %10 kayıp içeren veri setinde tam verilere en yakın değer regresyon yöntemi ile elde edilmiştir. Ayrıca tüm p değerleri >0.05 olduğu için analiz sonuçlarında 1. tip ve 2. tip hata gözlenmemiştir.



Şekil 4-3. 50 birimlik düşük korelasyonlu veri setinde %20 kayıp için t testi değerleri

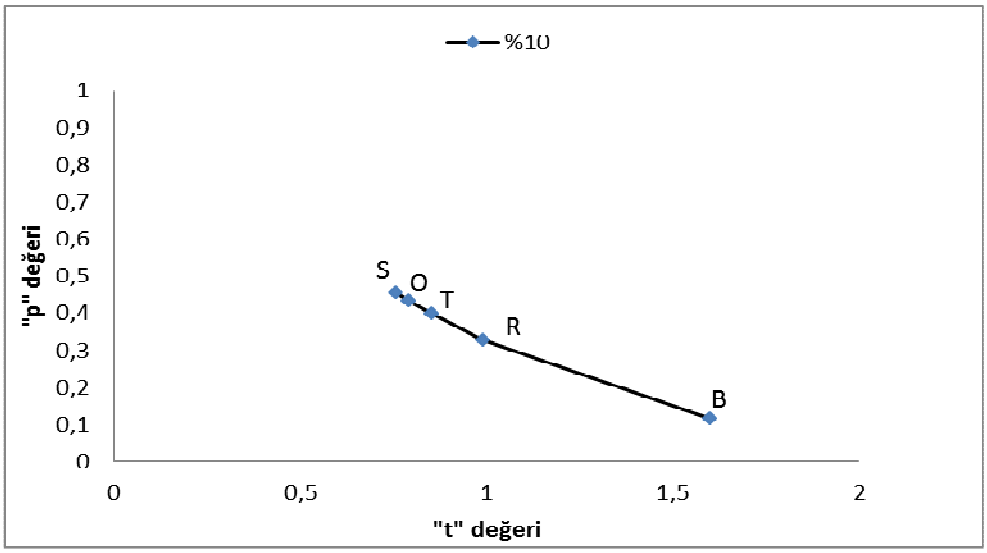
Şekil 4-3.'e göre 50 birimlik düşük korelasyona sahip, %20 kayıp içeren veri setinde kullanılan yöntemlerin tam veri setine ait t ve p değerlerine yakın sonuçlar vermediği görülmüştür. Buna rağmen tüm p değerleri >0.05 olduğu için analiz sonuçlarında 1. tip ve 2. tip hata gözlenmemiştir.



Şekil 4-4. 50 birimlik yüksek korelasyonlu veri setinde %5 kayıp için t testi değerleri

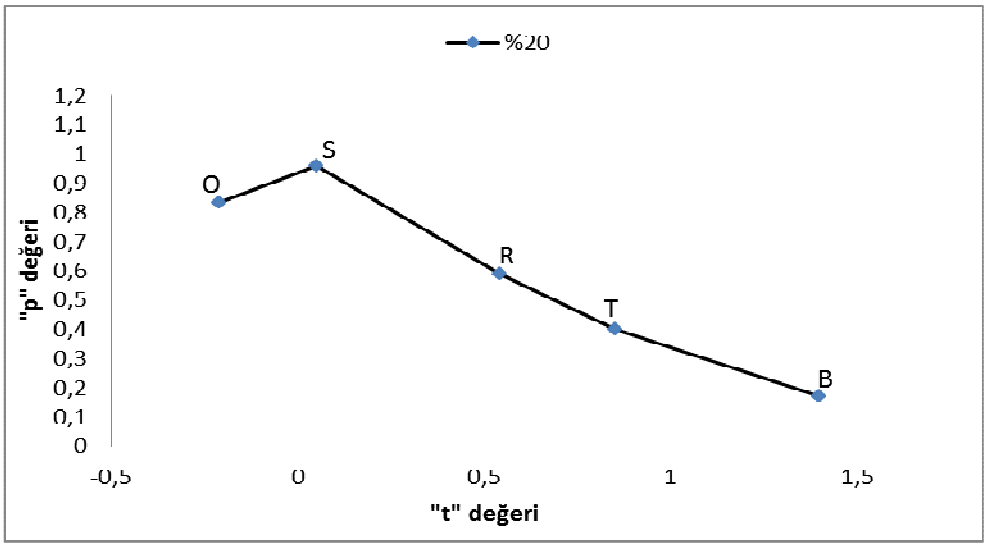
Şekil 4-4.'e göre 50 birimlik yüksek korelasyona sahip, %5 kayıp içeren veri setinde tam verilere en yakın değer BM yöntemi ile elde edilmiştir. Ayrıca tüm p

değerleri >0.05 olduğu için analiz sonuçlarında 1. tip ve 2. tip hata gözlenmemiştir.



Şekil 4-5. 50 birimlik yüksek korelasyonlu veri setinde %10 kayıp için t testi değerleri

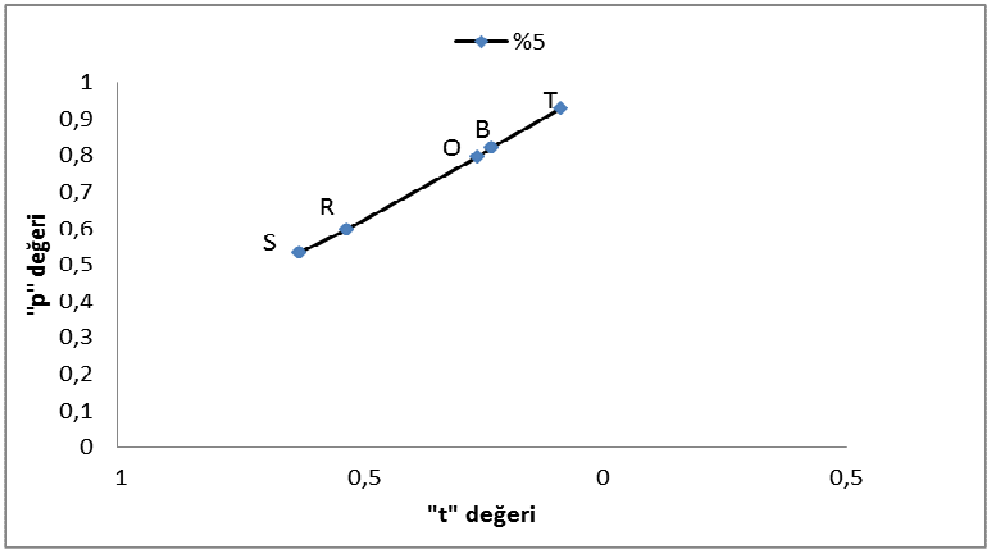
Şekil 4-5.'e göre 50 birimlik yüksek korelasyona sahip, %10 kayıp içeren veri setinde tam verilere en yakın değer silme yöntemi, yerine ortalama koyma yöntemi ile elde edilmiştir. Ayrıca tüm p değerleri >0.05 olduğu için analiz sonuçlarında 1. tip ve 2. tip hata gözlenmemiştir.



Şekil 4-6. 50 birimlik yüksek korelasyonlu veri setinde %10 kayıp için t testi değerleri

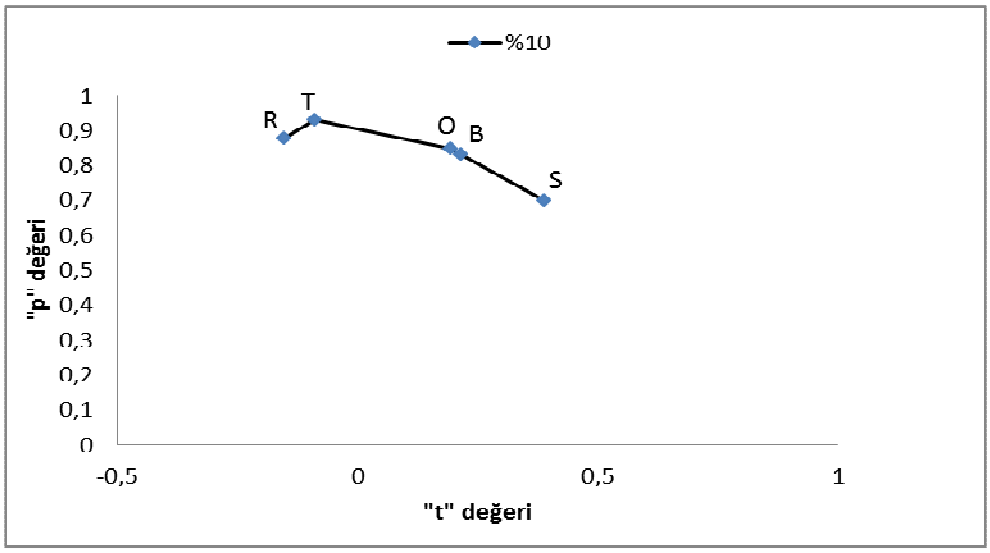
Şekil 4-6.'ya göre 50 birimlik yüksek korelasyona sahip, %20 kayıp içeren veri

setinde çok da etkin olmamaklar birlikte, tam verilere en yakın değer regresyon yöntemi ile elde edilmiştir. Ayrıca tüm p değerleri >0.05 olduğu için analiz sonuçlarında 1. tip ve 2. tip hata gözlenmemiştir.



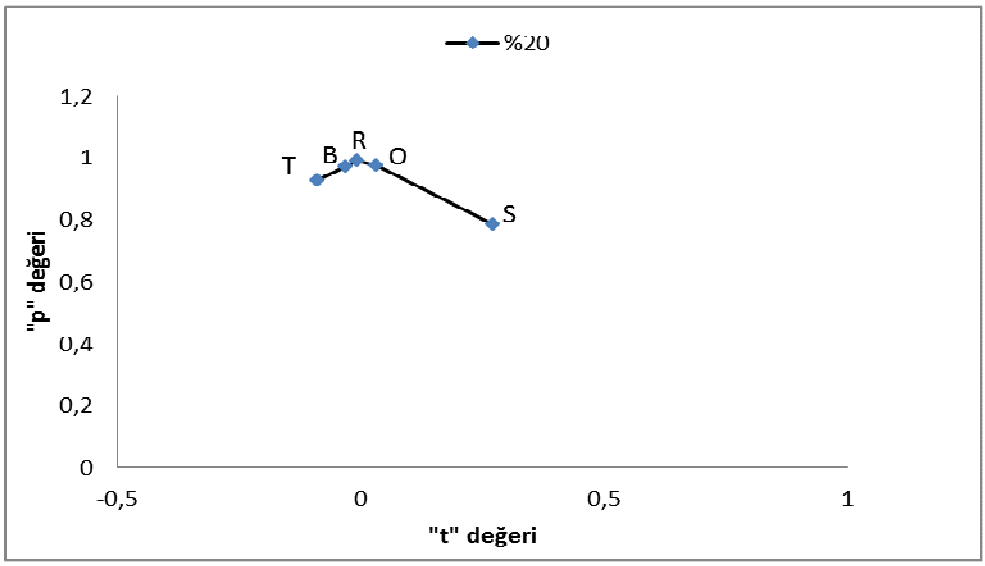
Şekil 4-7. 100 birimlik düşük korelasyonlu veri setinde %5 kayıp için t testi değerleri

Şekil 4-7.'ye göre 100 birimlik düşük korelasyona sahip, %5 kayıp içeren veri setinde uygulanan yöntemlerle tam verilere yakın değer elde edilememiştir. Buna rağmen tüm p değerleri >0.05 olduğu için analiz sonuçlarında 1. tip ve 2. tip hata gözlenmemiştir.



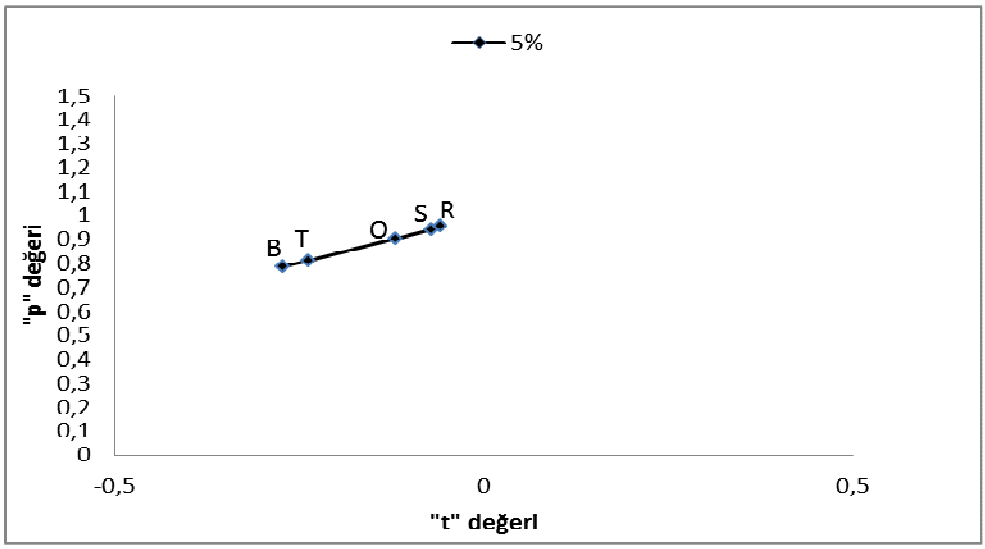
Şekil 4-8. 100 birimlik düşük korelasyonlu veri setinde %10 kayıp için t testi değerleri

Şekil 4-8.'e göre 100 birimlik düşük korelasyona sahip, %10 kayıp içeren veri setinde tam verilere en yakın değer regresyon yöntemi ile elde edilmiştir. Ayrıca tüm p değerleri >0.05 olduğu için analiz sonuçlarında 1. tip ve 2. tip hata gözlenmemiştir.



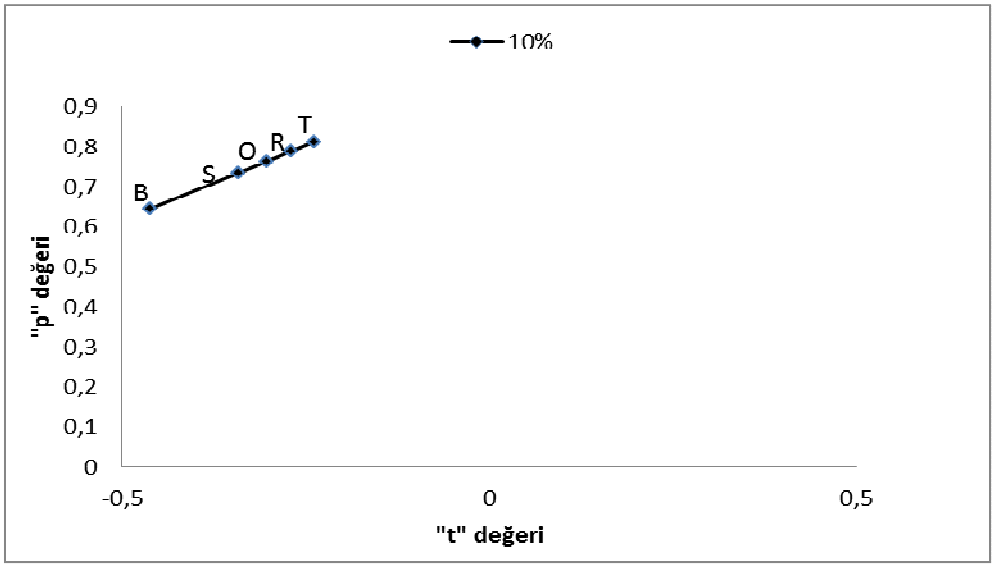
Şekil 4-9. 100 birimlik düşük korelasyonlu veri setinde %20 kayıp için t testi değerleri

Şekil 4-9.'a göre 100 birimlik düşük korelasyona sahip, %20 kayıp içeren veri setinde tam verilere en yakın değer BM ve regresyon yöntemleri ile elde edilmiştir. Ayrıca tüm p değerleri >0.05 olduğu için analiz sonuçlarında 1. tip ve 2. tip hata gözlenmemiştir.



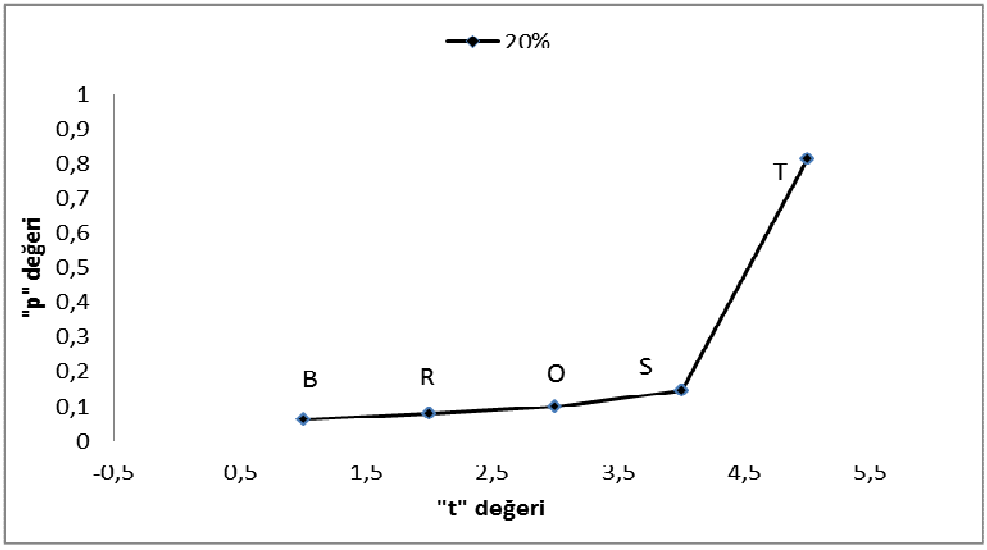
Şekil 4-10. 100 birimlik yüksek korelasyonlu veri setinde %5 kayıp için t testi değerleri

Şekil 4-10.'a göre 100 birimlik yüksek korelasyona sahip, %5 kayıp içeren veri setinde tam verilere en yakın değer BM yöntemi ile elde edilmiştir. Ayrıca tüm p değerleri >0.05 olduğu için analiz sonuçlarında 1. tip ve 2. tip hata gözlenmemiştir.



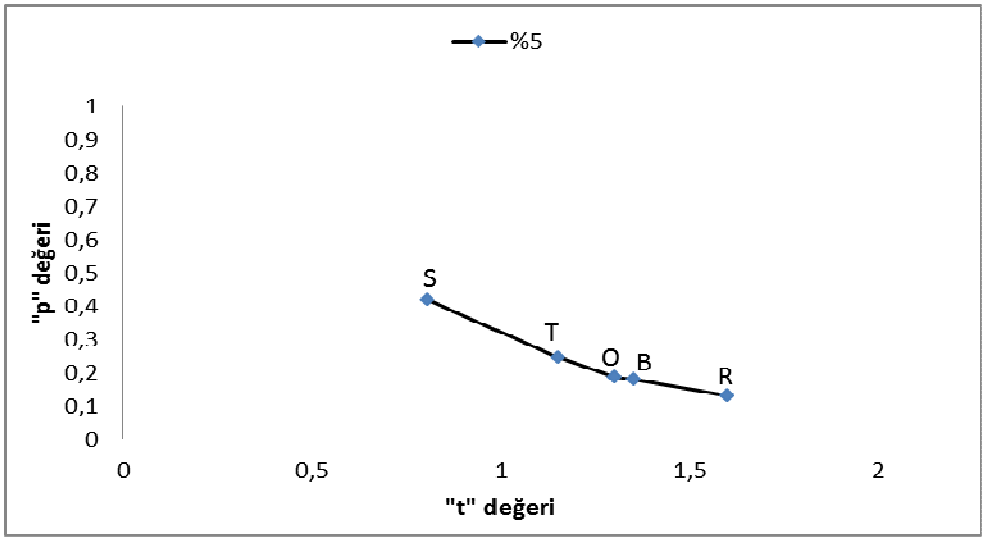
Şekil 4-11. 100 birimlik yüksek korelasyonlu veri setinde %10 kayıp için t testi değerleri

Şekil 4-11.'e göre 100 birimlik yüksek korelasyona sahip, %10 kayıp içeren veri setinde tam verilere en yakın değer regresyon ve yerine ortalama koyma yöntemi ile elde edilmiştir. Ayrıca tüm p değerleri >0.05 olduğu için analiz sonuçlarında 1. tip ve 2. tip hata gözlenmemiştir.



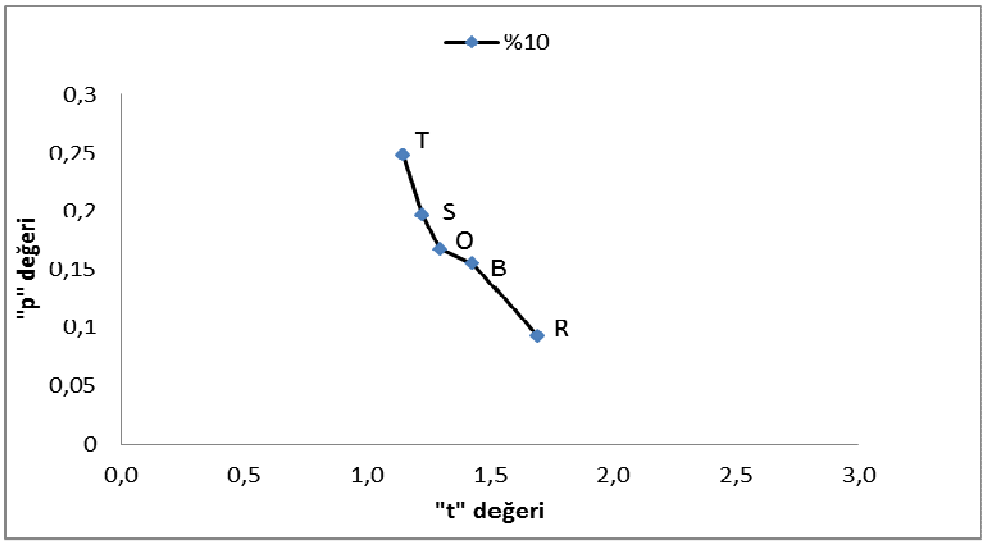
Şekil 4-12. 100 birimlik yüksek korelasyonlu veri setinde %20 kayıp için t testi değerleri

4-12.'ye göre 100 birimlik yüksek korelasyona sahip, %20 kayıp içeren veri setinde tam verilere en yakın değerler uyguladığımız yöntemlerle elde edilememiştir. Ayrıca yerine ortalama koyma ve regresyon yöntemlerinde p değerleri incelenirse bu yöntemlerin kullanılmasında 1. tip hata yapma olasılıklarının da arttığı görülmektedir.



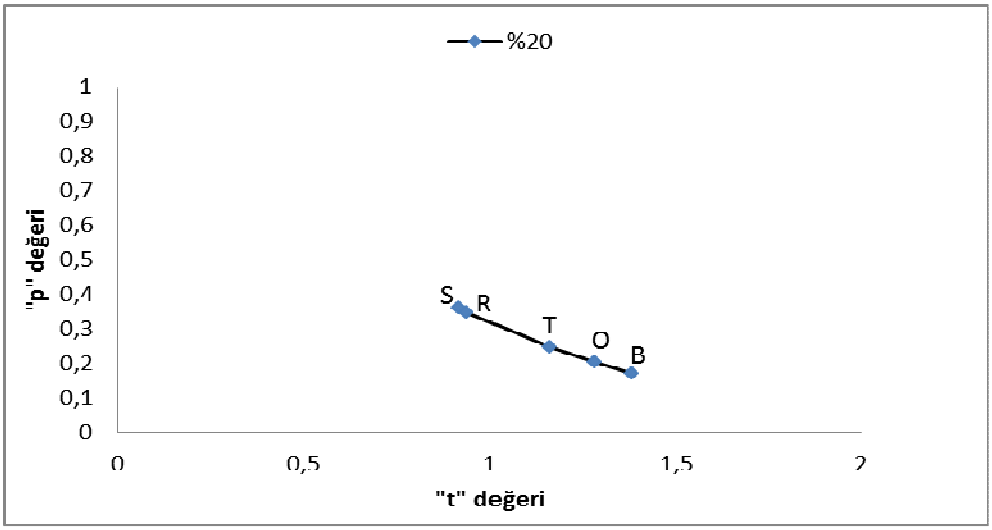
Şekil 4.13. 200 birimlik düşük korelasyonlu veri setinde %5 kayıp için t testi değerleri

Şekil 4-13.'e göre 200 birimlik düşük korelasyona sahip, %5 kayıp içeren veri setinde tam verilere en yakın değer BM yöntemi ve yerine ortalama koyma yöntemi ile elde edilmiştir. Ayrıca tüm p değerleri >0.05 olduğu için analiz sonuçlarında 1. tip ve 2. tip hata gözlenmemiştir.



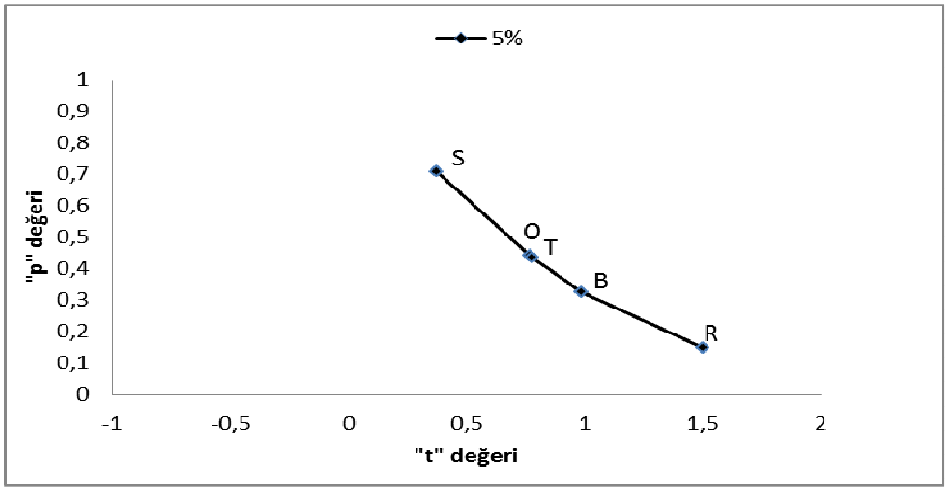
Şekil 4-14. 200 birimlik düşük korelasyonlu veri setinde %10 kayıp için t testi değerleri

Şekil 4-14.'e göre 200 birimlik düşük korelasyona sahip, %10 kayıp içeren veri setinde tam verilere en yakın değer silme yöntemi ve yerine ortalama koyma yöntemi ile elde edilmiştir. Ayrıca tüm p değerleri >0.05 olduğu için analiz sonuçlarında 1. tip hata gözlenmemiştir ancak regresyon yöntemi için bulunan p değerinin anlamlılık değeri olan 0.05 e yakınlığı, bu yöntemin kullanılmasında 1. tip hata yapılma olasılığını da göstermektedir.



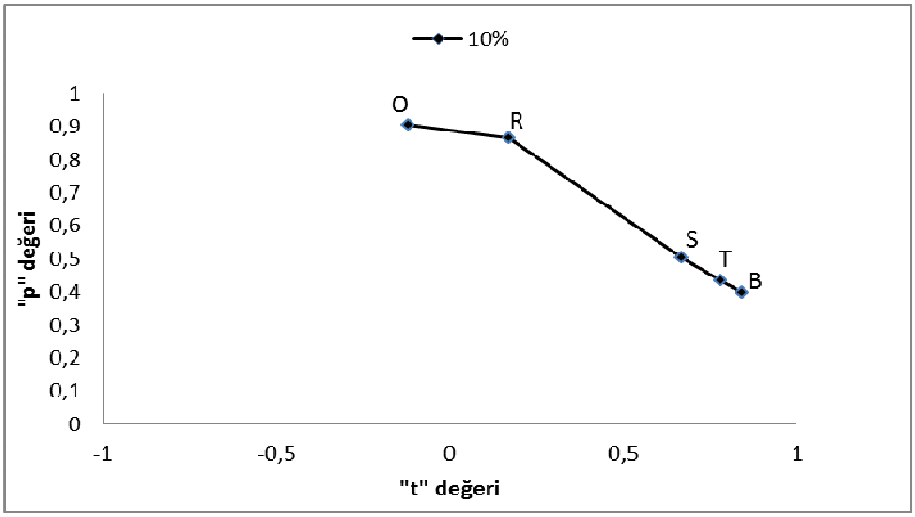
Şekil 4-15. 200 birimlik düşük korelasyonlu veri setinde %20 kayıp için t testi değerleri

Şekil 4-15.'e göre 200 birimlik düşük korelasyona sahip, %20 kayıp içeren veri setinde tam verilere en yakın değer yerine ortalama koyma yöntemi ile elde edilmiştir. Ayrıca tüm p değerleri >0.05 olduğu için analiz sonuçlarında 1. tip ve 2. tip hata gözlenmemiştir.



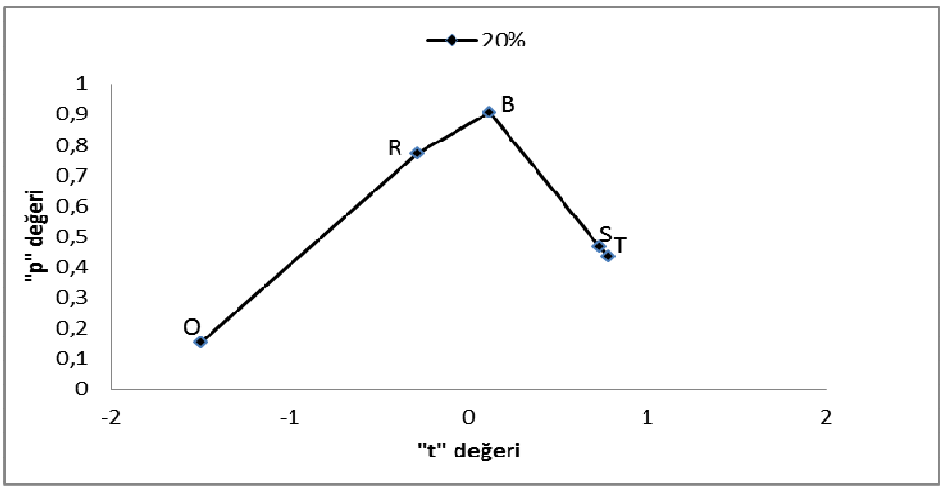
Şekil 4-16. 200 birimlik yüksek korelasyonlu veri setinde %5 kayıp için t testi değerleri

Şekil 4-16.'ya göre 200 birimlik yüksek korelasyona sahip, %5 kayıp içeren veri setinde tam verilere en yakın değer yerine ortalama koyma yöntemi ile elde edilmiştir. Ayrıca tüm p değerleri >0.05 olduğu için analiz sonuçlarında 1. tip ve 2. tip hata gözlenmemiştir.



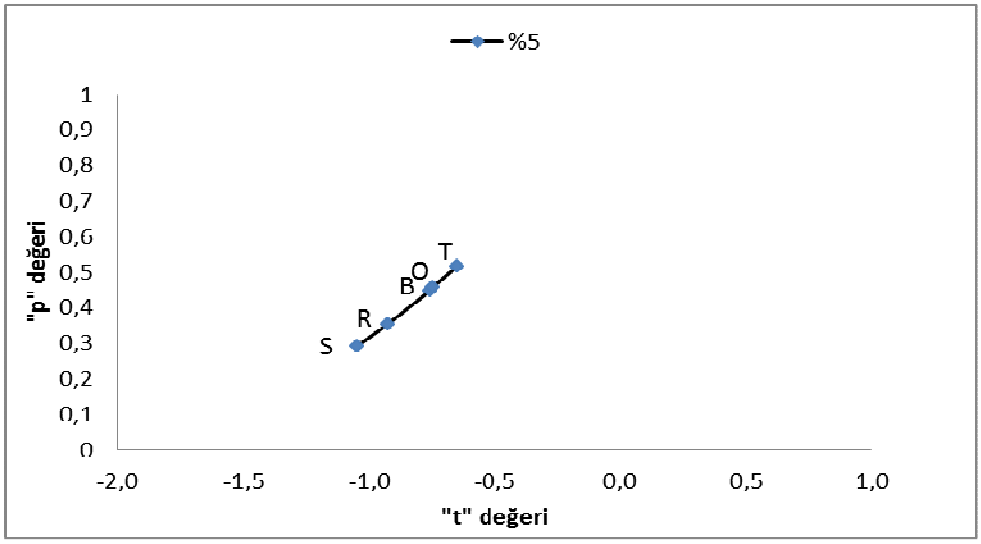
Şekil 4-17. 200 birimlik yüksek korelasyonlu veri setinde %10 kayıp için t testi değerleri

Şekil 4-17.'ye göre 200 birimlik yüksek korelasyona sahip, %10 kayıp içeren veri setinde tam verilere en yakın değer BM yöntemi ve silme yöntemi ile elde edilmiştir. Ayrıca tüm p değerleri >0.05 olduğu için analiz sonuçlarında 1. tip ve 2. tip hata gözlenmemiştir.



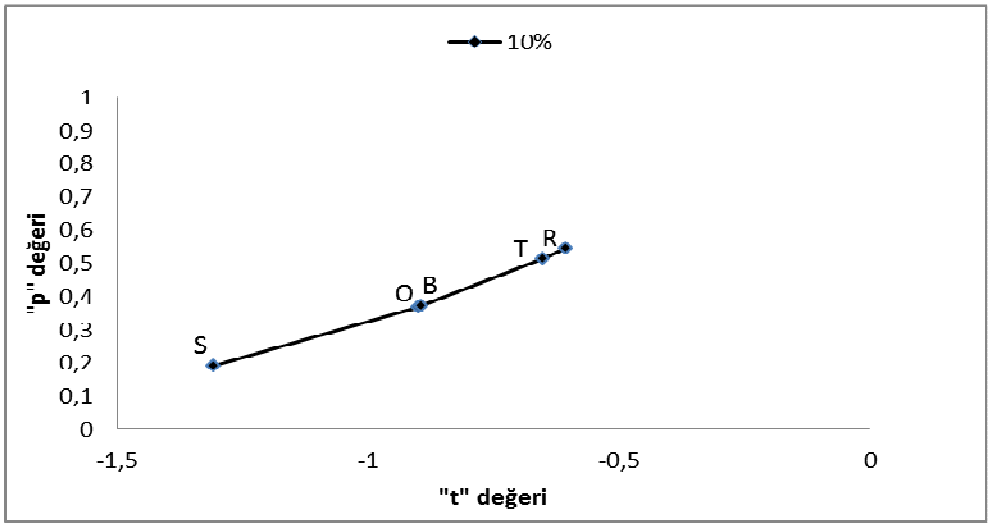
Şekil 4-18. 200 birimlik yüksek korelasyonlu veri setinde %20 kayıp için t testi değerleri

Şekil 4-18.'e göre 200 birimlik yüksek korelasyona sahip, %20 kayıp içeren veri setinde tam verilere en yakın değer silme yöntemi ile elde edilmiştir. Ayrıca tüm p değerleri >0.05 olduğu için analiz sonuçlarında 1. tip ve 2. tip hata gözlenmemiştir.



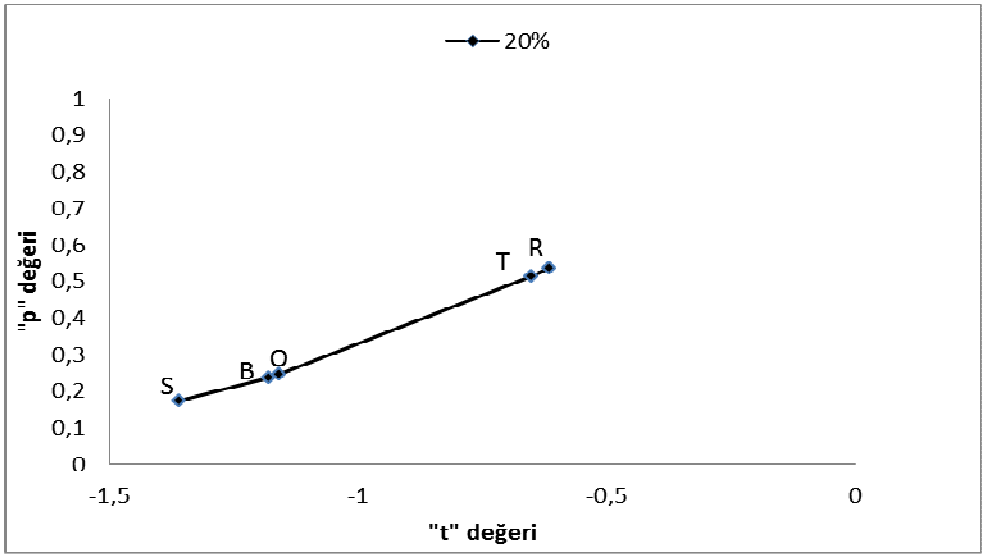
Şekil 4-19. 400 birimlik düşük korelasyonlu veri setinde %5 kayıp için t testi değerleri

Şekil 4-19.'a göre 400 birimlik düşük korelasyona sahip, %5 kayıp içeren veri setinde tam verilere en yakın değer BM ve yerine ortalama koyma yöntemi ile elde edilmiştir. Ayrıca tüm p değerleri >0.05 olduğu için analiz sonuçlarında 1. tip ve 2. tip hata gözlenmemiştir.



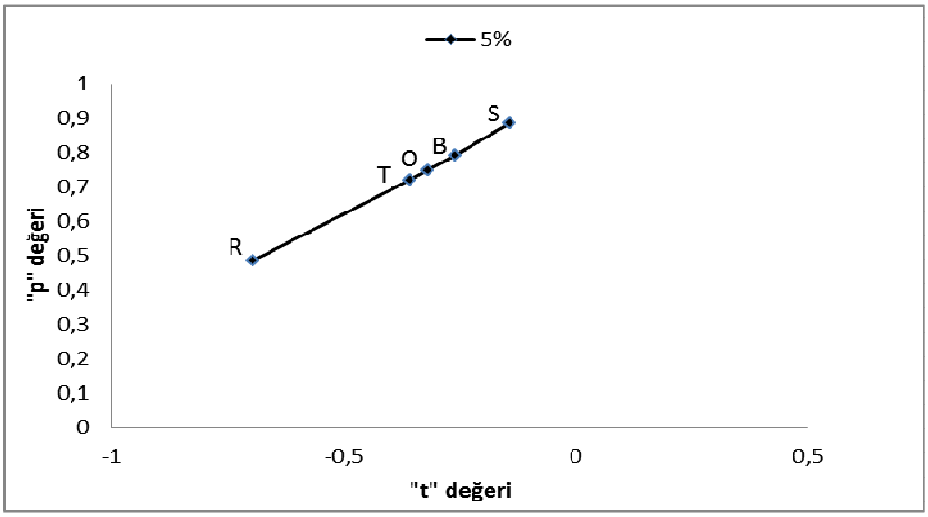
Şekil 4-20. 400 birimlik düşük korelasyonlu veri setinde %10 kayıp için t testi değerleri

Şekil 4-20.'ye göre 400 birimlik düşük korelasyona sahip, %10 kayıp içeren veri setinde tam verilere en yakın değer regresyon yöntemi ile elde edilmiştir. Ayrıca tüm p değerleri >0.05 olduğu için analiz sonuçlarında 1. tip ve 2. tip hata gözlenmemiştir.



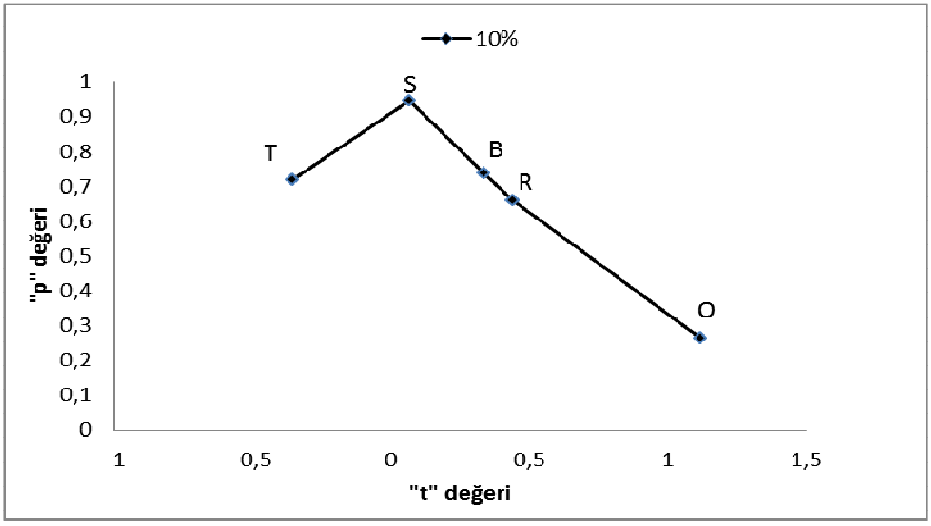
Şekil 4-21. 400 birimlik düşük korelasyonlu veri setinde %20 kayıp için t testi değerleri

Şekil 4-21.'e göre 400 birimlik düşük korelasyona sahip, %20 kayıp içeren veri setinde tam verilere en yakın değer regresyon yöntemi ile elde edilmiştir. Ayrıca tüm p değerleri >0.05 olduğu için analiz sonuçlarında 1. tip ve 2. tip hata gözlenmemiştir.



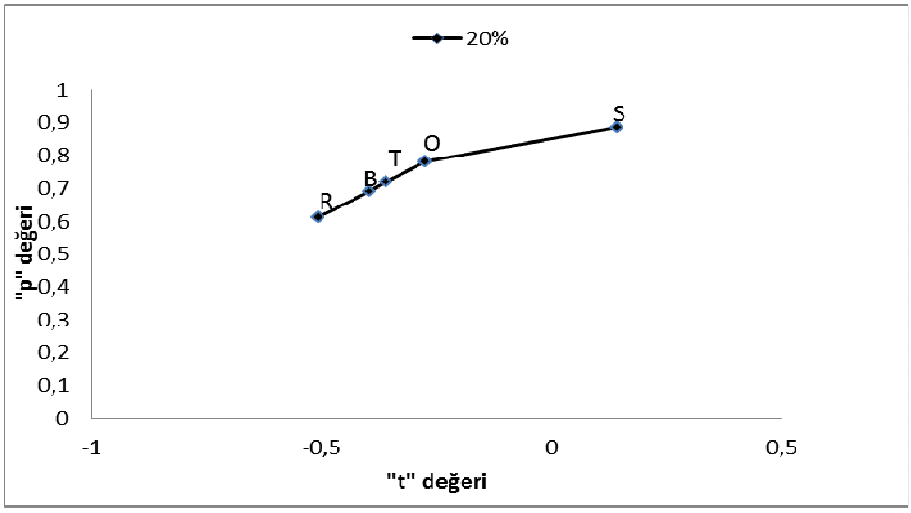
Şekil 4-22. 400 birimlik yüksek korelasyonlu veri setinde %5 kayıp için t testi değerleri

Şekil 4-22.'ye göre 400 birimlik yüksek korelasyona sahip, %5 kayıp içeren veri setinde tam verilere en yakın değer yerine ortalama koyma yöntemi ile elde edilmiştir. Ayrıca tüm p değerleri >0.05 olduğu için analiz sonuçlarında 1. tip ve 2. tip hata gözlenmemiştir.



Şekil 4-23. 400 birimlik yüksek korelasyonlu veri setinde %10 kayıp için t testi değerleri

Şekil 4-23.'e göre 400 birimlik yüksek korelasyona sahip, %10 kayıp içeren veri setinde kullanılan yöntemlerin tam veri setine ait t ve p değerlerine yakın sonuçlar vermediği görülmüştür. Buna rağmen tüm p değerleri >0.05 olduğu için analiz sonuçlarında 1. tip ve 2. tip hata gözlenmemiştir.



Şekil 4-24. 400 birimlik yüksek korelasyonlu veri setinde %20 kayıp için t testi değerleri

Şekil 4-24.'e göre 400 birimlik yüksek korelasyona sahip, %20 kayıp içeren veri setinde tam verilere en yakın BM ve yerine ortalama koyma yöntemi ile elde edilmiştir. Ayrıca tüm p değerleri >0.05 olduğu için analiz sonuçlarında 1. tip ve 2. tip hata gözlenmemiştir.

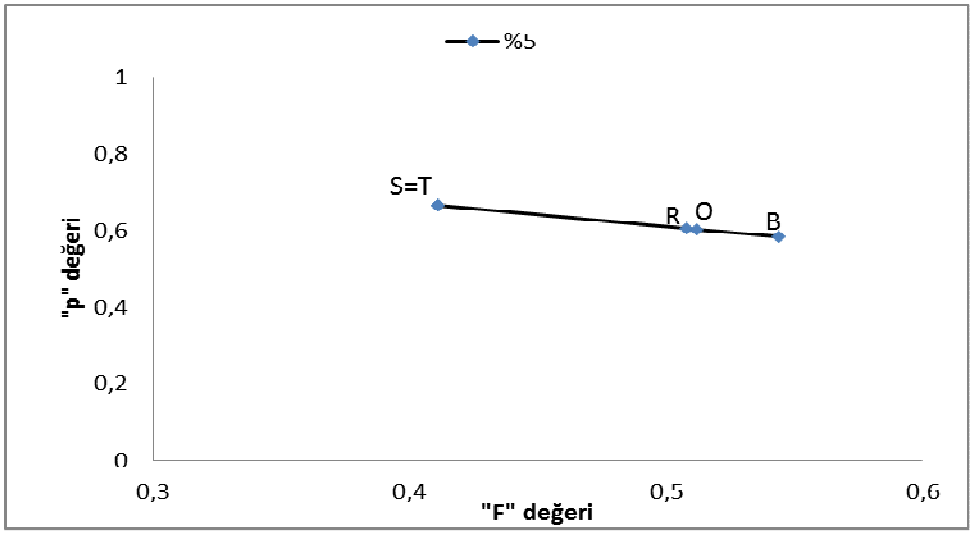
4.2. ANOVA Testine Ait Bulgular ve Yorumlar

Tablo 4-2. Tüm veri setlerine ait ANOVA testinin F ve p değerleri

			TAM		SİLME		ORTALAMA		REGRESYON		BM	
	N	%	F	p	F	p	F	p	F	P	F	p
Düşük	50	0	0,411	0,665								
		5			0,411	0,665	0,512	0,603	0,508	0,605	0,544	0,584
		10			0,704	0,502	0,659	0,522	0,544	0,584	0,696	0,504
		20			0,279	0,760	0,770	0,469	0,247	0,782	0,890	0,418
	100	0	4,236	0,017								
		5			5,013	0,009	4,961	0,009	4,641	0,012	5,036	0,008
		10			4,121	0,021	0,194	0,824	2,671	0,074	2,533	0,085
		20			1,010	0,374	2,069	0,132	1,725	0,184	1,665	0,195
	200	0	0,305	0,738								
		5			0,399	0,672	0,400	0,671	0,351	0,704	0,388	0,679
		10			0,185	0,831	0,123	0,885	0,139	0,871	0,114	0,893
		20			0,138	0,871	0,127	0,881	0,129	0,879	0,900	0,914
	400	0	2,061	0,129								
		5			1,981	0,139	1,984	0,139	2,518	0,082	2,026	0,133
		10			1,710	0,183	1,442	0,238	1,023	0,361	1,483	0,228
		20			2,055	0,132	1,092	0,336	2,972	0,052	1,157	0,316
Yüksek	50	0	0,662	0,521								
		5			0,758	0,474	0,778	0,465	0,572	0,568	0,642	0,531
		10			1,032	0,367	0,504	0,608	0,818	0,447	0,693	0,505
		20			0,017	0,983	0,005	0,995	0,999	0,376	0,715	0,494
	100	0	0,495	0,611								
		5			0,357	0,700	0,361	0,698	0,499	0,609	0,492	0,613
		10			0,132	0,876	0,194	0,824	0,613	0,544	0,512	0,601
		20			0,245	0,784	0,734	0,483	0,501	0,607	0,623	0,538
	200	0	0,141	0,868								
		5			0,398	0,672	0,399	0,672	0,172	0,842	0,189	0,828
		10			0,340	0,712	0,328	0,706	0,119	0,888	0,148	0,863
		20			1,687	0,192	0,558	0,573	0,213	0,808	0,193	0,825
	400	0	0,746	0,059								
		5			0,608	0,545	0,609	0,544	0,724	0,485	0,761	0,468
		10			0,596	0,552	0,779	0,459	0,615	0,541	0,609	0,544
		20			0,126	0,881	0,783	0,458	0,932	0,394	0,650	0,522

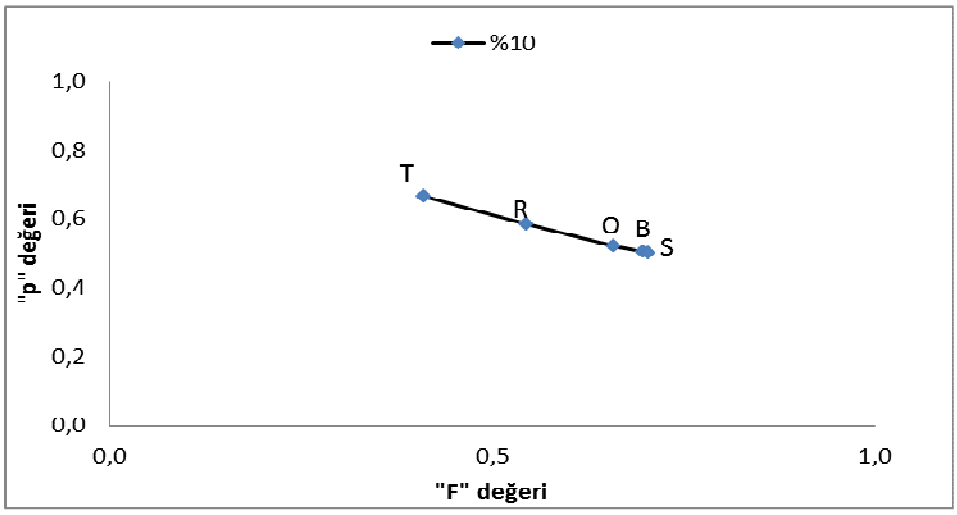
Tablo 4-2.'de tam (kayıp olmayan) verilerin ve belirli oranlarda kayıp söz konusu iken silme ve atama yolları ile tamamlanmış yeni veri setlerine uygulanmış ANOVA testine ait F ve p değerleri bulunmaktadır. Veri setlerindeki farklı miktardaki

kayıpların giderilmesine uygun yöntemin belirlenmesinde, tam verilere ait sonuçlara yakınlığı/benzerliği göz önüne alınmıştır. Grafiklerde ANOVA testine ait F değerleri yatay eksende, p değerlerinin ise düşey eksende karşılık geldiği noktalarla belirtilmiştir. Böylece tam (Tam veri=T) verilere ait nokta çiftlerine ne kadar benzer/yakın değerler oluşmuş ise (Silme yöntemi=S, Yerine ortalama koyma yöntemi=O, Regresyon yöntemi=R, Beklenti maksimizasyonu Yöntemi=B) seçilen yöntemin gerçeğe yakın olabileceği söylenebilir.



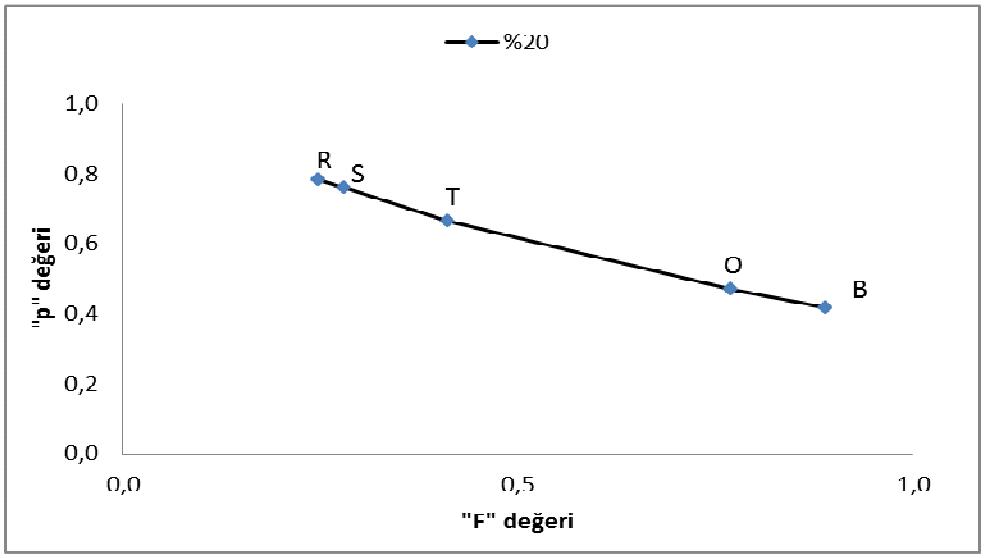
Şekil 4-25. 50 birimlik düşük korelasyonlu %5 kayıp için ANOVA değerleri

Şekil 4-25.'e göre 50 birimlik düşük korelasyona sahip, %5 kayıp içeren veri setinde tam verilere en yakın değer silme yöntemi ile elde edilmiştir.



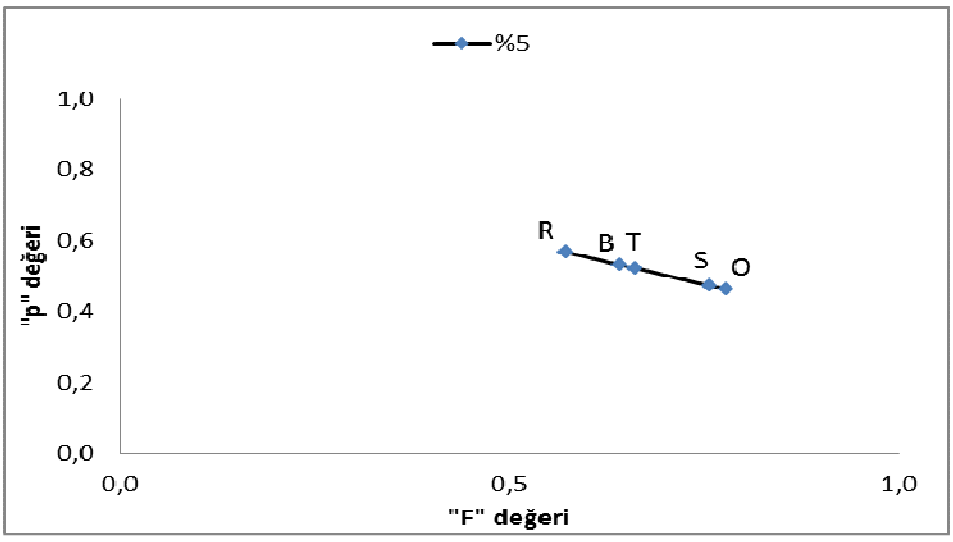
Şekil 4-26. 50 birimlik düşük korelasyonlu veri setinde %10 kayıp için ANOVA değerleri

Şekil 4-26.'ya göre 50 birimlik düşük korelasyona sahip, %10 kayıp içeren veri setinde tam verilere en yakın değer regresyon yöntemi ile elde edilmiştir.



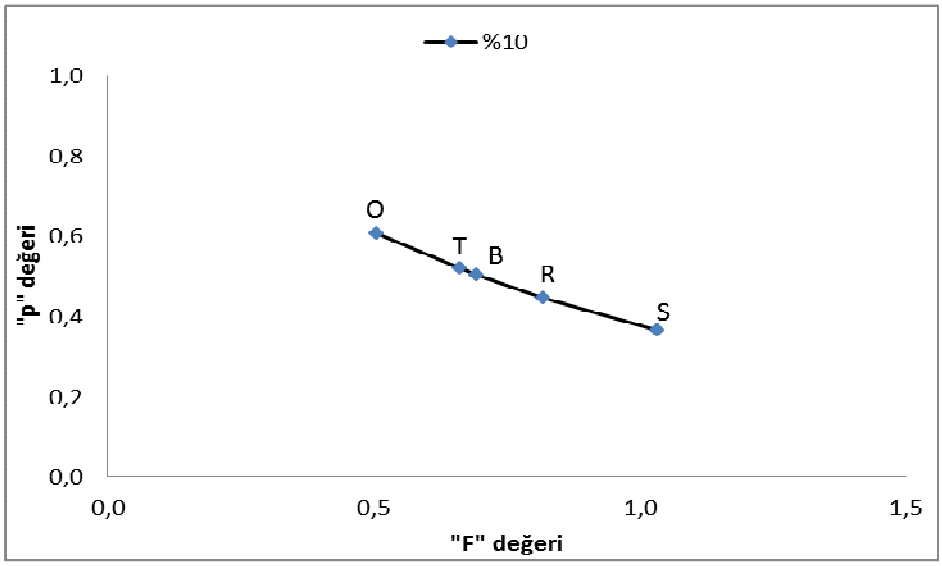
Şekil 4-27. 50 birimlik düşük korelasyonlu veri setinde %20 kayıp için ANOVA değerleri

Şekil 4-27.'ye göre 50 birimlik düşük korelasyona sahip, %20 kayıp içeren veri setinde tam verilere en yakın değer silme yöntemi ile elde edilmiştir.



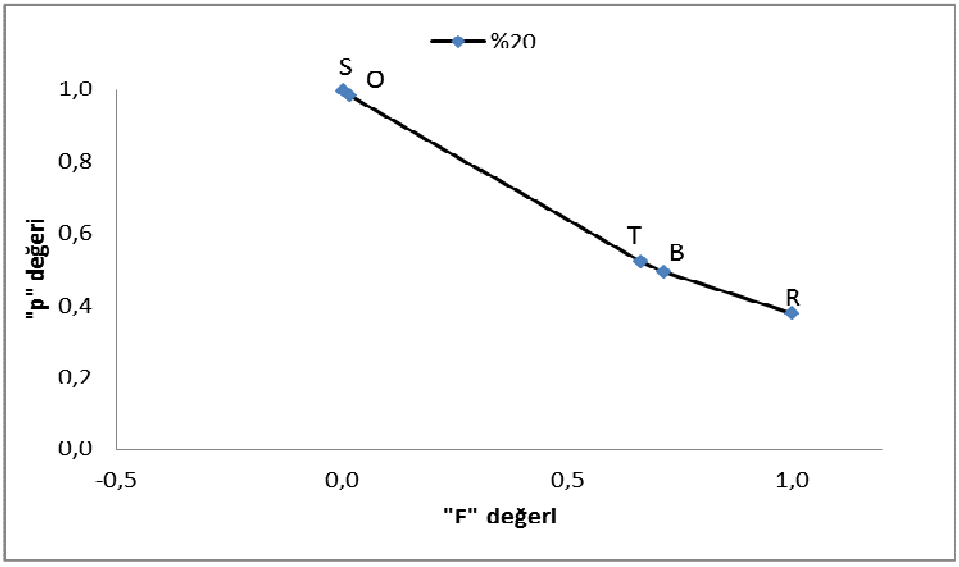
Şekil 4-28. 50 birimlik yüksek korelasyonlu veri setinde %5 kayıp için ANOVA değerleri

Şekil 4-28.'e göre 50 birimlik yüksek korelasyona sahip, %5 kayıp içeren veri setinde tam verilere en yakın BM yöntemi ile elde edilmiştir.



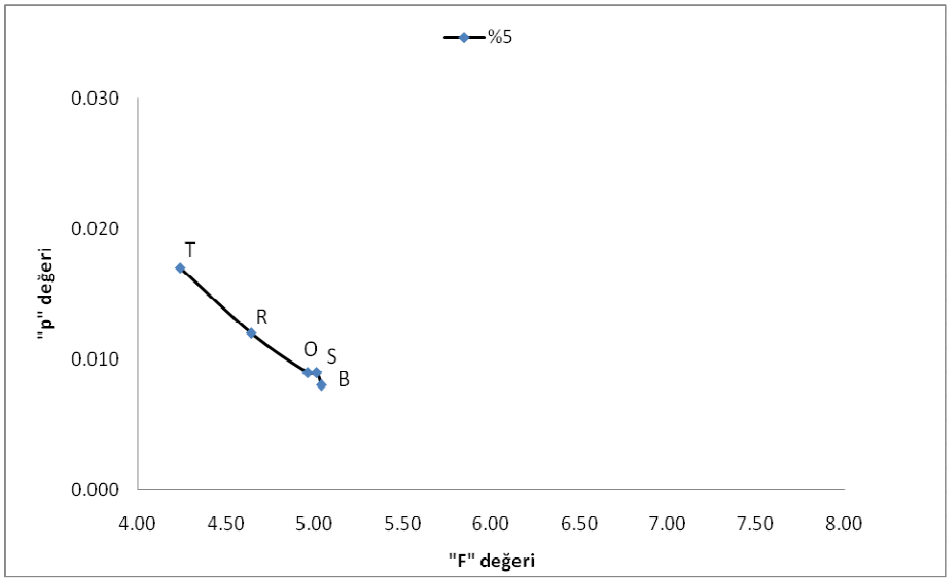
Şekil 4-29. 50 birimlik yüksek korelasyonlu veri setinde %10 kayıp için ANOVA değerleri

Şekil 4-29.'a göre 50 birimlik yüksek korelasyona sahip, %10 kayıp içeren veri setinde tam verilere en yakın BM yöntemi ile elde edilmiştir.



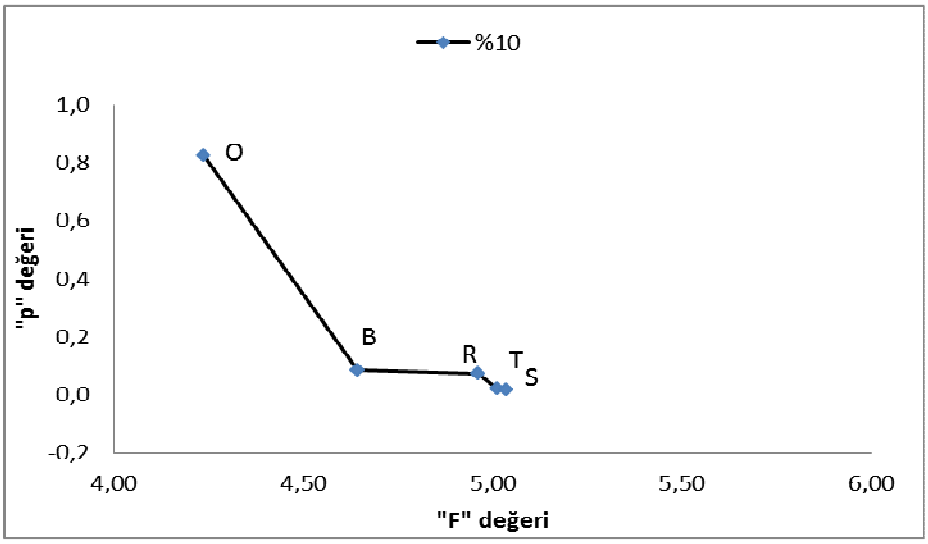
Şekil 4-30. 50 birimlik yüksek korelasyonlu veri setinde %20 kayıp için ANOVA değerleri

Şekil 4-30.'a göre 50 birimlik yüksek korelasyona sahip, %20 kayıp içeren veri setinde tam verilere en yakın BM yöntemi ile elde edilmiştir.



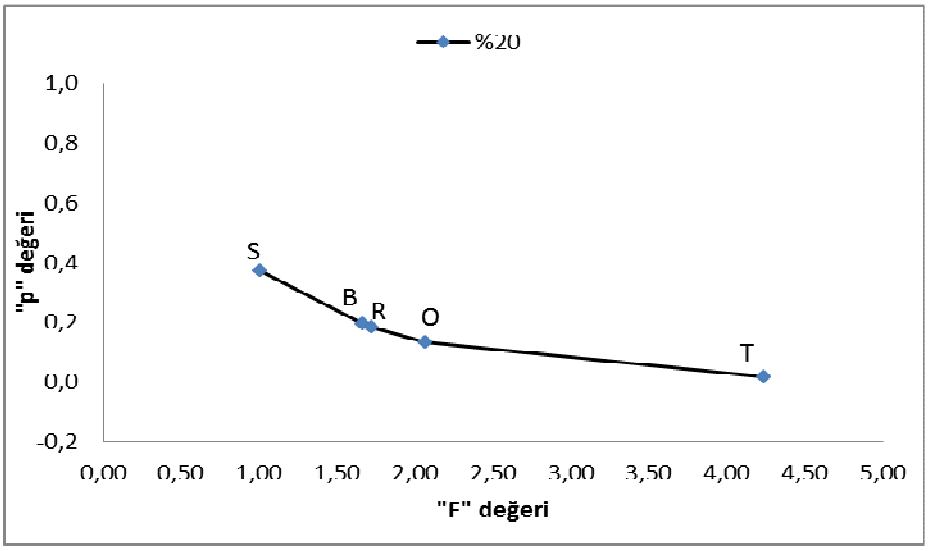
Şekil 4-31. 100 birimlik düşük korelasyonlu veri setinde %5 kayıp için ANOVA değerleri

Şekil 4-31.'e göre 100 birimlik düşük korelasyona sahip, %5 kayıp içeren veri setinde tam verilere en yakın yerine ortalama koyma ve regresyon yöntemi ile elde edilmiştir.



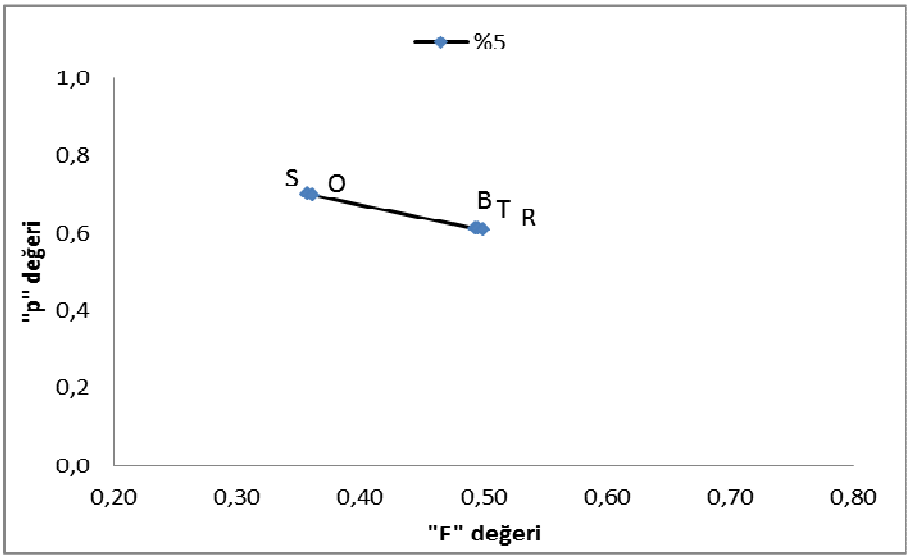
Şekil 4-32. 100 birimlik düşük korelasyonlu veri setinde %10 kayıp için ANOVA değerleri

Şekil 4-32.'ye göre 100 birimlik düşük korelasyona sahip, %10 kayıp içeren veri setinde tam verilere en yakın silme yöntemi ile elde edilmiştir.



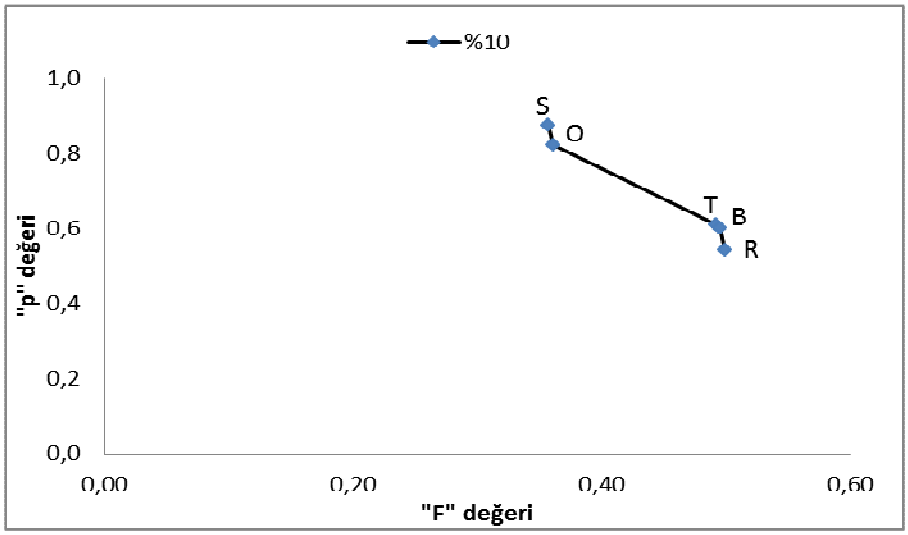
Şekil 4-33. 100 birimlik düşük korelasyonlu veri setinde %20 kayıp için ANOVA değerleri

Şekil 4-33.'e göre 100 birimlik düşük korelasyona sahip, %20 kayıp içeren veri setinde tam verilere yakın uygun değer hiçbir yöntemde elde edilememiştir.



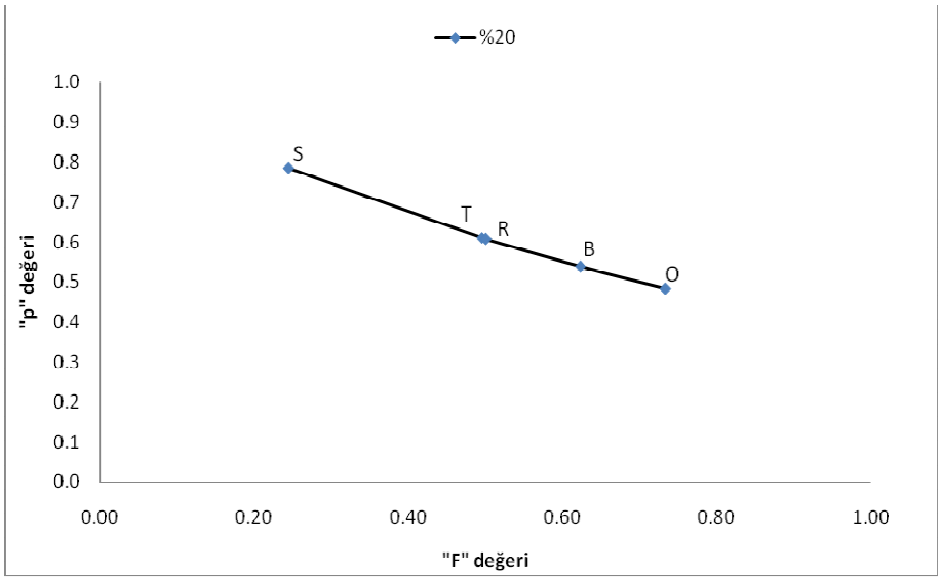
Şekil 4-34. 100 birimlik yüksek korelasyonlu veri setinde %5 kayıp için ANOVA değerleri

Şekil 4-34.'e göre 100 birimlik yüksek korelasyona sahip, %5 kayıp içeren veri setinde tam verilere en yakın BM ve regresyon yöntemi ile elde edilmiştir.



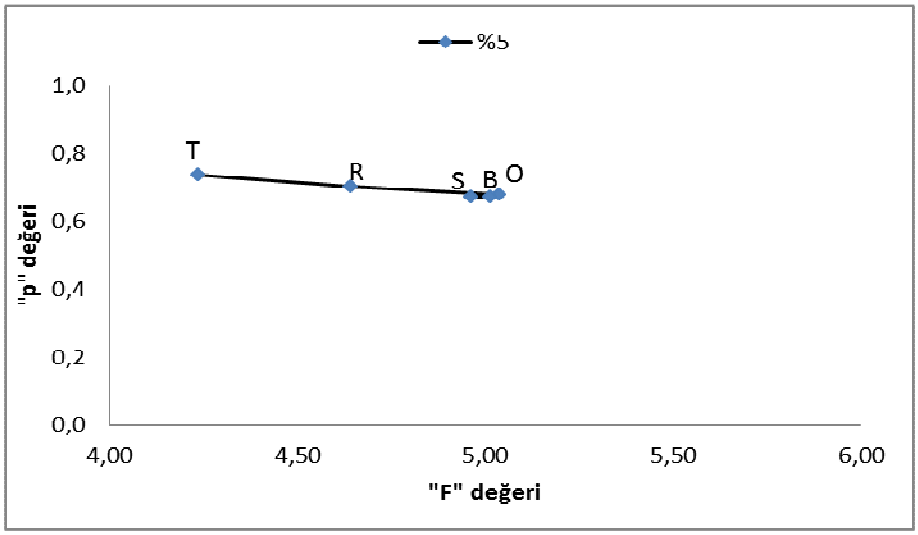
Şekil 4-35. 100 birimlik yüksek korelasyonlu veri setinde %10 kayıp için ANOVA değerleri

Şekil 4-35.'e göre 100 birimlik yüksek korelasyona sahip, %10 kayıp içeren veri setinde tam verilere en yakın BM yöntemi ile elde edilmiştir.



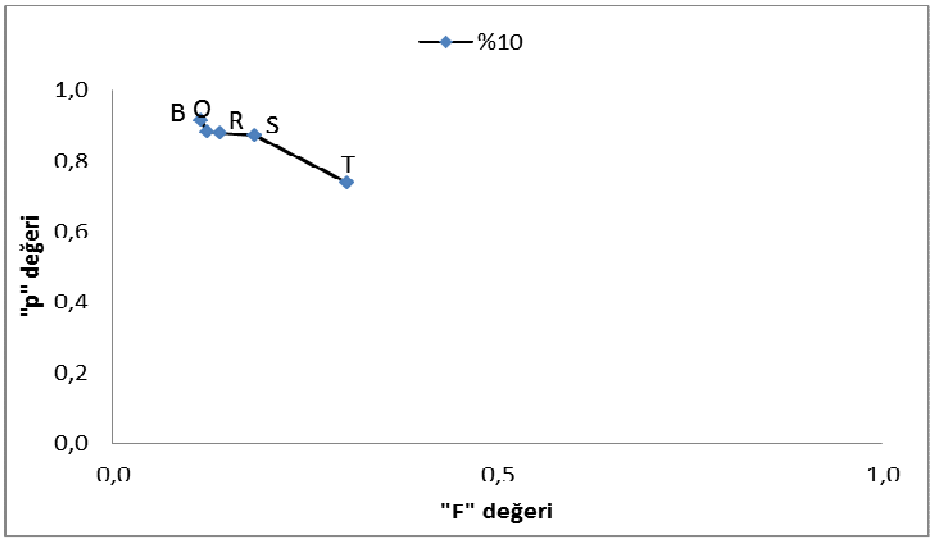
Şekil 4-36. 100 birimlik yüksek korelasyonlu veri setinde %20 kayıp için ANOVA değerleri

Şekil 4-36.'ya göre 100 birimlik yüksek korelasyona sahip, %20 kayıp içeren veri setinde tam verilere en yakın regresyon yöntemi ile elde edilmiştir.



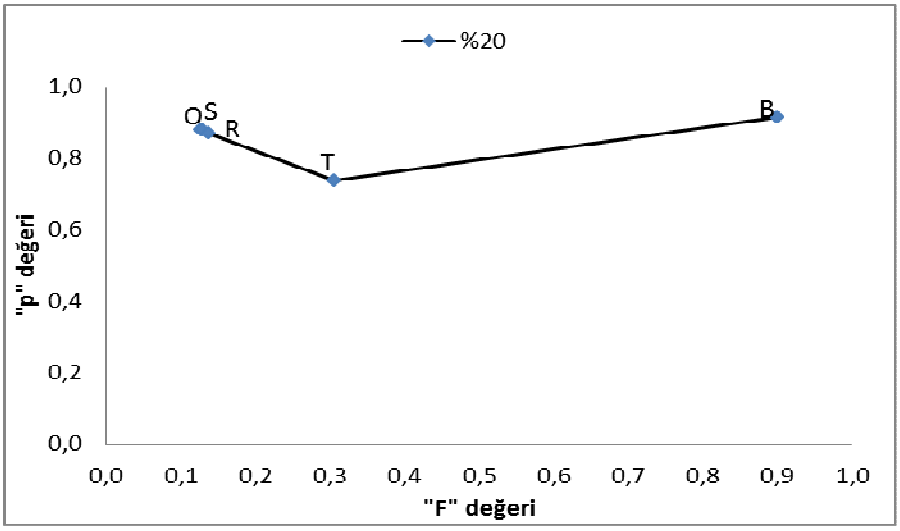
Şekil 4-37. 200 birimlik düşük korelasyonlu veri setinde %5 kayıp için ANOVA değerleri

Şekil 4-37.'e göre 200 birimlik düşük korelasyona sahip, %5 kayıp içeren veri setinde tam verilere yakın değerler hiçbir yöntemle elde edilememiştir.



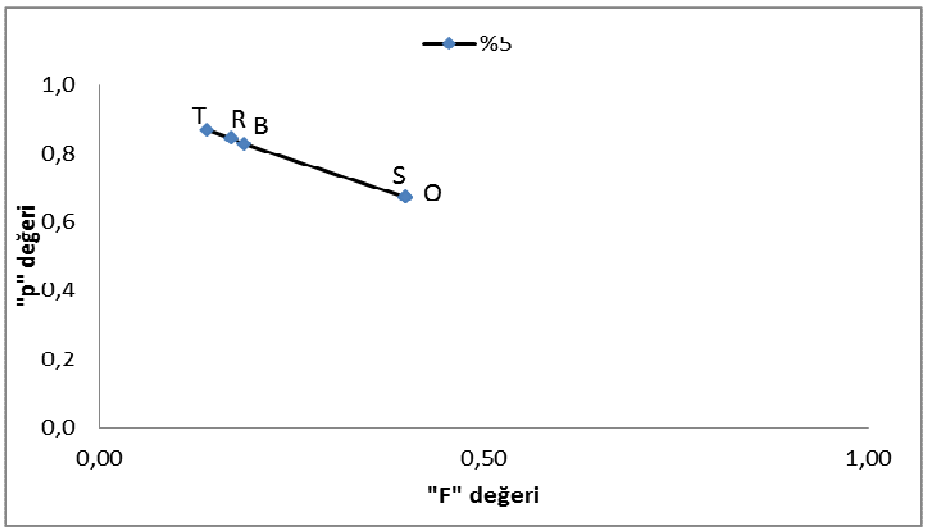
Şekil 4-38. 200 birimlik düşük korelasyonlu veri setinde %10 kayıp için ANOVA değerleri

Şekil 4-38.'e göre 200 birimlik düşük korelasyona sahip, %10 kayıp içeren veri setinde tam verilere yakın değerler hiçbir yöntemle elde edilememiştir.



Şekil 4-39. 200 birimlik düşük korelasyonlu veri setinde %20 kayıp için ANOVA değerleri

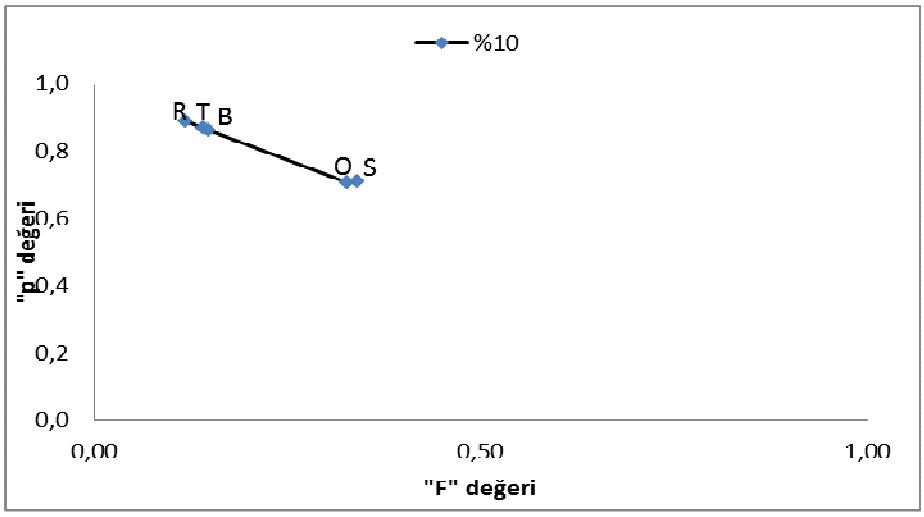
Şekil 4-39.'e göre 200 birimlik düşük korelasyona sahip, %20 kayıp içeren veri setinde tam verilere yakın değerler hiçbir yöntemle elde edilememiştir.



Şekil 4-40. 200 birimlik yüksek korelasyonlu veri setinde %5 kayıp için ANOVA değerleri

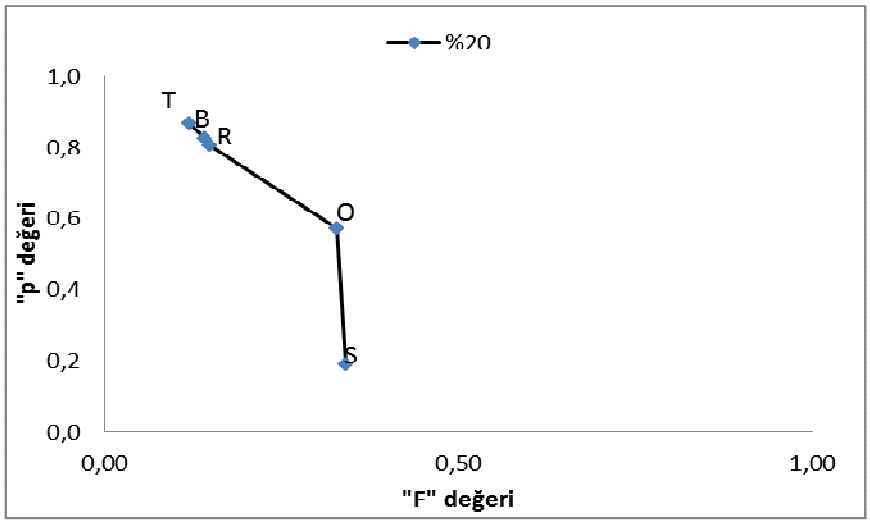
Şekil 4-40.'a göre 200 birimlik yüksek korelasyona sahip, %5 kayıp içeren veri

setinde tam verilere en yakın değerler regresyon ve BM yöntemi ile elde edilmiştir.



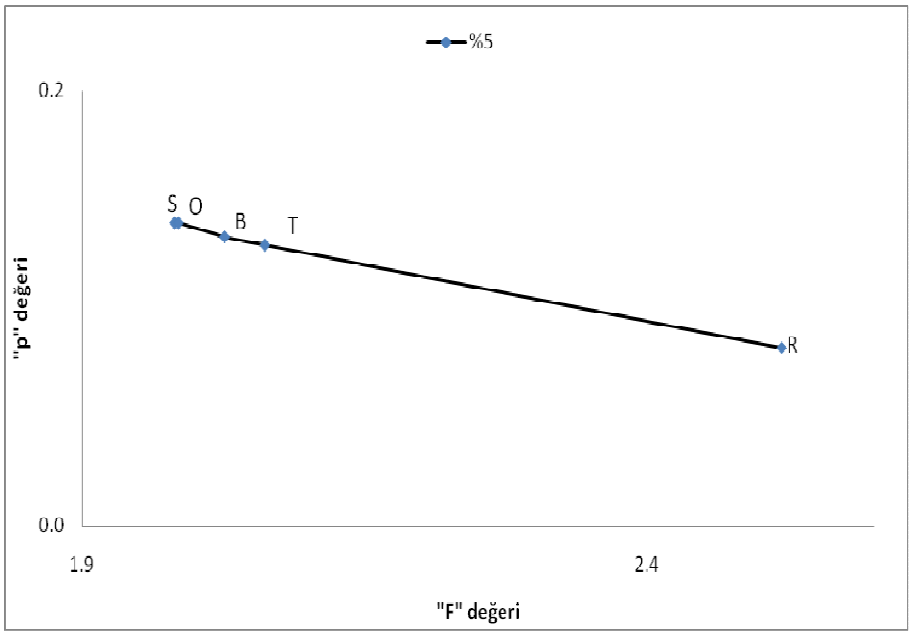
Şekil 4-41. -200 birimlik yüksek korelasyonlu veri setinde %10 kayıp için ANOVA değerleri

Şekil 4-41.'e göre 200 birimlik yüksek korelasyona sahip, %10 kayıp içeren veri setinde tam verilere en yakın değerler regresyon ve BM yöntemi ile elde edilmiştir.



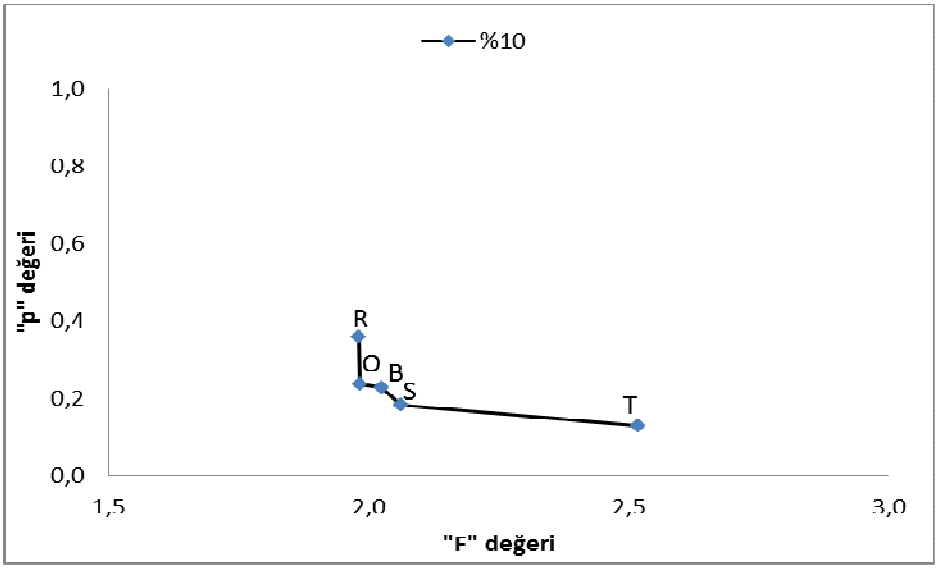
Şekil 4-42. 200 birimlik yüksek korelasyonlu veri setinde %20 kayıp için ANOVA değerleri

Şekil 4-42.'ye göre 200 birimlik yüksek korelasyona sahip, %20 kayıp içeren veri setinde tam verilere en yakın değerler regresyon ve BM yöntemi ile elde edilmiştir.



Şekil 4-43. 400 birimlik düşük korelasyonlu veri setinde %5 kayıp için ANOVA değerleri

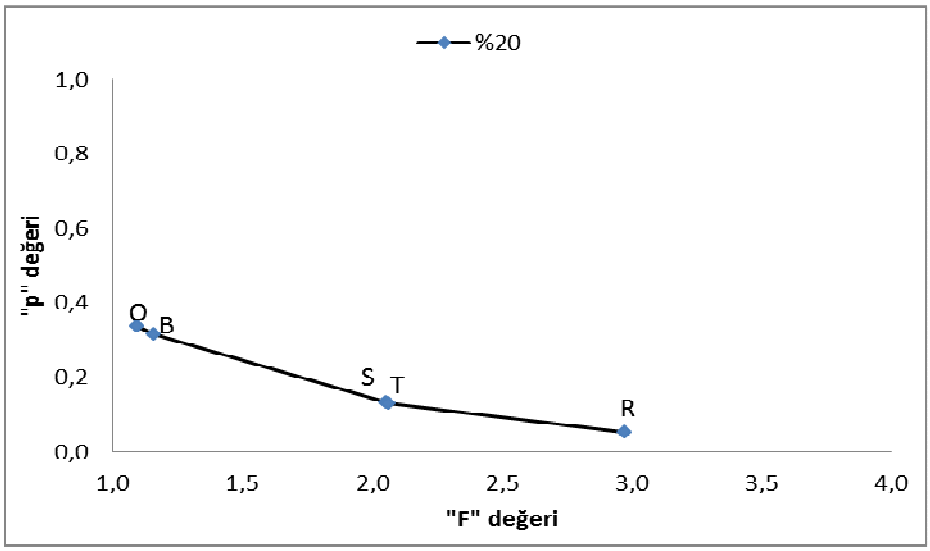
Şekil 4-43.'e göre 400 birimlik düşük korelasyona sahip, %5 kayıp içeren veri setinde tam verilere yakın değer BM yöntemi ile elde edilmiştir.



Şekil 4-44. 400 birimlik düşük korelasyonlu veri setinde %10 kayıp için ANOVA değerleri

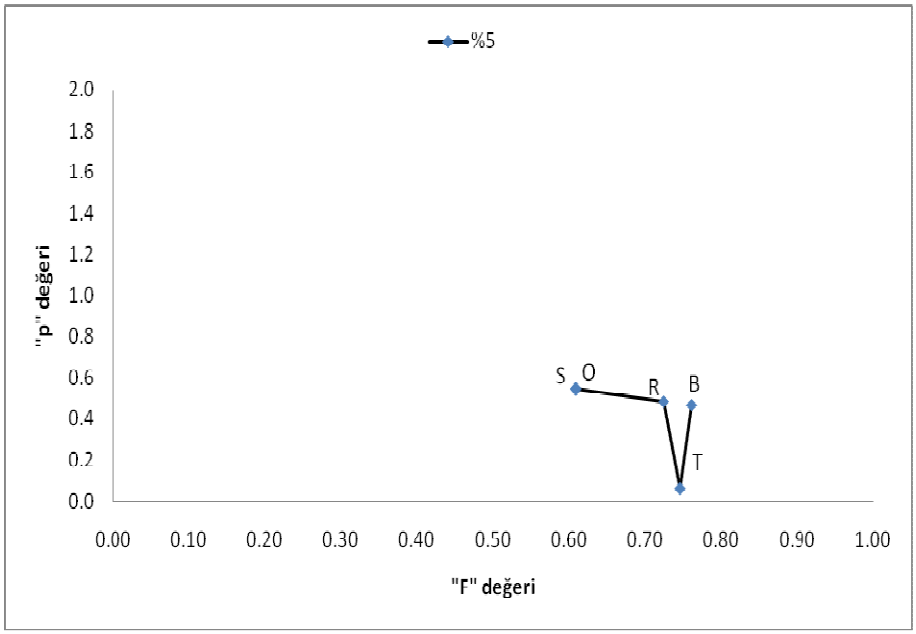
Şekil 4-44.'e göre 400 birimlik düşük korelasyona sahip, %10 kayıp içeren veri

setinde tam verilere yakın değerler hiçbir yöntemle elde edilememiştir.



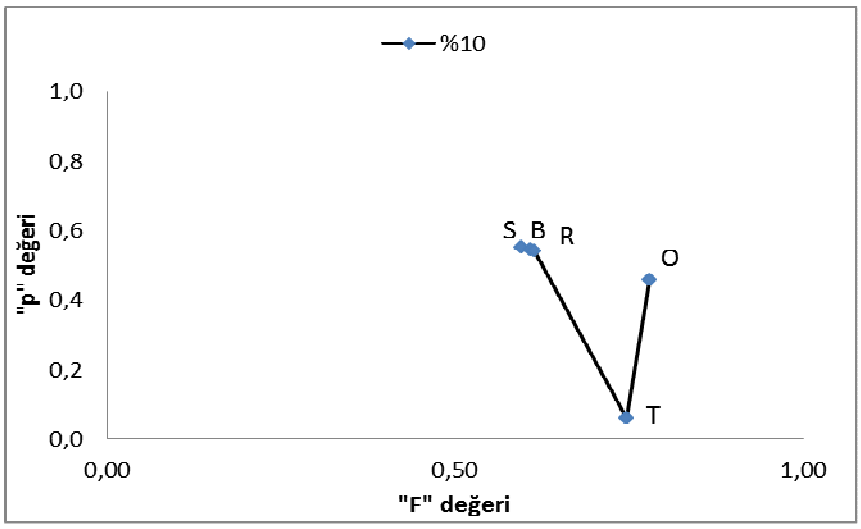
Şekil 4-45. 400 birimlik düşük korelasyonlu veri setinde %20 kayıp için ANOVA değerleri

Şekil 4-45.'e göre 400 birimlik düşük korelasyona sahip, %20 kayıp içeren veri setinde tam verilere en yakın değerler silme yöntemi ile elde edilmiştir.



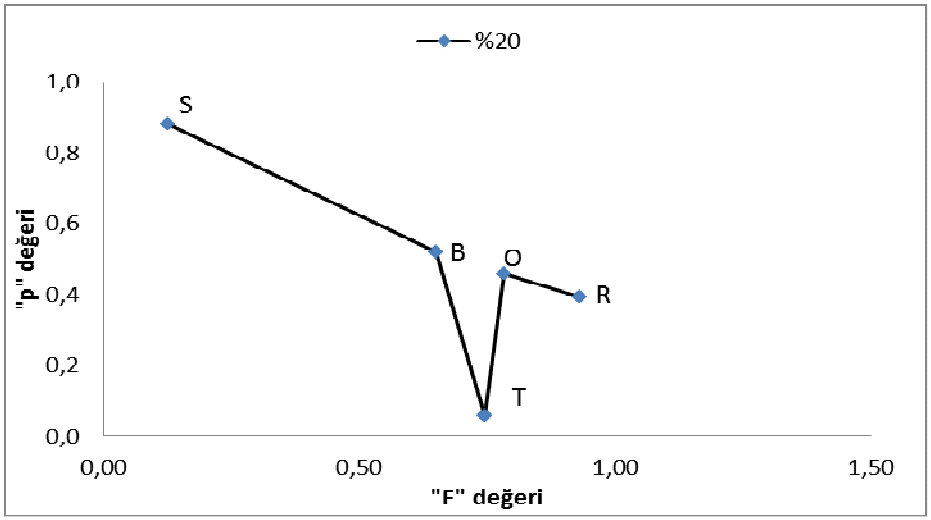
Şekil 4-46. 400 birimlik yüksek korelasyonlu veri setinde %5 kayıp için ANOVA değerleri

Şekil 4-46.'ya göre 400 birimlik yüksek korelasyona sahip, %5 kayıp içeren veri setinde tam verilere yakın değerler regresyon ve BM yöntemi ile elde edilmiştir.



Şekil 4-47. 400 birimlik yüksek korelasyonlu veri setinde %10 kayıp için ANOVA değerleri

Şekil 4-47.'ye göre 400 birimlik yüksek korelasyona sahip, %10 kayıp içeren veri setinde tam verilere yakın değerler hiçbir yöntemle elde edilememiştir.



Şekil 4-48. 400 birimlik yüksek korelasyonlu veri setinde %20 kayıp için ANOVA değerleri

Şekil 4-48.'e göre 400 birimlik yüksek korelasyona sahip, %20 kayıp içeren veri setinde tam verilere yakın değerler hiçbir yöntemle elde edilememiştir.

BÖLÜM V

5. Tartışma ve Sonuç

Türetilen tüm veri setlerinin silme yöntemiyle veya yerine ortalama koyma, BM ve regresyon gibi atama yöntemleri sonucu elde edilen yeni veri setlerinin analizleri sonucunda, kullanılan yöntemler arasındaki farklar, farklı korelasyona sahip farklı büyüklükteki veri setlerinde oldukça farklılaşmıştır. Örneğin, düşük korelasyonlu değişkenlere sahip verilerin analizleri sonucu farklı gruplara ait ortalamaların saptanmasında, 50 birim ve 100 birim gibi veri setinin hacminin küçük olma durumunda regresyon ve BM yöntemleri işe yarar yöntemler olarak kullanılabilir. 200 birim ve 400 birim gibi veri hacminin arttığı durumlarda etkin yöntemler regresyon ve yerine ortalama koyma olarak görülmektedir. Ayrıca düşük korelasyonlu veri setlerinde silme yönteminin etkin bir yöntem olmadığı da görülmüştür. Yüksek korelasyonlu veri setlerinde kullanılan yöntemlerin araştırmamızda kullanılan örnek veri setleri için genelleme yapılabilecek bir farklılık oluşturmadığı görülmüştür. Bu durum Baygül (2005)' ün yaptığı çalışma ile benzerlik göstermektedir. Ancak bu durum farklı büyüklükteki farklı kayıp miktarına sahip veri setleri için farklılıkların ortaya çıkabileceğinin de ipuçlarını vermiştir.

50 birimlik veri setlerinin %5 kayıp olduğu durumlarda yapılan analizlerde, farklı büyüklükteki diğer veri setlerine göre daha anlamlı sonuçlar elde edilmiştir. t testi için BM yöntemi, tam veri setinden elde edilen sonuçlara oldukça yakın değerler vermiştir. Benzer şekilde 100 birimlik veri setlerinin %10 kayıp olması durumunda, tam değerlere en yakın sonuçlar regresyon yöntemi, 200 birimlik veri setlerinde %5 kayıp olma durumunda yerine ortalama koyma yöntemi, 400 birimlik veri setlerinde %5 kayıp olma durumunda ise yerine ortalama koyma yöntemi diğer yöntemlere göre daha etkin sonuçlar vermiştir. Bu durum da Bal (2003)' ın yaptığı çalışma ile benzerlik

göstermektedir.

Farklı büyüklükteki farklı kayıp miktarına sahip veri setlerinde kullanılan yöntemler sonucu ANOVA testinin değerleri incelendiğinde, düşük korelasyonlu veri grubu için herhangi bir genelleme yapılamazken; yüksek korelasyonlu veri setlerinde BM ve regresyon yönteminin tam değerler oldukça yakın sonuçlar verdiği görülmüştür. Böylece BM' nin etkin yöntem olduğu literatür ile benzerlik göstermektedir.

Analizler sonucunda ortaya çıkan p değerleri göz önüne alındığında istatistikte sıkça rastlanılan 1. tip hata durumu ile karşılaşılmamıştır. Ancak 200 birimlik düşük korelasyona sahip veri setinde %10 kayıp olması durumunda regresyon yönteminin, benzer şekilde 100 birimlik yüksek korelasyona sahip veri setinde %20 kayıp durumunda ise regresyon ve yerine ortalama koyma yöntemlerinin1. tip hataya daha yakın değerler verdiği dikkate alınmalıdır.

5.1. Öneriler

Bu çalışmada farklı büyüklükteki farklı kayıp miktarı içeren veri setlerine silme yöntemi, yerine ortalama koyma yöntemi, regresyon yöntemi, BM yöntemi uygulanmıştır. Araştırmacılar diğer yöntemleri kullanarak en etkili yöntem konusunda farklı çalışmalar yapabilirler.

Bu çalışmada düşük ve yüksek korelasyonlu 50 birim, 100 birim, 200 birim, 400 birimlik veri setleri kullanılmıştır. Daha genel sonuçlar elde edilmesi açısından örneklemdeki veri setlerinin birim sayıları arttırılabilir.

Bu çalışmada normal dağılıma sahip 50 birim, 100 birim, 200 birim, 400 birimlik veri setleri kullanılmıştır. Araştırmacılar normal dağılıma sahip olmayan veri setleri üzerinde de çalışabilirler.

Bu çalışmada farklı veri setlerinin %5, %10, %20 kayıp olması durumu

incelenmiştir. Araştırmacıların farklı kayıp oranları üzerinde çalışmaları önerilebilir.

Bu çalışmada kayıp olma durumu TROK yapısı olarak belirlenmiştir. Araştırmacılar ROK yapısına uygun veri setlerindeki kayıp sorununu gidermeye yönelik çalışmalar yapabilirler.

Bu çalışmada farklı büyüklükteki farklı kayıp miktarı içeren veri setlerine t testi ve ANOVA analizleri uygulanmıştır. Araştırmacılar kullanılan silme ve atama yöntemlerinin farklı istatistiksel yöntemler üzerindeki sonuçlarını inceleyebilirler.

Bu çalışmada farklı büyüklükteki farklı kayıp içeren türetilmiş veri setlerinin tek faktör çatısı altında incelemesi yapılmıştır. Araştırmacılar farklı faktör sayılarında, farklı yöntemler kullanarak en etkili yöntemi belirleyebilirler.

KAYNAKÇA

Afiffi A. And Elashoff R. M. (1966). Missing Observations in Multivariate Statistics: I. Review of the Literature, Journal of the American Statistical Association, Vol. 61, No. 315, pp. 595-604.

Allison P. D. (2001). Missing Data, Sage University Papers Series on Quantitative Applications in the Social Sciences, Thousands Oaks, CA, Sage.

Alpar, R. (2003). Uygulamalı Çok Değişkenli İstatistiksel Yöntemlere Giriş-1, Nobel Kitabevi.

Bal C. (2003). Çok Gruplu Veri Setlerinde Kayıp Gözlem Sorununun Çözümlenmesi ve Sağlık Alanında Bir Uygulama, Doktora Tezi, Eskişehir: Eskişehir Osmangazi Üniversitesi Sağlık Bilimleri Enstitüsü.

Baygül A, (2007). Kayıp Veri Analizinde Sıklıkla Kullanılan Etkin Yöntemlerin Değerlendirilmesi, Yüksek Lisans Tezi, İstanbul: İstanbul Üniversitesi Sağlık Bilimleri Enstitüsü.

Cheema, J. (2012). Handling Missing Data in Educational Researchusing SPSS. Unpublished Doctora lDissertation. George Mason University, USA.

Çokluk Ö. , Kayrı M., (2011). Kayıp Değerlere Yaklaşık Değer Atama Yöntemlerinin Ölçme Araçlarının Geçerlik Ve Güvenirliği Üzerindeki Etkisi, Kuram ve Uygulamada Eğitim Bilimleri, Kış; 289-309.

Dempster, A.P., Laird, N. M, and Rubin, D. (1977). Maximum Like Lihood Estimation From İncomplete Data Viathe EM Algorithm, Journal Royal Statistical Soc. Vol: 39

Enders C. K. (2011). Analyzing Longitudinal Data With Missing Values.

Evens, Fiona H. (2003). Detectingfish in Underwater Video Using The EM Algorithm, Proceeding Of The IEEE International Conference on Image Processing, Barcelona.

Gildea, L. And Hofmann, T. (1999). Topic-Based language models using EM, (http://www.cs.brown.edu/people/th/papers/GildeaHofmann-EUROSPEECH99.pdf , 31. 01. 2013, Erişildi).

Iturria, S.J. and Blangero, J. (2000). An EM Algorithm Forobtaining Maximum Like Lihoo Destimates In Themulti-Pheno Type Variance Components Link Age Model, Ann. Hum. Genet. Vol. 64.

Kalaycı Ş. (2008), SPSS Uygulamalı Çok Değişkenli İstatistik Teknikleri.

Karasar N. (2004), Bilimsel Araştırma Yöntemleri, Nobel Yayıncılık.

Krishner, S., Cadez, I.V., Smyth, P., and Kamath, C.(2003), Learning toclassify galax yshapes using the EM algorithm, Advances in Neural Information Processing.

Little R. J. A. (1998). A Test of Missing Completely at Random for Multivariate Data With MissingValues. Journal of The American Statistical Association 38, 1198-1202.

Little R. J. A. , Rubin D. R.(2002). Statistical Analysis With Missing Data, Second Edition, Wiley, New York.

Oğuzlar A. (2001). Alan Araştırmalarında Kayıp Değer Problemi Ve Çözüm Önerileri, V. Ulusal Ekonometri ve İstatistik Sempozyumu, Çukurova Üniversitesi İİBF Ekonometri Bölümü, Adana, 19-22 Eylül.

Özdamar K. (2002). Paket Programlarla İstatistiksel Veri Analizi-1, Kaan Kitabevi, Eskişehir.

Özdamar K. (2010). Paket Programlarla İstatistiksel Veri Analizi-1, Kaan Kitabevi, Eskişehir.

Rubin, D.R. (1976). Inference and missing data. Biometrika, 63 (3), 581-592.

Satıcı, E. Ve Kadılar, C. (2009), Kayıp Gözlem Olması Durumunda Kitle Ortalamasının Tahmini, Anadolu Üniversitesi Bilim ve Teknoloji Dergisi, 10(2), 549-556.

Schafer J. Multiple Imputatıon FAQ page. (http://sites.stat.psu.edu/~jls/mifaq.html, 31.01.2013, Erişildi)

Sezgin E. , Çelik Y.(2013). Veri Madenciliğinde Kayıp Veriler İçin Kullanılan Yöntemlerin Karşılaştırılması, Akademik Bilişim Konferansı, Akdeniz Üniversitesi, 23-25 Ocak 2013.

Shukla D., Singhai R. (2011). Thakur N. S., A New Imputation Methodfor Missing Attribute Values in Data Mining, Journal of Applied Computer Science & Mathematics, no:10 (5).

Stehoven D. J. Bühlmann P., (2011). Miss Forest-Nonparametric Missing Value Imputation for Mixed- Type Data.

Tabachnick, B.G. ve Fidel (2001). L.S. Using multivariate statistics (4^{th} ed.). Needham Heights, MA: Allyn& Bacon.

Tarsitano, A. ve Falcone, M. (2011). Missing Values Adjust mentfor Mixed Tyre Data.

Yazıcıoğlu, Y. ve Erdoğan, S. (2007). Spss Uygulamalı Bilimsel Araştırma Yöntemleri. Ankara, Detay Yayıncılık.

Yazıcı, F. (2005). EM Algoritması ve Uzantıları, Yüksek Lisans Tezi, Ankara: Hacettepe Üniversitesi Fen Bilimleri Enstitüsü.

Yozgatligil, C., Purutcuoglu, V., Yazıcı, C. ve Batmaz, İ. (2011). Validity of Homogeneity Testsfor Meteorological Time Series Data: A Simulation Study. "Proceedings of the 58th World Statistics Congress (ISI2011)".

EKLER

Ek 1. R-Studıo Programında Türetilen Verilere Ait Örnek Program Kodu

```
library(mvtnorm)
covariance <- 0.95 * sqrt(0.9) * sqrt(0.001)*sqrt(0.001)
sigma <- matrix(c(0.9, covariance, covariance, covariance, 0.001, covariance, covariance, covariance, 0.001),
nrow=3, byrow=TRUE)
xx <- rmvnorm(n=1000,
mean=c(1,1,1),
sigma=sigma) cor(xx[, 1], xx[, 2]) cor(xx[, 1], xx[, 3]) cor(xx[, 2], xx[, 3])
plot(xx[, 1], xx[, 2])
plot(xx[, 1], xx[, 3])
plot(xx[, 2], xx[, 3])
```