T.C.
HACETTEPE ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

MAKİNE ÖĞRENİMİNE DAYALI ÇOKLU ATAMA YÖNTEMLERİNİN PARAMETRİK OLMAYAN ÇOKLU KARŞILAŞTIRMALARA UYGULANMASI VE BENZETİM ÇALIŞMASI

Tuncay YANARATEŞ

Biyoistatistik Programı
DOKTORA TEZİ

ANKARA
2025

T.C.

HACETTEPE ÜNİVERSİTESİ

SAĞLIK BİLİMLERİ ENSTİTÜSÜ

MAKİNE ÖĞRENİMİNE DAYALI ÇOKLU ATAMA YÖNTEMLERİNİN PARAMETRİK OLMAYAN ÇOKLU KARŞILAŞTIRMALARA UYGULANMASI VE BENZETİM ÇALIŞMASI

Tuncay YANARATEŞ

Biyoistatistik Programı

DOKTORA TEZİ

TEZ DANIŞMANI

Prof. Dr. Erdem KARABULUT

ANKARA

2025

Makine Öğrenimine Dayalı Çoklu Atama Yöntemlerinin Parametrik Olmayan Çoklu Karşılaştırmalara Uygulanması ve Benzetim Çalışması

Öğrenci: Tuncay YANARATEŞ

Danışman: Prof. Dr. Erdem KARABULUT

Bu tez çalışması 10.03.2025 tarihinde jürimiz tarafından "Biyoistatistik Programı" nda doktora tezi olarak kabul edilmiştir.

Jüri Başkanı: **Prof. Dr. Mehtap AKÇİL OK**

(Başkent Üniversitesi)

Üye: **Doç. Dr. Beyza DOĞANAY**

(Ankara Üniversitesi)

Üye: **Doç. Dr. Osman DAĞ**

(Hacettepe Üniversitesi)

Üye: **Dr. Öğr. Üyesi Sevilay KARAHAN**

(Hacettepe Üniversitesi)

Üye: **Dr. Öğr. Üyesi H. Yağmur ZENGİN**

(Hacettepe Üniversitesi)

Bu tez Hacettepe Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin ilgili maddeleri uyarınca yukarıdaki jüri tarafından uygun bulunmuştur.

19.03.2025

Prof. Dr. Müge YEMİŞCİ ÖZKAN

Enstitü Müdürü

YAYIMLAMA VE FİKRİ MÜLKİYET HAKLARI BEYANI

Enstitü tarafından onaylanan lisansüstü tezimin/raporumun tamamını veya herhangi bir kısmını, basılı (kağıt) ve elektronik formatta arşivleme ve aşağıda verilen koşullarla kullanıma açma iznini Hacettepe Üniversitesine verdiğimi bildiririm. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet haklarım bende kalacak, tezimin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanım hakları bana ait olacaktır.
Tezin kendi orijinal çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanılması zorunlu metinlerin yazılı izin alınarak kullandığımı ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederim.
Yükseköğretim Kurulu tarafından yayınlanan *"Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge"* kapsamında tezim aşağıda belirtilen koşullar haricince YÖK Ulusal Tez Merkezi / H.Ü. Kütüphaneleri Açık Erişim Sisteminde erişime açılır.

Enstitü / Fakülte yönetim kurulu kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren 2 yıl ertelenmiştir. (1)

Enstitü / Fakülte yönetim kurulunun gerekçeli kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren ... ay ertelenmiştir. (2)

Tezimle ilgili gizlilik kararı verilmiştir. (3)

12/03/2025

Tuncay YANARATEŞ

i

[1] *"Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge" Madde 6. 1. Lisansüstü tezle ilgili patent başvurusu yapılması veya patent alma sürecinin devam etmesi durumunda, tez danışmanının önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü veya fakülte yönetim kurulu iki yıl süre ile tezin erişime açılmasının ertelenmesine karar verebilir.*

Madde 6. 2. Yeni teknik, materyal ve metotların kullanıldığı, henüz makaleye dönüşmemiş veya patent gibi yöntemlerle korunmamış ve internetten paylaşılması durumunda 3. şahıslara veya kurumlara haksız kazanç imkanı oluşturabilecek bilgi ve bulguları içeren tezler hakkında tez danışmanının önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü veya fakülte yönetim kurulunun gerekçeli kararı ile altı ayı aşmamak üzere tezin erişime açılması engellenebilir.

*Madde 7. 1. Ulusal çıkarları veya güvenliği ilgilendiren, emniyet, istihbarat, savunma ve güvenlik, sağlık vb. konulara ilişkin lisansüstü tezlerle ilgili gizlilik kararı, tezin yapıldığı kurum tarafından verilir *. Kurum ve kuruluşlarla yapılan işbirliği protokolü çerçevesinde hazırlanan lisansüstü tezlere ilişkin gizlilik kararı ise, ilgili kurum ve kuruluşun önerisi ile enstitü veya fakültenin uygun görüşü üzerine üniversite yönetim kurulu tarafından verilir. Gizlilik kararı verilen tezler Yükseköğretim Kuruluna bildirilir.*
Madde 7.2. Gizlilik kararı verilen tezler gizlilik süresince enstitü veya fakülte tarafından gizlilik kuralları çerçevesinde muhafaza edilir, gizlilik kararının kaldırılması halinde Tez Otomasyon Sistemine yüklenir

** Tez danışmanının önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü veya fakülte yönetim kurulu tarafından karar verilir.*

ETİK BEYAN

Bu çalışmadaki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi, görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu, kullandığım verilerde herhangi bir tahrifat yapmadığımı, yararlandığım kaynaklara bilimsel normlara uygun olarak atıfta bulunduğumu, tezimin kaynak gösterilen durumlar dışında özgün olduğunu, Prof. Dr. Erdem KARABULUT danışmanlığında tarafımdan üretildiğini ve Hacettepe Üniversitesi Sağlık Bilimleri Enstitüsü Tez Yazım Yönergesine göre yazıldığını beyan ederim.

Tuncay YANARATEŞ

TEŞEKKÜR

Yüksek lisans eğitimimden sonra başladığım doktora programına beni kabul eden Hacettepe Üniversitesi Biyoistatistik Anabilim Dalı öğretim üyeleri başta olmak üzere tüm Hacettepe Üniversitesi Biyoistatistik Anabilim Dalı öğretim üyelerine teşekkürlerimi sunarım.

Bu süreçte benden desteğini esirgemeyen tez danışmanım Prof. Dr. Erdem KARABULUT'a teşekkürlerimi sunarım.

Hayatımın her döneminde olduğu gibi tez döneminde de beni teşvik eden anneme teşekkürlerimi borç bilirim.

Bu tez çalışmasında aşağıda belirtilen makaleden yararlanılmıştır.

Yanarateş, T and Karabulut, E. Applying machine learning-based multiple imputation methods to nonparametric multiple comparisons in longitudinal clinical studies. Journal of Biopharmaceutical Statistics. 2024. Copyright Taylor & Francis. https://doi.org/10.1080/10543406.2024.2444243.

ÖZET

Yanarateş, T. Makine Öğrenimine Dayalı Çoklu Atama Yöntemlerinin Parametrik Olmayan Çoklu Karşılaştırmalara Uygulanması ve Benzetim Çalışması. Hacettepe Üniversitesi Sağlık Bilimleri Enstitüsü Biyoistatistik Programı Doktora Tezi, Ankara, 2025. Aynı bireyler üzerinde tekrarlanan ölçümlerin yapıldığı uzunlamasına veri (bağımlı örneklemler), bireyler arasındaki potansiyel farklılıkları ortadan kaldırır. Uzunlamasına veride eksik veriler tasarım gereği veya rastgele oluşabilir. Skillings-Mack testi, normal dağılım göstermeyen eksik gözlemlere sahip k-bağımlı örneklemler için Friedman testi yerine kullanılır. Gruplar arasında anlamlı bir fark varsa, parametrik olmayan çoklu karşılaştırmalar yapılması gerekir. Bu çalışmada, normal dağılım göstermeyen eksik veriye sahip k-bağımlı örneklemlerin parametrik olmayan çoklu karşılaştırmalarına dört yöntem uygulayarak yenilikçi bir yaklaşım önerilmektedir. Dört yöntem, makine öğrenmesine dayalı iki parametrik olmayan çoklu atama yöntemi, bir parametrik olmayan atama yöntemi (rastgele sıcak deste ataması) ve liste bazında silme yöntemidir. Dört yöntem iki eksik veri mekanizması altında karşılaştırılmaktadır. Benzetim çalışmasında farklı senaryolar uygulandıktan sonra, liste bazında silme yöntemi (tam gözlemlerin kullanımı yöntemi) diğer yöntemlerden güç bakımından daha düşük bulunmuştur. İki parametrik olmayan çoklu atama yöntemi, iyi kontrol edilen tip 1 hatası olan orta ve küçük örnek boyutları için diğer yöntemlerden güç bakımından daha üstündür. Bu nedenle, eksik gözlemlere sahip k-bağımlı örneklemlerin parametrik olmayan çoklu karşılaştırmaları için makine öğrenimine dayalı çoklu atama yöntemlerinin kullanılmasını öneriyoruz. Ayrıca, önerilen yaklaşım gerçek bir veri seti üzerinde de uygulanmıştır. Bu örnekte, önerilen çoklu atama yöntemleri, başlangıçta eksik gözlemlerin olmadığı gerçek veri setindeki sonuçlara yakın sonuçlar vermiştir.

Anahtar Kelimeler: Uzunlamasına veri, eksik veri, tamamen rastgele eksik, rastgele eksik, çoklu atama

ABSTRACT

Yanarateş, T. Applying machine learning-based multiple imputation methods to nonparametric multiple comparisons and simulation study. Hacettepe University Graduate School of Health Sciences, PhD Thesis in Biostatistics, Ankara, 2025. Longitudinal data, in which repeated measurements are made on the same subjects, eliminate potential differences among the subjects. In longitudinal data, missing data can occur by design or completely random. The Skillings-Mack test is used instead of the Friedman test for longitudinal data with missing observations that are non-normally distributed. Nonparametric multiple comparisons need to be performed if a significant difference exists among groups. In this study, we propose a new approach by applying four methods to nonparametric multiple comparisons of longitudinal data that are non-normally distributed. The four methods are two nonparametric multiple imputation methods based on machine learning, one nonparametric imputation method (random hot deck imputation), and the listwise deletion method. We assume two missing data mechanisms. After implementing different scenarios in a simulation study, the listwise deletion method is inferior to the other methods. The two nonparametric multiple imputation methods are superior to the other methods for moderate and small sample sizes with well-controlled type 1 error. Therefore, we propose the two multiple imputation methods for nonparametric multiple comparisons of longitudinal data with missing observations. Moreover, the proposed approach was also applied on a real data set. In this example, the proposed multiple imputation methods yielded results similar to those of the real dataset without missing observations at the beginning.

Keywords: Longitudinal data, missing data, missing completely at random, missing at random, multiple imputation

İÇİNDEKİLER

SİMGELER VE KISALTMALAR

MCAR Tamamen Rastgele Eksik

MAR Rastgele Eksik

MNAR Rastgele Olmayan Eksik

LD Listesel Silme Yöntemi

HDI Sıcak Deste Ataması

FCS Tamamen Koşullu Belirleme

MICE Zincirlenmiş Denklemlerle Çoklu Atama

MICE-CART Sınıflandırma ve Regresyon Ağaçlarını Kullanan Zincirlenmiş Denklemlerle Çoklu Atama

MICE-RF Rastgele Orman Kullanan Zincirlenmiş Denklemlerle Çoklu Atama

GLMM Genelleştirilmiş Doğrusal Karışık Etki Modeli

μ Kitle ortalaması

$\sum$ Kovaryans matrisi

S^2 Örneklem varyansı

$\bar{X}$ Örneklem ortalaması

S Örneklem standart sapması

r Örneklem korelasyonu

k Deneme sayısı

n Örneklem için gözlem sayısı

ŞEKİLLER

TABLOLAR

1.GİRİŞ

Uzunlamasına veri (bağımlı örneklemler), bireyler üzerinde belirli bir zaman periyodu boyunca tekrarlı ölçümlerin yapıldığı verilerdir. Klinik çalışmalarda araştırmacılar, tedavinin hastalık süreci üzerinde zaman içindeki etkisini incelemek için bağımlı örneklemleri kullanırlar. Aynı bireylerde iki veya daha fazla farklı zamanda birden fazla ölçüm yapıldığında, bir örneklemdeki değerler diğer örneklemdeki değerleri etkiler ve bu durumda örneklemler bağımlı hale gelir. Bağımlı örneklemlerde, aynı bireyler üzerinde tekrarlanan ölçümler yapıldığı için bireyler arasındaki farklılıklar ortadan kalkar. İki bağımlı örneklem için normallik varsayımı karşılanırsa, bağımlı örneklemlerde t-testi kullanılır (1). Normallik varsayımı karşılanmazsa, Wilcoxon işaretli sıra testi kullanılır (2). K-bağımlı örneklemler için örneklem büyüklüğü yeterince büyükse veya normallik ve küresellik varsayımları karşılanıyorsa, tekrarlı ölçümlerde tek yönlü varyans analizi (ANOVA) kullanılır; aksi takdirde, Friedman testi kullanılır (3).

Uzunlamasına veride sıklıkla eksik veriyle karşılaşılmaktadır. Eksik veri, uzunlamasına veride önemli bir sorundur. Uzunlamasına veride çeşitli nedenlerle eksik gözlemler olabilir. Örneğin, bireyler bir tedaviyi reddedip sonra diğer tedaviyi alabilir, bireyler ölebilir veya çalışmadan ayrılabilir. Literatürde üç eksik veri mekanizması vardır: Tamamen rastgele eksik (missing completely at random) (MCAR), rastgele eksik (missing at random) (MAR) ve rastgele olmayan eksik (missing not at random) (MNAR) (4). Bu çalışmada, MNAR'ın tasarlanması oldukça zor olduğundan MCAR ve MAR'a odaklanılmıştır, çünkü MNAR'da değişkenin gözlemlenen değerleri eksik değerlerin yanlı tahminlerini vermektedir (5). Ayrıca, üç eksik veri tasarımı vardır: tek değişkenli, monoton ve monoton olmayan. Tek değişkenli tasarımda, yalnızca bir zamanda ölçülen değişkende eksik verileri vardır. Monoton tasarımda, birey bir kez çalışmadan ayrıldığında, bir daha çalışmada olmayacaktır. Eksik veri tasarımı monoton değilse, monoton olmayan veya rastgele (arbitrary) olarak adlandırılır (6). Literatürde, eksik gözlemleri olan iki bağımlı örneklem için bir terim (kısmi eşli veri) oluşturulmuştur. Ayrıca, k-bağımlı örneklemlerde eksik gözlemlerle karşılaşabilmektedir. Eksik veriler tasarıma göre (örneğin, araştırmacılar kasıtlı olarak bazı bilgileri toplamamaya karar verebilir ve

ilgili eksik veri tasarıma göre eksik olur) veya tamamen rastgele oluşabilir (7). Literatürde çeşitli örnekler vardır (bkz.[8]).

Uzunlamasına veride eksik gözlem durumunda geleneksel yöntem olarak bilinen tam gözlemlerin kullanımı yöntemi (listesel silme yöntemi) en basit ve en çok bilinen yöntemdir. Fakat bu yöntemin dezavantajları avantajlarına göre fazladır. Birinci dezavantajı örneklem büyüklüğünün azalması nedeniyle yanlı tahminler oluşmasıdır. Ayrıca istatistiksel güç azalır ve evren hakkında genelleme yapmak zorlaşır.

İstatistiksel olarak atama terimi verideki eksikleri doldurmak anlamında kullanılmaktadır. Atama yöntemleri iki bölümde incelenebilir: Tekli ve çoklu atama.

Little ve Rubin'e (9) göre tekli atama için iki çeşit yaklaşım vardır. Tekli atama yöntemleri açık ve örtük modelleme yöntemlerini içerir. Parametrik olmayan atama yöntemleri örtük modelleme yöntemleridir ve sıcak deste atamasını içerir. Parametrik olmayan bir atama yöntemi olarak Wang ve diğerleri (10), rastgele sıcak deste atama yönteminin uzunlamasına verilerle kullanılabileceğini göstermiştir. Uzunlamasına veriler için iki temel çoklu atama yaklaşımı vardır. Bunlar, ortak modelleme çok değişkenli normal atama (joint modeling) ve tamamen koşullu belirlemedir (fully conditional specification). Ortak modelleme çok değişkenli normal atama, normal dağılmış verileri varsaydığı için parametrik atama yöntemlerini içerir. Tamamen koşullu belirleme, normal dağılım varsayımına dayanmadığı için daha fazla esnekliğe sahiptir. Standart tamamen koşullu belirlemede (MICE) iki adet parametrik olmayan tahmin yöntemi vardır. İki yöntem, ağaç tabanlı ve makine öğrenmesi tabanlı yöntemlerdir ve normal dağılmayan verilerle kullanılırlar. Literatürde, uzunlamasına verilerle iki ağaç tabanlı yöntemin kullanıldığı çalışmalar bulunmaktadır. Ancak her iki yöntem de daha önce uzunlamasına verilerde parametrik olmayan çoklu karşılaştırmalar için kullanılmamıştır.

Skillings ve Mack (11), normal dağılım göstermeyen eksik gözlemlere sahip k-bağımlı örneklemler için alternatif bir test önermiştir. K-bağımlı örneklemlerde eksik gözlemler varsa, Skillings-Mack testinden sonra parametrik olmayan çoklu karşılaştırmalar yapılmalıdır, çünkü anlamlı bir Skillings-Mack test değeri gruplar arasında anlamlı bir fark olduğunu gösterir. Bu çalışmada, normal dağılım

göstermeyen eksik k-bağımlı örneklemlerin parametrik olmayan çoklu karşılaştırmalarına üç parametrik olmayan atama yöntemini ve liste bazında silme yöntemini uygulayarak yeni bir yaklaşım önerilmektedir. Bu yöntemler uygulandıktan sonra, k-bağımlı örneklemlerin parametrik olmayan çoklu karşılaştırmaları için bazı testler kullanılabilir. En iyi bilinen testlerden biri, Nemenyi'nin tezinde (12) bulunabilecek olan Nemenyi testidir. Bu çalışmada eksik k-bağımlı örneklemlerin parametrik olmayan çoklu karşılaştırmaları için dört farklı atama yönteminin performansı karşılaştırılmaktadır.

2. GENEL BİLGİLER

Rastgele blok tasarımı, iki faktör içeren k-bağımlı örneklemleri ele alır. Rastgele blok tasarımı için yaygın bir model aşağıdaki gibidir:

$Y_{ij} = \mu + \beta_i + \tau_j + \in_{ij}$

Y_{ij} , i'inci bloktaki j'inci denemeye karşılık gelen yanıttır, μ genel ortalamadır, β_i i'inci blok etkisi ve τ_j j'inci deneme etkisidir. Hatalar, ε_{ij}, bağımsızdır ve sürekli bir dağılım fonksiyonuyla özdeş olarak dağılmıştır. Ayrıca, toplam k deneme ve n blok olduğunda, n_{ij}, i'inci bloktaki j'inci deneme için bir gözlem olup olmadığını gösterir. Genel yokluk hipotezi, deneme etkilerinin hepsinin özdeş olmasıdır (yani, H_0: $\tau_1 = \cdots = \tau_k$).

Friedman testi, eşit örneklem büyüklüklerine sahip bu tür k-bağımlı deneme grubu için parametrik olmayan alternatiftir (3).

Friedman testi istatistiği eşitlik 2.1'de gösterilmektedir.

$$\chi_R^2=[\frac{12}{nk(k+1)} \sum_{j=1}^{k}(RA_j)^2]\text{-}3n(k+1) \qquad (2.1)$$

RA_j: Her bir sütundaki rank toplamıdır.

K-bağımlı örneklemlerde, veriler tasarım gereği veya tamamen rastgele eksik olabilir. Skillings ve Mack, veriler normal olmayan bir şekilde dağıldığında monoton olmayan eksik veri yapısıyla eksik blok tasarımı için bir test önermiştir. Skillings-Mack testinde, yalnızca bir gözlemi olan herhangi bir blok kaldırılır. Her blok içinde, gözlemler 1'den k_i'ye kadar sıralanır, burada k_i, i. bloktaki tedavi sayısıdır ve eşitlik varsa, sıraların ortalaması alınır. Y_{ij} için sıra r_{ij} olacaktır, ancak bir gözlem eksik

olduğunda, (k_i+1) / 2 değeri kullanılır. Daha sonra ayarlanmış bir tedavi toplamı, A_j hesaplanır. A_j nin hesaplanması eşitlik 2.2'de gösterilmektedir.

$$A_j=\sum_{i=1}^{n}\left(\frac{12}{k_i+1)}\right)^{1/2}\left(r_{ij}-\frac{k_i+1}{2}\right), \quad j=1,\ldots,k \text{ için} \qquad (2.2)$$

Daha sonra $A_1,\ldots,A_k$ üzerinde tedavi toplamlarının kovaryansı için bir kovaryans yapısı üretilir. **A =** $(A_1,\ldots,A_{k-1})$, olarak ayarlandığında, yokluk hipotezi H_0 altında A için kovaryans matrisi şu şekilde verilir:

$$\sum_0 = \begin{bmatrix} \sum_{t=2}^{k}\lambda_{1t} & -\lambda_{12} & -\lambda_{13} & \ldots & -\lambda_{1,k-1} \\ -\lambda_{12} & \sum_{t\neq 2}^{k}\lambda_{2t} & -\lambda_{23} & \ldots & -\lambda_{2,k-1} \\ \vdots & \vdots & \vdots & \vdots\ \cdots & \vdots \\ -\lambda_{1,k-1} & -\lambda_{2,k-1} & -\lambda_{3,k-1} & \ldots & \sum_{t\neq k-1}^{k}\lambda_{k-1,t} \end{bmatrix}$$

$t \neq q = 1,\ldots,k$ (k deneme) için, burada $\lambda_{qt} = \lambda_{tq}$ = (hem q hem de t'nin gözlemlendiği blok sayısı).

Skillings-Mack test istatistiği eşitlik 2.3'te gösterilmektedir.

$$SM=A\Sigma_0^{-1}A' \qquad (2.3)$$

Burada Σ_0^{-1} , Σ_0'nin ters matrisidir.

H_0: $[\tau_1 = \cdots = \tau_k]$

H_1: en az iki deneme etkisi farklıdır $\tau_i \neq \tau_j$

SM değeri, α önemlilik düzeyinde elde edilecek sm_α kritik tablo değerinden büyük ise H_0 hipotezi reddedilir (7,8,11).

Yokluk hipotezi reddedilirse, gruplar arasında genel olarak bir farklılık vardır. Hangi gruplar arasında farklılık olduğunu bulmak için atama ve silme yöntemleri uygulandıktan sonra parametrik olmayan çoklu karşılaştırmalara bakılabilir. Parametrik olmayan çoklu karşılaştırmalar için farklı testler bulunmaktadır. Bu çalışmada, atama ve silme yöntemlerini uyguladıktan sonra parametrik olmayan çoklu karşılaştırmalar için en iyi bilinen testlerden biri olan Nemenyi testi kullanılmıştır. Nemenyi testi eşitlik 2.4'te gösterilmektedir:

$$|\bar{R}_i\text{-}\bar{R}_j| > \frac{q_{\infty;k;\alpha}}{\sqrt{2}} \sqrt{\frac{k(k+1)}{6n}} \qquad (2.4)$$

Bu denklemde, $\bar{R}_i\text{-}\bar{R}_j$:ortalama rank toplamlarının farkı, n: toplam örneklem büyüklüğü, k: tedavi grubu sayısı ve q: student aralık dağılımıdır. Genel yokluk hipotezi, iki tedavi etkisinin aynı olmasıdır (H_0: $\tau_i = \tau_j$). Nemenyi, deneysel hatayı hesaba katmak için bu testi geliştirmiştir. Bu nedenle herhangi bir düzeltmeye (örneğin Bonferroni düzeltmesi, Holm düzeltmesi) gerek duyulmamaktadır (12,13).

2.1 Eksik Veri Mekanizmaları

Rubin'e (4) göre, üç eksik veri mekanizması vardır. Bunlar tamamen rastgele eksik (MCAR), rastgele eksik (MAR) ve rastgele olmayan eksik (MNAR)'dır. MCAR mekanizmasında eksiklik, gözlenen ve eksik değerlere sahip olan değişkenlere bağlı değildir. MAR mekanizmasında eksiklik, değişkenin gözlenen değerlerine bağlıdır ve verilerdeki değişkenin eksik değerlerine bağlı değildir. MNAR mekanizmasında eksiklik, verideki değişkenin eksik değerlerine bağlıdır (9,14,15). Bu tür bir eksikliği tasarlamak oldukça zordur, çünkü değişkenin gözlenen değerleri eksik değerlerin yanlı

tahminlerini üretir (5). Bu nedenle bu çalışma için eksik veri mekanizmasının MCAR veya MAR olduğu varsayılmıştır.

MCAR, MAR ve MNAR Kavramları:

Rubin (4) eksik veri sorununu üç kategoriye ayırmıştır. Onun teorisinde her noktada verinin eksik olma olasılığı vardır. Bu olasılıkları idare eden sürece eksik veri mekanizması veya yanıt mekanizması denir. Eksik veri modeli, sürecin modeline verilen isimdir (6).

Verilerin tamamen rastgele eksik olduğu durum, tüm durumlar için eksik olma olasılığı aynı olduğunda oluşur. Bu, eksik verilerin nedenlerinin verilerle ilgisi olmadığı anlamına gelir. Sonuç olarak, bilginin kaybı dışında verilerin kaybolması nedeniyle ortaya çıkan karmaşıklıkların çoğu göz ardı edilebilir. MCAR kullanışlı olsa da, veriler için genellikle gerçekçi değildir (6).

Eksik olma olasılığı gözlemlenen verilerle belirtilen gruplar içinde aynıysa, veriler MAR'dır. MAR, MCAR'dan çok daha geniş bir sınıfa sahiptir ve daha gerçekçidir. MAR varsayımı, modern eksik veri yöntemlerinin başlangıcıdır. Ne MCAR ne de MAR geçerli değilse, o zaman MNAR söz konusudur. Literatürde aynı kavram için NMAR kısaltması da kullanılabilir (6).

R matrisi Y matrisinde bulunan eksik gözlemlerin yerini belirtmek üzere, R'nin dağılımı Y=($Y_{gözlenen}$, $Y_{kayıp}$)'e ya tasarım ya da rastlantısal olarak bağlıdır. Bu ilişki eksik veri modeli olarak tanımlanabilir. φ, eksik veri modelinin parametrelerini göstermektedir. Eksik veri modelinin genel ifadesi Pr (R|$Y_{gözlenen}$, $Y_{kayıp}$, φ) olarak belirtilebilir.

Eğer veri MCAR ise,

Pr (R|$Y_{gözlenen}$, $Y_{kayıp}$, φ) = Pr (R=0|φ)'dir.

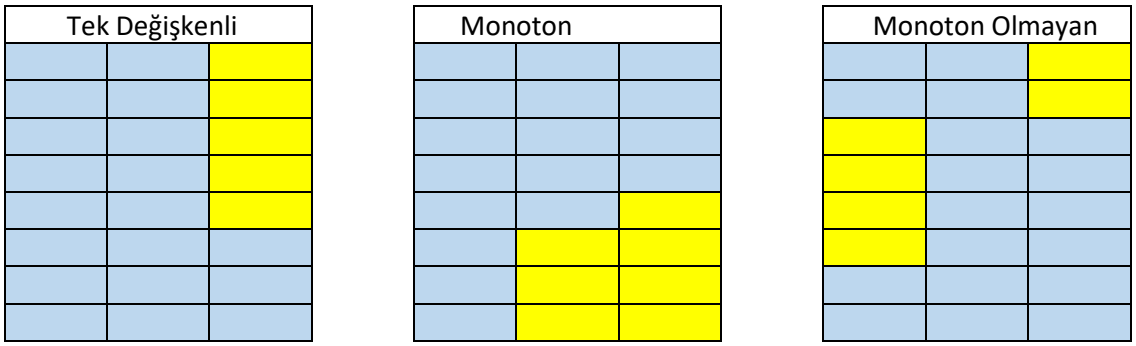
Eğer veri MAR ise,

Pr (R|$Y_{gözlenen}$, $Y_{kayıp}$, φ) = Pr (R=0|$Y_{gözlenen}$,φ)'dir.

Eğer veri MNAR ise,

Pr (R|$Y_{gözlenen}$, $Y_{kayıp}$, φ)=Pr (R=0|$Y_{gözlenen}$, $Y_{kayıp}$, φ)'dir (9).

2.2. Eksik Veri Tasarımları



Şekil 2.1. Eksik Veri Tasarımları (Mavi renkteki gözlenen, sarı renkteki eksik veri)

Eksik veri tasarımları üç kısımda incelenebilir. Tek değişkenli tasarımda yalnızca bir değişkende eksik veri vardır. Şekil 2.1'de görüldüğü gibi birinci ve ikinci sütunlarda (değişkenlerde) eksik veri yoktur. Üçüncü sütunda eksik veri gözlenmiştir. Yani tek değişkende eksik veri gözlendiği için tek değişkenli eksik veri tasarımından söz edilebilir. Şekil 2.1'de ikinci durumda ikinci değişkenin son üç gözlemi eksik olarak görülmektedir. Üçüncü değişkende ise son dört gözlem eksik olarak görülmektedir. Burada görüldüğü gibi son üç gözlem çalışmadan ayrılmış ve çalışma boyunca bir daha çalışmaya dahil olmamıştır. Böyle bir eksik veri tasarımı monotondur. Üçüncü durumda ise birinci değişkendeki dört gözlem eksiktir. Ancak bu dört gözlem sonra çalışmaya dahil olmuştur. Üçüncü değişkende de ilk iki gözlem eksik görülmektedir. Burada görüldüğü gibi monoton eksik veri tasarımındaki eksik gözlemlerin bir daha gözlenmemesi durumu yoktur. Böyle bir eksik veri tasarımı monoton olmayandır.

2.3. Uzunlamasına Veride Atama ve Silmeye Dayalı Yöntemler

2.3.1. Silmeye Dayalı Yöntemler

Eksik değerlerle başa çıkmanın en yaygın ve en kolay tekniği liste bazında silmedir (Listwise Deletion) (LD). Bu yönteme tam gözlemlerin analizi de denir. Listesel silme yöntemi, tüm değişkenlerdeki eksik olmayan gözlemlerin mevcut olduğu durumları ele alır. Bu yöntemin temel avantajı, istatistiksel analizlerin değişiklik yapılmadan uygulanabilmesi nedeniyle basitliğidir. Bu yöntemin temel iki dezavantajı vardır. Bunlardan biri kesinliğin kaybı, diğeri ise eksik veri mekanizması MCAR olmadığında oluşan yanlılıktır (9).

LD, birçok istatistik paketinde (SPSS, SAS ve Stata) eksik verileri işlemenin varsayılan yoludur. na.omit() işlevi R yazılımında aynı işlemi gerçekleştirir. Prosedür, analiz değişkenlerinde bir veya daha fazla eksik değere sahip tüm gözlemleri ortadan kaldırır (6).

Tam gözlemlerin analizinin en büyük avantajı kolaylıktır. Veriler MCAR ise, liste bazında silme, ortalamaların, varyansların ve regresyon ağırlıklarının yansız tahminlerini üretme durumu ortaya çıkarır. MCAR altında, liste bazında silme yöntemi, genellikle tüm mevcut verilere göre daha büyük olan standart hatalar ve anlam düzeyleri üretir (6).

LD'nin bir dezavantajı, potansiyel olarak verilerin israf olmasıdır. Veriler MCAR değilse, LD, ortalamaların, regresyon katsayılarının ve korelasyonların tahminlerini önemli düzeyde yanlı hale getirebilir (6).

2.3.2. Uzunlamasına Verilerde Kullanılabilen Parametrik Olmayan Atama Yöntemleri

Rastgele Sıcak Deste Atama Yöntemi

Atama (Imputation), eksik verilerle başa çıkmanın bir başka yoludur. Little ve Rubin'e (9) göre, tekli atamada iki genel yaklaşım vardır. Biri açık modellemedir çünkü varsayımlar açıktır ve tahmini dağılım kuramsal bir istatistiksel modele (örneğin, çok değişkenli normal dağılım) dayanmaktadır. Diğer yaklaşım ise varsayımlar kapalı

olduğu ve yaklaşım bir algoritmaya odaklandığı için örtük modellemedir. Örtük modelleme yöntemleri arasında sıcak deste ataması bulunur (9). Parametrik olmayan atama yöntemleri örtük modelleme yöntemleridir ve normal dağılmayan verilerde kullanılırlar.

Sıcak deste ataması (HDI), bir yanıt vermeyen (alıcı) için bir veya daha fazla değişkenin eksik değerlerini, yanıt vermeyenle benzer olan bir yanıt verenden (donör) alınan gözlemlenen değerlerle değiştirmeyi kapsar. İki önemli sıcak deste atama yöntemi türü vardır. Bir yöntem türünde, donör, donör havuzu adı verilen olası donör kümesinden rastgele seçilir; bu yöntemlere rastgele sıcak deste yöntemleri adı verilir. Diğer yöntemde, tek bir donör belirlenir ve değerler bu durumdan atanır; Bu yöntemlere deterministik sıcak deste yöntemleri denir çünkü donörün seçiminde rastgelelik yoktur. Uygulanan üç tür basit rastgele sıcak deste ataması vardır. İlk türde, her değişken (sütun) ayrı ayrı ele alınır. Bir nesnenin (satır) bir değişkende (sütun) eksik bir değeri varsa, aynı değişkendeki gözlenen değerlerden biri rastgele seçilir ve eksik değer bununla değiştirilir. Bu, tüm eksik değerler için yapılır (16). Bu çalışmada bu türü kullanılmaktadır. Wang ve diğerleri (10), rastgele sıcak deste atama yönteminin uzunlamasına veriler için kullanılabileceğini göstermiştir. Varsayımsal yönden, rastgele sıcak deste atamasının uyumluluğu MCAR altında gösterilmiştir (16).

Makine Öğrenmesine Dayalı Parametrik Olmayan Çoklu Atama Yöntemleri

Çoklu atama (Multiple Imputation), 1970'lerde Donald B. Rubin tarafından geliştirilmiş bir yöntemdir. Çoklu atama, Mevcut Nüfus Anketi'nin (CPS) Mart Gelir Ekinde eksik gelir verileriyle ilgili pratik bir soruna çözüm olarak geliştirilmiştir. 1977'de Scheuren, Sosyal Güvenlik İdaresi ve ABD Nüfus Sayımı Bürosu'nun ortak bir projesi üzerinde çalışırken Nüfus Sayımı Bürosu o zamanlar sıcak deste atama yöntemini kullanıyordu. Scheuren, varyansın düzgün bir şekilde hesaplanamayacağını belirtmesi üzerine Rubin, 1970'lerin başında daha önce keşfettiği tam veri setinin birden fazla versiyonunu kullanma fikrini ortaya attı (17). Fikri tanıtan orijinal 1977 raporu 2004 yılında American Statistician'ın tarih köşesinde yayınlandı (6,17,18).

Rubin, eksik değer için tek bir değer tahmin etmenin (tek bir atama) genel olarak doğru olamayacağını gözlemledi. Gözlemlenmeyen verileri gözlemlenen verilerle

ilişkilendirmek için bir modele ihtiyacı vardı ve belirli bir model için bile tahmin edilen değerlerin kesin olarak hesaplanamayacağını belirtti. Bu nedenle eksik verilerin belirsizliğini yansıtan birden fazla tahmin oluşturmak fikrini ortaya koydu (6).

Rubin'in önerisi, birleşik tahminleri hesaplamak için formüller içermeyip bunun yerine, varyasyon çalışmasına odaklanıp tahmin edilen değerlerdeki belirsizliğe çözüm bulmayı amaçlamıştır. Bu fikir, Bayesçi çıkarım çerçevesine dayanmaktadır (6).

Eksik veriye sahip uzunlamasına verilere yönelik iki genel çoklu atama yaklaşımı vardır. Bunlar, birleşik modelleme çok değişkenli normal atama ve tamamen koşullu belirlemedir. Birleşik modelleme, çok değişkenli normal dağılmış verileri varsaydığı için parametrik bir yaklaşımdır. Tamamen koşullu belirleme, normal dağılım varsayımına dayanmadığı için daha fazla esnekliğe sahiptir ve çoklu atama yapmak için zincirleme denklemlerle çoklu atamalar (multiple imputation by chained equations) (MICE) yöntemi kullanılır. MICE'da, çok değişkenli eksik veriler, tamamen koşullu belirleme adı verilen her bir değişkeni ayrı ayrı ele alan bir şekilde tahmin edilir (19,20). Tamamen koşullu belirleme, esasen kesitsel veriler için önerilmiştir ancak zamana bağlı değişkenlerin tekrarlanan ölçümlerini ayrı bir değişken olarak değerlendirerek eşit aralıklarla toplanan uzunlamasına verileri tahmin etmek için de kullanılabilir (21). Huque ve diğerleri (21) uzunlamasına çalışmalarda tamamen koşullu belirlemeyi kullanmıştır. Standart tamamen koşullu belirleme (MICE) uzunlamasına çalışmalarda iyi performans göstermiş ve eşit aralıklarla toplanan uzunlamasına verilerde genelleştirilmiş doğrusal karışık etki modeli (GLMM) tabanlı çoklu tahmin yaklaşımlarından daha uygun olduğu belirlenmiştir. Standart tamamen koşullu belirleme, düzensiz zaman aralıklarında toplanan uzunlamasına veriler için uygun değildir. Standart tamamen koşullu belirlemenin uzantıları olan GLMM tabanlı yaklaşımlar, düzensiz zaman aralıklarında toplanan uzunlamasına veriler için daha uygundur (21). Bu çalışmada, standart tamamen koşullu belirlemede (MICE) parametrik olmayan çoklu tahmin yöntemleri kullanılmıştır. Bu nedenle, bu çalışmadaki parametrik olmayan çoklu tahmin yöntemleri, düzensiz zaman aralıklarında toplanan uzunlamasına veriler için uygun değildir.

Birleşik modelleme

Birleşik modelleme (Joint Modeling), verilerin çok değişkenli bir dağılımla tanımlanabileceği varsayımından başlar. Göz ardı edilebilirlik varsayılarak, atamalar, seçilmiş bir dağılımından çekilerek oluşturulur. Model, genel olarak çok değişkenli normal dağılıma dayanarak uygulanır (6).

Sürekli verilerde ortak modelleme ile eksik verilerin atanması çok değişkenli normal dağılım varsayımına dayanır (Y~ N(μ,Σ)). Bu atama modellerinin φ parametreleri θ = N(μ,Σ))'nin fonksiyonlarıdır. Tarama operatörü sonuç değişkenlerini tahmin edici değişkenlere çevirerek θ'yı φ'ye dönüştürür. Ters tarama operatörü ters işleme izin verir. θ parametreleri genellikle bilinmezler. Monoton olmayan eksik veri tasarımında θ'yı tahmin etmek çok zordur (6).

Algoritma: Çok değişkenli normal dağılıma uyan verilerde birleşik bir modelle eksik verilerin atanması için algoritmanın adımları aşağıda verilmiştir.

1. Y'nin satırları S eksik veri tasarımına göre sınıflandırılır ($Y_{[s]}$, s =1,…,S.)

2. Uygun bir başlama değeri seçerek θ^0= N(μ^0, Σ^0) işlemine başlanır.

3. t=1,…,T'ye kadar süreç tekrar edilir.

4. s=1,…,S'ye kadar süreç tekrar edilir.

5. θ^{t-1}'in dışında s deseninin tahmin edicilerini süpürerek, $\phi_{[s]}$=SWP(θ^{t-1},s) parametreleri hesaplanır.

6. s desenindeki eksik veri sayısı olarak p_s'i hesaplanır. Daha sonra o_s = p - p_s'yi hesaplanır.

7. s desenindeki eksik veriye karşılık gelen $\phi_{[s]}$'nin p_s x p_s altmatrisinin C_8 Choleski ayrışması hesaplanır.

8. p_s uzunluğunda z~ N(0,1)) raslantı vektörü düzenlenir.

9. Regresyon ağırlıklarının ϕ_s'nin o_s x p_s altmatrisleri olarak $\hat{B}_s$'i alınır.

10. $\hat{Y}_{[s]}^{t}$'nin s desenindeki gözlemlenen veri olarak, $\hat{Y}_{[s]}^{t} = Y_{[s]}^{obs}\hat{B}_s + C_8'z$ atamaları hesaplanır.

11. s kadar tekrar edilir.

12. Schafer'e göre normal ters Wishart dağılımdan $\hat{\theta}^{t} = (\dot{\mu}, \dot{\Sigma})$'i hesaplanır.

13. t kadar tekrar edilir (6).

Tamamen koşullu belirleme (Fully Conditional Specification)

Tamamen koşullu belirleme (FCS) (19,22) çok değişkenli eksik verileri değişken bazında atamayı amaçlar. Yöntem, her eksik değişken için bir atama modelinin belirlenmesini gerektirir ve yinelemeli bir şekilde değişken başına atamalar oluşturur (6).

Birleşik modellemenin aksine FCS, çok değişkenli dağılımı P(Y, X,R|θ), P(Y_j|X,Y_{-j},R, ϕ_j) koşullu yoğunluk kümesi aracılığıyla belirler. Bu koşullu yoğunluk, X, Y_{-j} ve R verildiğinde Y_j'yi atamak için kullanılır. Marjinal dağılımdan basit rastgele çekimlerle başlayarak, FCS kapsamında atama, koşullu olarak belirtilen atama modelleri üzerinde yineleme yapılarak yapılır. FCS, tek değişkenli atamanın doğal bir genellemesidir (6).

Rubin (23) atamaları oluşturmak için gereken işi üç göreve bölmüştür. Modelleme görevi, veriler için belirli bir model seçer. Hesaplama görevi, model verildiğinde arka parametre dağılımını formüle eder ve atama görevi, parametre ve veri dağılımlarından ardışık olarak çekerek eksik veriler için rastgele çekimler yapar. FCS, çekimlerin yapılacağı koşullu dağılımları doğrudan belirtir ve bu nedenle veriler için çok değişkenli bir model belirtme ihtiyacını ortadan kaldırır (6,24).

Zincirli Denklemlerle Çoklu Atama (MICE) algoritması

Koşullu olarak belirtilen modeller altında hesaplamayı uygulamanın birkaç yolu vardır. MICE algoritması gözlemlenen verilerden rastgele bir çekilişle başlar ve eksik verileri her bir değişkeni ayrı ele alacak şekilde hesaplar. Bir yineleme, tüm Y_j

boyunca bir döngüden oluşur. Yineleme sayısı T genellikle düşük olabilir, örneğin 5 veya 10 (6).

MICE algoritması, aşağıda verilen algoritmayı m kez paralel olarak çalıştırarak birden fazla hesaplama üretir. MICE algoritması, durum alanının tüm atanan değerlerin toplanması olduğu bir Markov zinciri Monte Carlo (MCMC) yöntemidir. MICE algoritması, ortak dağılımdan örnekler elde etmek için koşullu dağılımlardan örnek alan bir Bayes benzetim tekniği olan bir Gibbs örnekleyicisidir. Gibbs örnekleyicisinin geleneksel uygulamalarında, tamamen koşullu dağılımlar ortak olasılık dağılımından türetilir (6,25).

Algoritma : Çok değişkenli eksik verilerin atanması için MICE algoritması.

1. j=1,…,p ile Y_j değişkeni için P($Y_j^{kayıp}|Y_j^{gözlenen}$, Y_{-j}, R) atama modeli belirlenir.

2. Her bir j için, $Y_j^{gözlenen}$'den rastgele örneklem çekilerek $\dot{Y}_j^0$ atamaları yapılır.

3. t=1,,…,T için tekrar yapılır.

4. j=1,,…,p için tekrar yapılır.

5. Y_j dışındaki mevcut tam veri olarak $Y_{-j}^{\dot{t}}$=($Y_1^{\dot{t}}$,… , $Y_{j-1}^{\dot{t}}$ $Y_{j-1}^{\dot{t}-1}$, $Y_{j+1}^{\dot{t}-1}$, $Y_p^{\dot{t}-1}$)

6.$\dot{\phi}_j^t$~ P($\phi_j^t|Y_j^{gözlenen}$, $\dot{Y}_{-j}^t$,R)'i çekilir.

7. $\dot{Y}_j^t$~ P($Y_j^{kayıp}|Y_j^{gözlenen}$, $\dot{Y}_{-j}^t$,R, $\dot{\phi}_j^t$) atamaları yapılır.

8. j kadar tekrar edilir.

9. t kadar tekrar edilir (6).

Zincirli Denklemlerle Çoklu Atama (MICE)

Bir dizi değişkeni, $y_1, \ldots, y_n$, bazılarının veya hepsinin eksik değerleri olduğunu düşünelim. Bu verilerde MICE kullanarak atama yapmak üç ana adımdan oluşur: çoklu atama oluşturma, atanan verileri analiz etme ve analiz sonuçlarını bir araya getirme

(22). Ana fikir, her eksik değişkeni kendi atama modelini kullanarak atamaktır. Tüm eksik değerler başlangıçta rastgele doldurulur. En az bir eksik değeri olan ilk değişken, diyelim ki y_1, daha sonra kalan değişkenler, $y_2, \ldots, y_j$ üzerinde regresyona tabi tutulur. Bu, y_1 için gözlemlenen değerlere sahip bireylerle sınırlıdır. y_1'deki eksik değerler artık y_1'in arka tahmini dağılımından simüle edilmiş çekimlerle değiştirilir. Eksik değerleri olan bir sonraki değişken, diyelim ki y_2, daha sonra diğer tüm değişkenler, $y_1, y_3, \ldots, y_j$ üzerinde regresyona tabi tutulur. Bu tahmin, gözlemlenen y_2'ye sahip bireylerle sınırlıdır ve y_1'in tahmini değerlerini kullanır. Yine, y_2'deki eksik değerler, y_2'nin arka tahmini dağılımından yapılan çekimlerle değiştirilir. Bu işlem, sırayla eksik değerlere sahip diğer tüm değişkenler için tekrarlanır. Sonuçları sabitlemek için bu döngü birkaç kez yinelenir ve bir tahmini veri kümesi üretilir. Tüm prosedür m kez tekrarlanır ve m tahmini veri kümesi üretilir. Her bir tam veri kümesi, MICE tarafından ayrı ayrı analiz edilir ve ardından sonuçlar bir araya getirilir. MICE'de iki temel parametrik olmayan çoklu tahmin yöntemi vardır. Bunlar MICE-sınıflandırma ve regresyon ağaçları ve MICE-rastgele ormandır. İki yöntem de değişkenler arasında korelasyon mevcut olduğunda yinelemeli bölümlemeye dayanır (20).

MICE-sınıflandırma ve regresyon ağaçları (MICE-CART): Bu yöntemde yanıt değişkeni kategorik veya sürekli olabilir (20,26). MICE-CART, MICE-rastgele ormanının aksine yalnızca bir ağaç üretir. Kök düğüm, tüm üyeleri içeren ağacın tepesindedir. En iyi değişkeni bulmak için verileri arayarak başarılı olur ve bireyleri en iyi şekilde ayıran bir kesme ile iki alt düğüm oluşturulur. Optimal bir bölme, Gini indeksi gibi bir homojenlik ölçüsüne göre alt gruplar oluşturur. Belirli bir durdurma kriterine ulaşıldığında, her alt düğümün bölümlenmesi tamamlanır (27). Düğümleri bölmek için dahil etme kriteri Gini indeksidir. Aşırı uyumu önlemek için kullanılan düzenleme tekniği çapraz doğrulamadır. Karmaşıklık parametresine göre uyumu iyileştirmeyen herhangi bir bölme muhtemelen çapraz doğrulama ile budanacaktır. Karmaşıklık parametresinin temel rolü, açıkça değersiz olan bölmeleri budayarak hesaplama süresinden tasarruf etmektir (20,28). MICE-CART'ta, minimum buket, herhangi bir yaprak düğümündeki minimum gözlem sayısıdır. Minimum buketi 1 gibi çok küçük bir değere ayarlanırsa modelde aşırı uyumluluk riski ortaya çıkabilir. Bu

nedenle, minimum buket 5 olarak belirlendi. Bir bölünmenin denenmesi için bir düğümde bulunması gereken minimum gözlem sayısı 15'tir. Son ağaçtaki herhangi bir düğümün maksimum derinliği 30'dur. Karmaşıklık parametresi 0,0001'dir. k katlı çapraz doğrulama sayısı 10'dur (20,28).

Algoritma: MICE'de sınıflandırma ve regresyon ağaç yönteminin uygulanması.

Y diye adlandırılan bir veri matrisi, Y_j kısmi olarak gözlenen değişkenlerin j'inci sütunu (modelin mümkün olduğunca çok bilgiyi taşıması için eksik değerler artan sırada dizilir), p eksik değere sahip olan değişken sayısı, $Y_J^{gözlenen}$ ve $Y_J^{kayıp}$ sırasıyla j'inci sütundaki gözlenen ve eksik veridir. $\dot{Y}$ ise Y'nin sonradan atama ile tamamlanmış veri matrisidir.

1. j=1,…,p olmak üzere, $Y_J^{gözlenen}$'den rastgele çekim yaparak $\dot{Y}_J^0$ atamaları yapılır ve $\dot{Y}$ adında bir veri matrisi tanımlanır

2. j=1,…,p olmak üzere $\dot{Y}_J^0$'i izleyen adımlarla değiştirilir ve bir atama yapılmış veri seti oluşturulur.

a) CART yöntemi kullanılarak $Y_J^{gözlenen}$'nin gözlemleriyle sınırlı olmak üzere $\dot{Y}$ üzerinde bir ağaç oluşturulur. Bu, bir ağaçta birçok yaprağa sebep olur (her biri $Y_J^{gözlenen}$'nin alt kümesidir ve bunlara donör adı verilir) .

b) $Y_J^{kayıp}$'nin üyeleri içinde adım 2a'daki oluşturulan ağaca göre sonuçlanacak yaprak belirlenir.

c) $Y_J^{kayıp}$'nin üyeleri içinde adım 2b'deki yaprağın donörlerinden rastgele $Y^{gözlenen}$'den bir değer seçilir. $\dot{Y}_J^0$'nin eksik değerlerini bu atama değerleriyle değiştirilir ve j'yi artırmadan önce $\dot{Y}_J$'nin tam versiyonu $\dot{Y}$'e eklenir.

3. t tekrar sayısını göstermek üzere 2'inci adım t kadar tekrar edilir.

4. m adet atama kümesi oluşturmak üzere adım1-3'ü m adet tekrar edilir (20).

MICE-rastgele orman (MICE-RF): Bu yöntem yalnızca bir ağaç yerine çok sayıda ağaç üretir. Birçok ağaç üretip ortalamasını almak, tutarsız ağaçların varyansını ve yaygınlığını azaltır (20,26). Bireysel ağaçlarda çeşitlilik sağlanır; bu daha sağlam bir çözümle sonuçlanır ve bu yöntem daha doğru hale gelir (20). Önyükleme (bootstrapping) ve rastgele girdi seçimi, bu çeşitlilik ile birleştirilebilen prosedürlerdir (20,29). Her bir ağacı büyütmeden önce, önyükleme ile veri kümesinin üyelerinden değiştirmeli rastgele bir seçim yapılır. Rastgele girdi seçimi ile en iyi bölünmeyi bulmak için küçük bir girdi değişkeni grubu seçilir (20). Bireysel ağaçların sonuçları ortalama alınarak toplanır. Shah ve diğ. (30), yanıt değişkeni sürekli olduğunda MICE-RF'nin kullanılabileceğini göstermiştir. Shah ve diğ. (30) tarafından yapılan benzetim çalışmasında 10 ve 100 ağaç için tahmin kalitesinin aynı olduğu kanıtlanmıştır. Bu nedenle, bu çalışmadaki ağaç sayısı 10 olarak belirlenmiştir. Çoklu atama yöntemleri, MAR altında yansız ve geçerli tahminler sağlar (31).

Algoritma: MICE'de rastgele orman yönteminin uygulanması.

Y diye adlandırılan bir veri matrisi, Y_j kısmi olarak gözlenen değişkenlerin j'inci sütunu (Modelin mümkün olduğunca çok bilgiyi taşıması için eksik değerler artan sırada dizilir), p eksik değere sahip olan değişken sayısı, $Y_J^{gözlenen}$ ve $Y_J^{kayıp}$ sırasıyla j'inci sütundaki gözlenen ve eksik veridir. $\dot{Y}$ ise Y'nin sonradan atama ile tamamlanmış veri matrisidir.

1. j=1,…,p olmak üzere, $Y_J^{gözlenen}$'den rastgele çekim yaparak $\dot{Y}_J^0$ atamaları yapılır ve $\dot{Y}$ adında bir veri matrisi tanımlanır.

2. j=1,…,p olmak üzere $\dot{Y}_J^0$'i izleyen adımlarla değiştirilir ve bir atama yapılmış veri seti oluşturulur.

a) $Y_J^{gözlenen}$'nin gözlemleriyle sınırlı olmak üzere $\dot{Y}$ üzerinden k bootstrap örneklem çekilir.

b) Adım 2a'daki çekilen her bir bootstrap örneklemden bir ağaç oluşturulur. Bu, k ağaçla sonuçlanır ve her bir ağaç birçok yaprağa sahiptir. Her bir yaprak $Y_J^{gözlenen}$ 'nin bir alt kümesi olan ve donör olarak adlandırılan veriye sahiptir.

c) $Y_J^{kayıp}$'nin üyeleri içinde adım 2b'deki oluşturulan k adet ağaca göre sonuçlanacak yaprak belirlenir. Bu, $Y_J^{kayıp}$'nin her bir üyesinin donörleri ile k adet yaprak ile sonuçlanır.

d) $Y_J^{kayıp}$'nin üyeleri içinde adım 2c'deki yaprağın donörlerinden rastgele $Y^{gözlenen}$'den bir değer seçilir. $\dot{Y}_J^0$'nin eksik değerlerini bu atama değerleriyle değiştirilir ve j'yi artırmadan önce $\dot{Y}_J$'nin tam versiyonu $\dot{Y}$'e eklenir.

3. t tekrar sayısını göstermek üzere 2'inci adımı t kadar tekrar edilir.

4. m adet atama kümesi oluşturmak üzere adım1-3'ü m adet tekrar edilir (20).

3. GEREÇ VE YÖNTEM

3.1. Üç parametrik olmayan atama yöntemini ve liste bazında silme yöntemini parametrik olmayan çoklu karşılaştırmalara uygulayarak yeni bir yaklaşım:

Eksik veri içeren k-bağımlı örneklemlerin (uzunlamasına veri) parametrik olmayan çoklu karşılaştırmaları için üç parametrik olmayan atama yöntemini ve liste bazında silme yöntemini uygulamak için izlenecek adımlar:

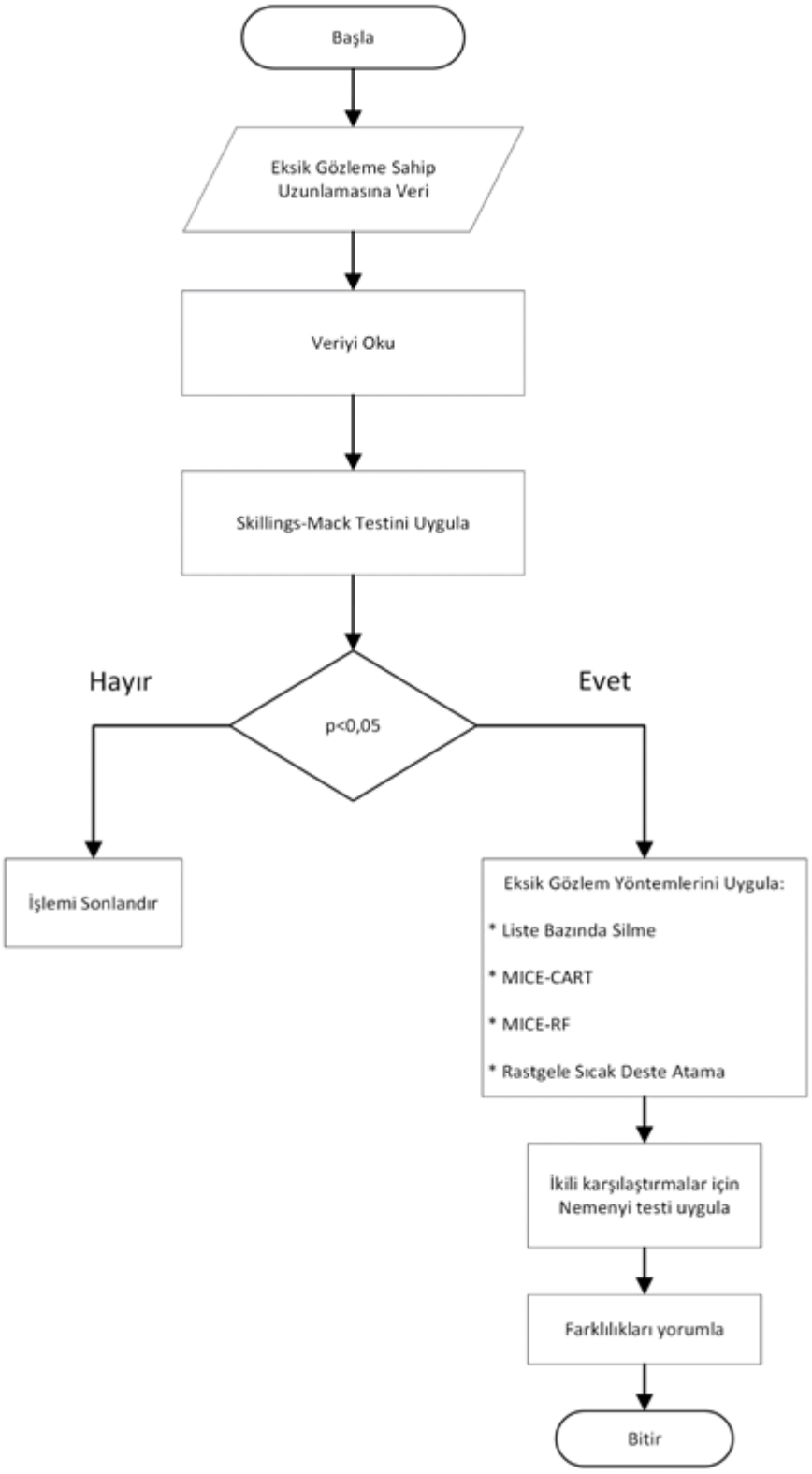
1. Adım: Normal dağılım göstermeyen, eksik veri içeren k-bağımlı örnekleme sahip veri olduğunda, önce Skillings-Mack test istatistiği ve ilgili p değeri hesaplanır. Anlamlı bir farklılık bulunursa, çoklu karşılaştırmaları incelemek için bir sonraki adıma geçilir.

2. Adım: Zamana bağımlı değişkenlerin tekrarlanan ölçümlerini ayrı değişken olarak değerlendirerek eşit aralıklarla toplanan uzunlamasına verileri tahmin etmek için kullanılabilen üç parametrik olmayan atama yöntemi ve eksik gözlemleri atan liste bazında silme yöntemi uygulanır ve tamamlanmış gözlemlerle işleme devam edilir.

3. Adım: Parametrik olmayan çoklu karşılaştırmaları incelemek için Nemenyi testi uygulanır.

4. Adım: İkili karşılaştırmalar ile farklılıklar değerlendirilir.

Önerilen yaklaşımın aşamaları Şekil 3.1’de gösterilmektedir.



Şekil 3.1. Önerilen Yaklaşımın Aşamaları

3.2. Benzetim Çalışması

Ortalama vektörü ve varyans-kovaryans matrisi kullanılarak çok değişkenli normal dağılımdan farklı senaryolar oluşturulmuş ve eksik gözlemlere sahip k-bağımlı örneklemlerin çoklu karşılaştırmaları için dört yöntemin performansı belirlenmeye çalışılmıştır. Küçük örneklemlerle çok değişkenli normal dağılım kullanılmıştır çünkü çok değişkenli normallik varsayımı küçük örneklerle zayıf olabilir ve bu varsayımın kabul edilebilirliğini küçük örneklemlerle test etme yeteneği oldukça zordur (14,32).

$$\mu=\begin{pmatrix}\mu_1\\ \mu_2\\ \mu_3\end{pmatrix}, \sum=\begin{pmatrix}{\sigma_1}^2 & \sigma_1\sigma_2\rho_{12} & \sigma_1\sigma_3\rho_{13}\\ \sigma_1\sigma_2\rho_{12} & {\sigma_2}^2 & \sigma_2\sigma_3\rho_{23}\\ \sigma_1\sigma_3\rho_{13} & \sigma_2\sigma_3\rho_{23} & {\sigma_3}^2\end{pmatrix}$$

μ_1, μ_2, μ_3 ortalamalar ${\sigma_1}^2, {\sigma_2}^2, {\sigma_3}^2$ varyanslar, ρ_{12}, ρ_{13}, ρ_{23} korelasyon katsayılarıdır. μ_1=μ_2=μ_3=0 için tip 1 hata ve μ_1=1,3, μ_2=0, μ_3=0 için güç değerleri bulunmaya çalışılmıştır. ρ_{12}=ρ_{13}=ρ_{23}=0,2 için zayıf korelasyon, ρ_{12}=ρ_{13}=ρ_{23}=0,5 için orta büyüklükte korelasyon, ρ_{12}=ρ_{13}=ρ_{23}=0,8 için güçlü korelasyon ve ρ_{12}=ρ_{13}=ρ_{23}=-0,2 negatif zayıf korelasyon olarak belirlenmiştir. ${\sigma_1}^2 = {\sigma_2}^2 = {\sigma_3}^2$ =1 ile eşit varyans belirlenmiştir. Bütün senaryolarda Friedman testi p değerleri 0,05'ten küçüktür ve anlamlıdır. Örneğin, n = 10, μ_1=1,3, μ_2=0, μ_3=0 ve ρ_{12}=ρ_{13}=ρ_{23}=-0,2 için üç ölçüm grubu arasında anlamlı fark vardır (Friedman testi p değeri= 0,04).

MCAR ve MAR altında, verilerin birinci ve ikinci tekrarlı ölçüm grupları silinmiştir. Üçüncü tekrarlı ölçüm grubunda eksik değer yoktur çünkü birinci ve ikinci gruptaki eksik olma olasılığı üçüncü gruptaki aynı satırdaki değerin sıralamasıyla orantılıdır (MAR varsayımı). MCAR altında, verilerin %10, %20 ve %30'u silinmiştir. MAR altında, üçüncü gruba göre verilerin %10, %20 ve %30'u silinmiştir. Monoton olmayan eksik veri tasarımı varsayılmıştır. Daha sonra, çoklu karşılaştırmaları incelemek için dört atama yöntemi uygulanmıştır. Her senaryo için, yöntemlerin 0,05 anlamlılık düzeyinde tip 1 hata oranını ve gücünü değerlendirmek için 1000 tekrar yapılmıştır. Yokluk hipotezi, iki deneme etkisinin aynı olduğu şeklindedir (H_0: $\tau_i = \tau_j$). Sıra numarası ortalamaları farkını ($\bar{R}_i$-$\bar{R}_j$) ve p değerlerini kritik anlamlılık düzeyine göre tahmin etmek amaçlanmıştır. Tip 1 hatayı ve gücü değerlendirirken birinci ve ikinci tekrarlarla ilgilenilmiştir. Örneklem büyüklükleri 10 ve 20 olarak belirlenmiştir. Veri

üretimi ve analizler için çalışmada dört R (4.3.2 versiyon) (13,33,34,35,36) paketi kullanılmıştır. Çok değişkenli normal dağılım oluşturmak için MASS paketi ve parametrik olmayan çoklu karşılaştırmalar için PMCMRplus paketi kullanılmıştır. MCAR ve MAR altında verileri silmek için missMethods paketi kullanılmıştır. MICE-CART ve MICE-RF için mice paketi ve rastgele sıcak deste ataması için missMethods paketi kullanılmıştır. Tezdeki benzetim çalışmasına ilişkin kodlar https://github.com/tuncyanar20008/thesis_codes adresinde verilmiştir.

4. BULGULAR

4.1. Benzetim Sonuçları

Tablo 4.1 ve 4.2, MCAR ve MAR altındaki dört yöntemin tip 1 hata olasılıklarını göstermektedir. Örneklem büyüklüğü küçük olduğunda, tüm yöntemler tip 1 hata olasılıklarını nominal değer olan 0,05'in altındaki değerlerle kontrol etmektedir. Örneklem büyüklüğü orta ve eksik yüzdesi 10 olduğunda, tüm yöntemlerin tip 1 hata olasılıkları nominal değer olan 0,05'in altındadır. Örneklem büyüklüğü orta ve eksik yüzdesi 20 olduğunda, MICE-CART, MICE-RF ve rastgele sıcak deste ataması bazı senaryolar için %5 ile %6 arasında tip 1 hata olasılıklarına sahiptir. Senaryoların çoğunda uygun tip 1 hata olasılıkları vardır. Ancak örneklem büyüklüğü orta ve eksik yüzdesi 30 olduğunda, MICE-CART, MICE-RF ve rastgele sıcak deste ataması senaryoların çoğu için şişirilmiş tip 1 hatalar (yüzde 5-9'a kadar) gözlenmiştir.

Tablo 4.1. MCAR altında dört yöntemin tip 1 hata olasılıkları. (α=0,05)

Parametreler			Yöntemler			
ρ	n	EY				
Küçük Örneklem Büyüklüğü			LD	HDI	MICE-CART	MICE-RF
-0,2	10	10	0,016	0,024	0,024	0,023
-0,2	10	20	0,011	0,037	0,043	0,037
0,2	10	10	0,022	0,022	0,028	0,030
0,2	10	20	0,016	0,041	0,038	0,041
0,5	10	10	0,013	0,014	0,015	0,025
0,5	10	20	0,015	0,029	0,035	0,043
0,8	10	10	0,019	0,020	0,019	0,025
0,8	10	20	0,014	0,025	0,025	0,039
Orta Örneklem Büyüklüğü						
-0,2	20	10	0,014	0,034	0,033	0,032
-0,2	20	20	0,019	0,042	0,038	0,049
-0,2	20	30	0,011	0,063	0,056	0,055
0,2	20	10	0,014	0,035	0,041	0,039
0,2	20	20	0,017	0,044	0,056	0,057
0,2	20	30	0,013	0,047	0,056	0,071
0,5	20	10	0,009	0,024	0,028	0,030
0,5	20	20	0,021	0,039	0,049	0,053
0,5	20	30	0,019	0,054	0,052	0,088
0,8	20	10	0,010	0,015	0,017	0,024
0,8	20	20	0,025	0,039	0,052	0,058
0,8	20	30	0,019	0,043	0,061	0,079

EY: Eksik Yüzdesi

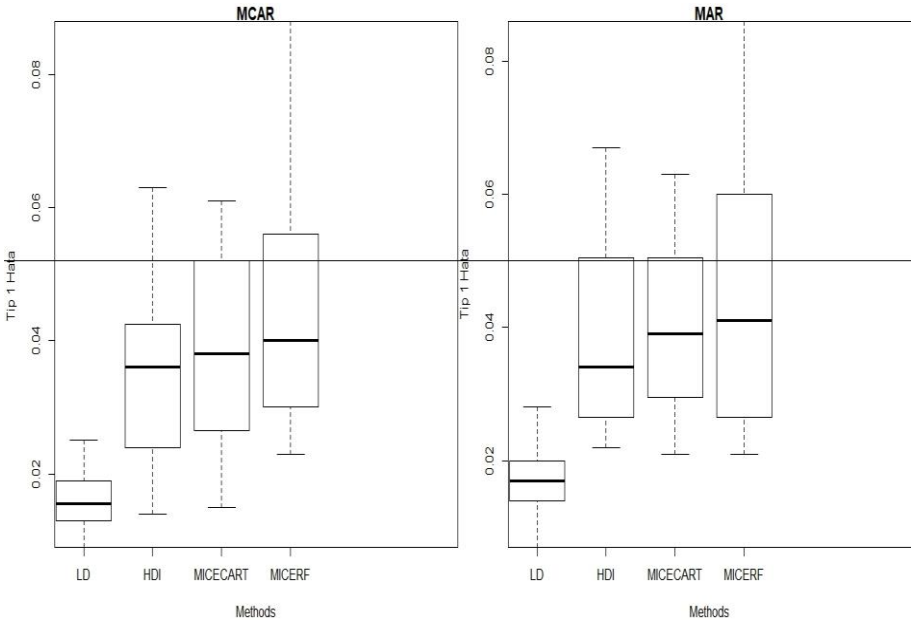
Liste bazında silme yöntemi, örneklem büyüklüğü orta ve eksik yüzdesi 30 olduğunda uygun tip 1 hata olasılıklarına sahiptir. Eksik yüzdesi arttıkça tip 1 hata

olasılıkları artmaktadır. Sonuç olarak dört yöntem aşırı şişirilmiş tip 1 hata olasılıklarına sahip değildir. Bu durum, Tablo 4.1 ve 4.2 ile Şekil 4.1.'de görülebilir.

Tablo 4.2. MAR altında dört yöntemin tip 1 hata olasılıkları. (α=0,05)

Parametreler			Yöntemler			
ρ	n	EY				
Küçük Örneklem Büyüklüğü			LD	HDI	MICE-CART	MICE-RF
-0,2	10	10	0,012	0,025	0,023	0,021
-0,2	10	20	0,014	0,037	0,038	0,046
0,2	10	10	0,016	0,025	0,023	0,027
0,2	10	20	0,019	0,030	0,042	0,044
0,5	10	10	0,021	0,032	0,029	0,032
0,5	10	20	0,014	0,030	0,034	0,033
0,8	10	10	0,007	0,026	0,030	0,026
0,8	10	20	0,016	0,027	0,032	0,031
Orta Örneklem Büyüklüğü						
-0,2	20	10	0,017	0,035	0,040	0,038
-0,2	20	20	0,020	0,052	0,046	0,049
-0,2	20	30	0,017	0,067	0,063	0,084
0,2	20	10	0,016	0,033	0,028	0,024
0,2	20	20	0,017	0,044	0,047	0,056
0,2	20	30	0,027	0,051	0,057	0,078
0,5	20	10	0,013	0,022	0,034	0,024
0,5	20	20	0,020	0,040	0,063	0,059
0,5	20	30	0,028	0,067	0,059	0,086
0,8	20	10	0,014	0,022	0,021	0,022
0,8	20	20	0,027	0,052	0,054	0,061
0,8	20	30	0,018	0,050	0,047	0,080

EY: Eksik Yüzdesi



Şekil 4.1. MCAR ve MAR altında dört yöntemin tip 1 hata olasılıkları.

Tablo 4.3 ve 4.4 ile Şekil 4.2., dört yöntemin μ_1=1,3, μ_2=0, μ_3=0 için MCAR ve MAR altındaki güç karşılaştırmasını göstermektedir. MICE-CART ve MICE-RF, MAR altında MCAR'a göre daha yüksek güce sahiptir. Çünkü MICE-CART ve MICE-RF, MAR altında daha yansız sonuçlara sahiptir. Liste bazında silme yöntemi, MCAR altında daha yüksek güce sahiptir. Çünkü liste bazında silme yöntemi, MCAR altında yansız sonuçlara sahiptir. Rastgele sıcak deste ataması, MCAR altında daha yüksek güç üretmiştir. Çünkü rastgele sıcak deste ataması yöntemi, MCAR altında yansız sonuçlara sahiptir. Liste bazında silme yöntemi diğer yöntemlere göre daha düşük güç değerlerine sahiptir ve eksik yüzdesi arttıkça liste bazında silme yönteminin gücü önemli ölçüde azalmaktadır. Bu nedenle, çok sayıda eksik gözlem olduğunda liste bazında silme yöntemi kullanılmamalıdır. Diğer yöntemler artan eksik yüzdesinden daha az etkilenmektedir. Küçük örneklem büyüklüğü ve zayıf negatif ve güçlü pozitif korelasyonlar için, MICE-CART, MICE-RF ve rastgele sıcak deste ataması, liste bazında silme yönteminden daha yüksek güç üretmiştir. MICE-CART ve MICE-RF, küçük örneklem büyüklüğü ve zayıf ve orta pozitif korelasyonlar için daha yüksek güç üretmiştir. Orta örneklem büyüklüğü ve zayıf negatif ve orta pozitif korelasyonlar için, MICE-CART, MICE-RF ve rastgele sıcak deste ataması, liste bazında silme yönteminden daha yüksek güce sahiptir. Rastgele sıcak deste ataması, pozitif korelasyondan çok negatif korelasyona karşı daha hassastır. Orta örneklem büyüklüğü ve zayıf pozitif korelasyon için, MICE-CART ve MICE-RF daha yüksek güç üretmiştir. Tüm yöntemlerin güçleri orta düzeyde örneklem büyüklüğü ve güçlü pozitif korelasyon için birbirine yakındır. Genel olarak, MICE CART ve MICE-RF, MAR altında diğer yöntemlerden üstündür. MICE-RF, MCAR altında diğer yöntemlerden üstündür. MICE-CART ve rastgele sıcak deste ataması MCAR altında benzer güce sahiptir. Şekil 4.2., makine öğrenimi tabanlı çoklu atama yöntemlerinin rastgele sıcak deste ataması ve liste bazında silme yöntemlerine karşı gücünü göstermektedir.

Tablo 4.3. MCAR altında dört yöntemin μ_1=1,3, μ_2=0, μ_3=0 için güç karşılaştırılması (α=0,05)

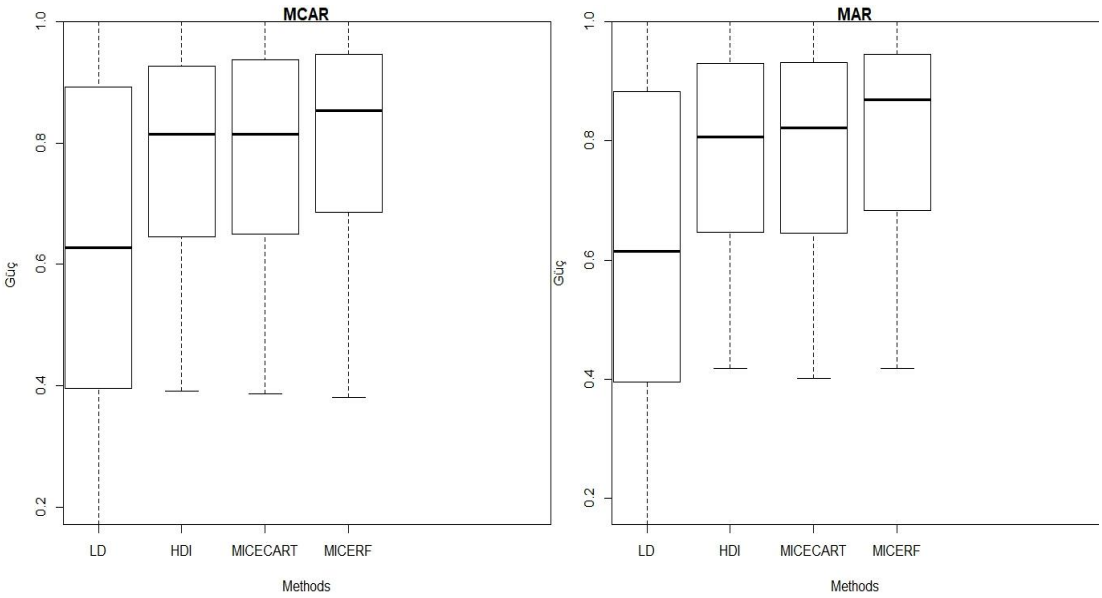
			MCAR			
ρ	n	EY	Yöntemler			
Küçük Örneklem Büyüklüğü			LD	HDI	MICE-CART	MICE-RF
-0,2	10	10	0,279	0,394	0,387	0,381
-0,2	10	20	0,172	0,391	0,395	0,402
0,2	10	10	0,392	0,517	0,522	0,532
0,2	10	20	0,272	0,518	0,532	0,522
0,5	10	10	0,552	0,669	0,673	0,694
0,5	10	20	0,400	0,621	0,626	0,678
0,8	10	10	0,894	0,888	0,886	0,923
0,8	10	20	0,660	0,814	0,812	0,875
Orta Örneklem Büyüklüğü						
-0,2	20	10	0,595	0,745	0,748	0,741
-0,2	20	20	0,516	0,758	0,747	0,754
-0,2	20	30	0,357	0,745	0,720	0,713
0,2	20	10	0,800	0,888	0,892	0,889
0,2	20	20	0,718	0,872	0,872	0,882
0,2	20	30	0,540	0,814	0,815	0,832
0,5	20	10	0,945	0,982	0,978	0,984
0,5	20	20	0,889	0,938	0,961	0,955
0,5	20	30	0,748	0,915	0,914	0,937
0,8	20	10	1,000	1,000	1,000	1,000
0,8	20	20	0,998	0,997	1,000	0,997
0,8	20	30	0,957	0,983	0,980	0,997

EY: Eksik Yüzdesi

Tablo 4.4. MAR altında dört yöntemin μ_1=1,3, μ_2=0, μ_3=0 için güç karşılaştırılması (α=0,05)

			MAR			
ρ	n	EY	Yöntemler			
Küçük Örneklem Büyüklüğü			LD	HDI	MICE-CART	MICE-RF
-0,2	10	10	0,278	0,425	0,418	0,419
-0,2	10	20	0,157	0,419	0,401	0,424
0,2	10	10	0,385	0,525	0,512	0,529
0,2	10	20	0,268	0,532	0,541	0,543
0,5	10	10	0,536	0,664	0,670	0,691
0,5	10	20	0,405	0,630	0,619	0,674
0,8	10	10	0,884	0,905	0,903	0,931
0,8	10	20	0,640	0,783	0,800	0,893
Orta Örneklem Büyüklüğü						
-0,2	20	10	0,590	0,737	0,725	0,736
-0,2	20	20	0,473	0,746	0,728	0,757
-0,2	20	30	0,353	0,768	0,745	0,727
0,2	20	10	0,791	0,877	0,873	0,884
0,2	20	20	0,688	0,867	0,880	0,881
0,2	20	30	0,495	0,829	0,842	0,856
0,5	20	10	0,930	0,965	0,982	0,975
0,5	20	20	0,881	0,951	0,960	0,960
0,5	20	30	0,718	0,908	0,892	0,913
0,8	20	10	1,000	0,999	1,000	1,000
0,8	20	20	0,993	0,997	1,000	0,999
0,8	20	30	0,966	0,966	0,975	0,992

EY: Eksik Yüzdesi



Şekil 4.2. MCAR ve MAR altında dört yöntemin güç karşılaştırması.

4.2. Gerçek Veri Örneği Sonuçları

Bu çalışmada, (bkz. [37] ve [38]) Hacettepe Üniversitesi Diş Hekimliği Fakültesi'ne başvuran kronik periodontitisli hastaların verileri kullanılmıştır. Bu hastaların periodontal sondalamalarına, ölçekleme ve kök düzeltme işleminin ardından bir hafta arayla iki kez %15 metronidazol benzoat içeren %1'lik kitosan jeli uygulanmıştır. Her iki seansta da tedavi edilen bölge periodontal macunla kapatılmıştır. İkinci jel uygulamasından bir hafta sonra macun çıkarılmış ve gerekli alanlar temizlenip cilalanmıştır. Kronik periodontitisli hastalarda mekanik tedaviye ek olarak uygulanan taşıyıcısı metronidazol olan kitosan jelinin sondalama derinliğine etkisi 6 haftalık süreler boyunca karşılaştırılmıştır. 10 kişinin üç zaman dilimindeki sondalama derinliği değerleri bu çalışmada ele alınmıştır. Üç grup arasında sondalama derinliği değerlerinde anlamlı bir fark olmadığını belirten hipotez yokluk hipotezidir. Friedman testi değeri 15.211 ve p değeri<0,001'dir. MCAR ve MAR altında tam verinin %10 ve %20'sini silindikten sonra (tedavi öncesi ve altıncı hafta gruplarında verinin %10 ve %20'si rastgele silinmiştir; onikinci hafta grubunda eksik değer yoktur), Skillings-Mack testi p değerleri kaydedilmiştir. Skillings-Mack testi p değerleri 0,05'ten küçük olduğundan (MCAR ve MAR için Skillings-Mack testi değerleri sırasıyla 12,439, 12,301, 13,243 ve 10,831 ve p değerleri <0,001'dir), çoklu karşılaştırmalara bakılması gerekir. Daha sonra verilere dört yöntem uygulanmış ve ardından çoklu karşılaştırmalar için Nemenyi testi uygulanmıştır. Nemenyi testine dayalı üç çoklu karşılaştırmanın p değerleri ve ortalama rank farkları Tablo 4.5, 4.6, 4.7 ve 4.8'de verilmiştir. Kritik anlamlılık düzeyi 0,05'tir. Yokluk hipotezi, iki tedavi etkisinin aynı olduğu yönündedir. MICE-CART, MICE-RF ve rastgele sıcak deste atama sonuçları, tam veri sonuçlarına benzerdir. Özellikle silinme yüzdesi 10 olduğunda MICE-RF'nin p değerleri tam veri sonuçlarıyla benzer çıkmıştır. Liste bazında silme yöntemi sonuçları, tam veri sonuçlarına yakın değildir. Ayrıca, Skillings-Mack testi bu yaklaşım için uygundur çünkü Skillings-Mack testi istatistikleri, tam verinin Friedman testi istatistiklerine yakındır.

Tablo 4.5. Gerçek veri örneğinde MCAR altında verinin %10'u silinerek hesaplanan ortalama rank farkı ve p değerleri.

	(Tedavi öncesi-6 hafta sonrası)		(6 hafta-12 hafta sonrası)		(Tedavi öncesi-12 hafta sonrası)	
Yöntem	$\bar{R}_i$-$\bar{R}_j$	p	$\bar{R}_i$-$\bar{R}_j$	p	$\bar{R}_i$-$\bar{R}_j$	p
Tam veri	0,8500	0,1384	0,8500	0,1384	1,7000	0,0004
LD	0,8125	0,2350	0,8125	0,2350	1.6250	0,0033
HDI	0,9000	0.1091	0,7500	0,2140	1,6500	0,0006
MICE-CART	0,6500	0,3136	0,9500	0,0849	1,6000	0,0010
MICE-RF	0,8500	0,1384	0,8500	0,1384	1,7000	0,0004

Tablo 4.6. Gerçek veri örneğinde MAR altında verinin %10'u silinerek hesaplanan ortalama rank farkı ve p değerleri.

	(Tedavi öncesi-6 hafta sonrası)		(6 hafta-12 hafta sonrası)		(Tedavi öncesi-12 hafta sonrası)	
Yöntem	$\bar{R}_i$-$\bar{R}_j$	p	$\bar{R}_i$-$\bar{R}_j$	p	$\bar{R}_i$-$\bar{R}_j$	p
Tam veri	0,8500	0,1384	0,8500	0,1384	1,7000	0,0004
LD	0,7500	0,2909	0,9375	0,1459	1,6875	0,0021
HDI	0,8000	0,1732	0,9500	0,0849	1,7500	0,0002
MICE-CART	0,7000	0,2608	0,8500	0,1384	1,5500	0,0015
MICE-RF	0,8000	0,1732	0,9500	0,0849	1,7500	0,0002

Tablo 4.7. Gerçek veri örneğinde MCAR altında verinin %20'si silinerek hesaplanan ortalama rank farkı ve p değerleri.

	(Tedavi öncesi-6 hafta sonrası)		(6 hafta-12 hafta sonrası)		(Tedavi öncesi-12 hafta sonrası)	
Yöntem	$\bar{R}_i$-$\bar{R}_j$	p	$\bar{R}_i$-$\bar{R}_j$	p	$\bar{R}_i$-$\bar{R}_j$	p
Tam veri	0,8500	0,1384	0,8500	0,1384	1,7000	0,0004
LD	0,8333	0,3186	1,0833	0,1455	1,9167	0,0025
HDI	0,7000	0,2608	0,8500	0,1384	1,5500	0,0015
MICE-CART	0,7000	0,2608	0,8500	0,1384	1,5500	0,0015
MICE-RF	0,7000	0,2608	0,8500	0,1384	1,5500	0,0015

Tablo 4.8. Gerçek veri örneğinde MAR altında verinin %20'si silinerek hesaplanan ortalama rank farkı ve p değerleri.

	(Tedavi öncesi-6 hafta sonrası)		(6 hafta-12 hafta sonrası)		(Tedavi öncesi-12 hafta sonrası)	
Yöntem	$\bar{R}_i$-$\bar{R}_j$	p	$\bar{R}_i$-$\bar{R}_j$	p	$\bar{R}_i$-$\bar{R}_j$	p
Tam veri	0,8500	0,1384	0,8500	0,1384	1,7000	0,0004
LD	0,8333	0,3186	0,8333	0,3186	1,6667	0,0108
HDI	0,7500	0,2140	0,9000	0,1091	1,6500	0,0007
MICE-CART	0,8000	0,1732	0,8000	0,1732	1,6000	0,0010
MICE-RF	0,9500	0,0849	0,8000	0,1732	1,7500	0,0002

5. TARTIŞMA

Uzunlamasına veri, bireyler üzerinde belirli bir zaman periyodu boyunca tekrarlı ölçümlerin alındığı verilerdir. Araştırmacılar uzunlamasına klinik denemelere sıklıkla başvururlar çünkü bir tedavinin belli bir zaman aralığında bir hastalık üzerinde etkisinin olup olmadığını ölçmek isterler. Ayrıca, araştırmacılar tedavinin hangi zaman aralığında etkin olup olmadığı konusunda bilgi sahibi olmak isterler ve çoklu karşılaştırmalara başvururlar. Uzunlamasına verilerde eksik veri durumu önemli bir sorundur. Bu durumla ilgili literatürde çalışmalar vardır. Araştırmacılar en bilinen yöntem olan liste bazında silme yöntemi ile tekli ve çoklu atama yöntemini karşılaştırmışlardır. Örneğin, Zhu (39) uzunlamasına veride listesel silme yöntemi, ortalama atama ve çoklu atamayı karşılaştırmış ve çoklu atamanın en yansız sonuçları verdiğini belirtmiştir. Dragset (40) uzunlamasına veride eksik veriyi ele aldığı tez çalışmasında çoklu atama ve listesel silme yöntemini ele almış ve çoklu atamayı önermiştir. Uzunlamasına veride eksik veriyle ilgili çalışmalar olmasına rağmen, uzunlamasına veride parametrik olmayan çoklu karşılaştırmalarla ilgili atama ve silme yöntemlerinin karşılaştırıldığı bir çalışma yoktur. Literatürdeki bu boşluk yeni bir yaklaşım önerilerek doldurulmak istenmektedir. Önerilen yaklaşımın normal dağılım göstermeyen veya küçük örneklem büyüklüğüne sahip eksik gözlemi olan uzunlamasına verilerde parametrik olmayan çoklu karşılaştırmalara bir çözüm olabileceği düşünülmektedir. Hem benzetim çalışmasından hem de gerçek veri örneğinden elde edilen sonuç bu türdeki verilerde tam gözlemlerin kullanımı yönteminin doğru bir çözüm olmayacağıdır. Verilen örnekte %10 ve %20 eksik veri oranlarının çok büyük bir eksik yüzdesi olmamasına rağmen tam gözlemlerin kullanımı yönteminin p değerlerinin, tam verinin p değerlerine yakın olmaması durumuyla ve uzunlamasına klinik çalışmalarda çoğunlukla tercih edilen bu yöntemin bir çözüm olamayacağı sonucuyla karşılaşılmıştır. 1970'li yıllarda Birleşik Devletlerde çoklu atama yöntemi bulunmamış olsaydı araştırmacılar tam gözlemlerin kullanımı yöntemi ve sıcak deste ataması yöntemini kullanıyor olacak ve böylece çalışmalarda daha yanlı sonuçlar elde edilecekti. Çoklu atama yöntemi, bulunduktan sonra ilerleyen süreçte birleşik modelleme ve tamamen koşullu belirleme olarak iki ana kısma ayrılmış ve verilerin dağılımına göre kullanılmaya başlanmıştır. Çok değişkenli sürekli verilerin normal dağılıma uymadığı durumlarda tamamen koşullu

belirleme bir alternatif olmuş, bu başlık altında birçok yöntem önerilmiştir. Literatüre incelendiğinde, tek bir karar ağacı kullanan MICE-CART yöntemine karşı çok ağaçlı bir yapıyı kullanan MICE-RF yöntemi daha üstün ve yansız tahminler üretmiştir. Schwerter ve diğerleri (41) 2024 yılında yaptıkları çalışmada ampirik ve uzunlamasına çalışmalarda sonuç çıkarımında MICE-RF'yi ağaç tabanlı yöntemler arasında en iyi yöntem olarak belirlemiştir. Javadi ve diğerleri (27), kukla ve sürekli değişkenler arasında etkileşim olan iki durumlu sonuç değişkeni eksik gözlemlere sahip olan yapı için MICE-CART ve MICE-RF'yi karşılaştırmıştır. Yaptıkları benzetim çalışmasında MICE-RF, MICE-CART'a göre daha yansız sonuçlar vermiştir. Shah ve diğerleri (30), MICE-RF'nin sürekli değişkenlerde de kullanılabileceğini göstermiş ve yaptıkları benzetim çalışmasında bu yöntemi parametrik MICE yöntemleri ile karşılaştırmıştır. Benzetim çalışmasına göre MICE-RF, parametrik MICE yöntemlerine göre daha yansız sonuçlar vermiştir. Önceki yapılan benzetim çalışmalarında MICE-RF için ağaç sayısı 10 olarak belirlendiği için bu çalışmada da ağaç sayısı 10 olarak belirlenmiştir. MICE-CART ile ilgili yapılan benzetim çalışmalarında ise karmaşıklık parametresi 0,0001 olarak kullanıldığı için bu değer seçilmiştir. Benzetim çalışmasına göre, liste bazında silme yöntemi, MCAR altında daha yansız sonuçlar vermektedir. MICE-CART ve MICE-RF, MAR altında daha yansız sonuçlar vermektedir. MICE-CART ve MICE-RF, iyi kontrol edilen tip 1 hata ile orta ve küçük örneklem büyüklükleri için diğer yöntemlerden üstündür.

6. SONUÇ VE ÖNERİLER

Bu çalışma, üç parametrik olmayan atama yöntemini ve listesel silme yöntemini, normal dağılım göstermeyen eksik k-bağımlı örneklemlerin parametrik olmayan çoklu karşılaştırmalarına uygulamayı amaçlamaktadır. Araştırmacılar zaman aralıkları arasındaki farklılıklar hakkında bilgi edinmek istedikleri için çoklu karşılaştırmalar konusu, uzunlamasına klinik çalışmalarda çok önemlidir. Bu çalışmada, normal dağılım göstermeyen eksik gözlemlere sahip k-bağımlı örneklemlerde önce Skillings-Mack testi uygulandı. Gruplar arasında anlamlı bir fark bulunduktan sonra, üç parametrik olmayan atama ve liste bazında silme yöntemi uygulandı. Daha sonra tamamlanmış gözlemlerle parametrik olmayan çoklu karşılaştırmalar için Nemenyi testi kullanıldı. Çalışmada bu konu için bir örnek verildi ve bir benzetim çalışması yapıldı. Benzetim çalışması sonuçlarına göre, liste bazında silme yöntemi güç bakımından diğer yöntemlerden daha düşük düzeydeydi ve çok sayıda eksik gözlemin olduğu durumlarda kullanılmamalıdır. Bu çalışmanın sınırlılıkları vardır. Uzunlamasına veriler için çok değişkenli bir dağılım oluşturmak oldukça zor bir konudur. Gelecekte araştırmacılar farklı çok değişkenli dağılımlar kullanabilir ve 50 ve 100 gibi farklı örneklem büyüklükleri belirleyebilirler. Bu çalışmada, örneklem büyüklükleri 10 ve 20 olan dört yöntemin performansıyla ilgili anlamlı bir fark yoktur. 1000 tekrar sayısı, dört yöntem arasındaki güç farkını tespit etmek için makul bir tekrar sayısıdır çünkü tip 1 hatanın %95 güven aralığı 1000 tekrar için [0,0364; 0,0635]'tir ve bu güven aralığı makuldür. Bu çalışmada uzunlamasına sürekli veriler için bir yaklaşım önerilmektedir. Uzunlamasına kategorik veriler çalışmanın kapsamı dışındadır. Çalışmamızın normal dağılım göstermeyen eksik gözleme sahip uzunlamasına verilerin parametrik olmayan çoklu karşılaştırmalarını incelemek için alternatif bir yol göstereceğini düşünüyoruz. Bu konu özellikle uzunlamasına klinik çalışmalarda sorunludur. Araştırmacılar önerilen yaklaşımı uzunlamasına klinik çalışmalarında kullanabilirler. Önerilen yaklaşımın adımlarının tümüyle uygulanması önemlidir. Örneğin, Skillings-Mack testi, normal dağılıma sahip olmayan eksik gözlemi olan uzunlamasına veride parametrik olmayan çoklu karşılaştırmalar için gruplar arasında tümel fark olup olmadığını gösteren bir testtir. Bu testi uygulamadan çoklu karşılaştırmalara bakılmaması gerekir. Skillings-Mack testi, R yazılımında çeşitli paketlerde bulunmaktadır. Bu test kullanılırken ilgili komut çalıştırılıp eksik

gözlemin olduğu veri programda tanıtılır ve sonuçlar yorumlanır. Nemenyi testi parametrik olmayan çoklu karşılaştırmaların incelenebileceği en bilinen testtir. Ayrıca bu test R gibi uygulama yazılımlarında bulunmaktadır. Verilen örnekte görüldüğü gibi makine öğrenimi tabanlı yöntemlerden biri olan MICE-RF'nin %10 ve %20 silinme yüzdesinde tam verinin sonuçlarıyla benzer sonuçlar vermesi, bu yaklaşımın özellikle çok yüksek olmayan eksik yüzdelerinde rahatlıkla kullanılabileceği sonucuna ulaştırmaktadır. %10 ve %20 silinme yüzdesinde Skillings-Mack testi değerlerinin, tam verinin Friedman testi değerleriyle yakın sonuç vermesi, çok yüksek olmayan eksik yüzdelerinde Skillings-Mack testinin gruplar arasında tümel fark olup olmadığını ölçülmesinde doğru sonuçlar vereceğini göstermektedir. Tezdeki benzetim çalışması ve gerçek veri örnek setine ilişkin kodlar https://github.com/tuncyanar20008/thesis_codes adresinde verilmiştir.

7. KAYNAKLAR

1. Goulden, C. H. Methods of Statistical Analysis, 2nd ed., New York: Wiley, pp. 50-55, 1956.

2. Wilcoxon, Frank (1945). *"Individual comparisons by ranking methods"*. Biometrics Bulletin. 1 (6): 80–83.

3. Friedman, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association. 1937;32(200):675-701.

4. Rubin, D. B. Inference and missing data. Biometrika. 1976;63:581-592.

5. Choudhury, A. Missing Data Imputation Using Machine Learning and Natural Language Processing for Clinical Diagnostic Codes. Ph.D. thesis, North Carolina University, USA, 2020.

6. Van Buuren, S. 2012. Flexible Imputation of Missing Data. Chapman & Hall/CRC, Boca Raton, FL

7. Chatfield, M., and Mander, A. The Skillings–Mack test (Friedman test when there are missing data). Stata J. 2009; 9(2): 299–305.

8. Hollander, M., Wolfe, D.A., Chicken, E. Nonparametric Statistical Methods, 3rd Ed., New York: J. Wiley, 2014.

9. Little, R. J. A., Rubin, D. B., Statistical Analysis with Missing Data, 2nd Ed., New York: J. Wiley, 2002.

10. Wang, C., Stokes T., Steele R. J., Wedderkopp N., Andridge, R.R., Little, R.J. A. (2010), Implementing Multiple Imputation for Missing Data in Longitudinal Studies When Models are Not Feasible: An Example Using the Random Hot Deck Approach,Clinical Epidemiology, 14:1387-1403.

11. Skillings JH, Mack GA. On the use of a Friedman-type statistic in balanced and unbalanced block designs. Technometrics. 1981;23:171–177.

12. Nemenyi, P. Distribution-free Multiple Comparisons. Ph.D. thesis, Princeton University, USA, 1963.

13. Thorsten Pohlert, https://CRAN.R-project.org/package=PMCMRplus,2021, r package version 1.9.6.

14. McNeish D. (2017). Missing data methods for arbitrary missingness with small samples, Journal of Applied Statistics, 44:1, 24-39.

15. Ibrahim, J.G., Molenberghs, G. (2009), Missing data methods in longitudinal studies: a review, Test(Madr), 18:1, 1-43.

16. Andridge, R.R., and Little, R. J. A. (2010), A Review of Hot Deck Imputation for Survey Non-response.NIH Public Access, 78(1):40-64.

17. Rubin, D. B. (2004). The design of a general and exible system for handling nonresponse in sample surveys. The American Statistician, 58(4):298{302.

18. Scheuren, F. J. (2004). Introduction to history corner. The American Statistician, 58(4):290{291.

19. Van Buuren, S. 2007. Multiple imputation of discrete and continuous data by fully conditional specification. Stat. Methods Med. Res. 16, 219–242.

20. Doove, L. L., Van Buuren, S., Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. Computational Statistics & Data Analysis, 72, 92-104.

21. Huque H., Carlin J. B., Simpson J. A., Lee K. J. (2018), A comparison of multiple imputation methods for missing data in longitudinal studies, BMC Medical Research Methodology, 18(1):168.

22. Van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., and Rubin, D. B. (2006). Fully conditional speci_cation in multivariate imputation. Journal of Statistical Computation and Simulation, 76(12):1049{1064.

23. Rubin, D. B. Multiple imputation for nonresponse in surveys. New York: Wiley,1987.

24. Bartlett, M. S. (1978). An Introduction to Stochastic Processes. Press Syndicate of the University of Cambridge, Cambridge, UK, 3rd edition.

25. Gilks, W. R. (1996). Full conditional distributions. In Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., editors, Markov Chain Monte Carlo in Practice, chapter 5, pages 75{88. Chapman & Hall, London.

26. Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, second ed., Springer Verlag, New York.

27. Javadi, S., Bahrampour, A., Saber, M. M., Garrusi, B., Baneshi, M. R., Evaluation of Four Multiple Imputation Methods for Handling Missing Binary Outcome Data in the Presence of an Interaction between a Dummy and a Continuous Variable. Journal of Probability and Statistics, vol. 2021. 14 pages. 2021.

28. Terry Therneau, Beth Atkinson, Brian Ripley,2023, https://github.com/bethatkinson/rpart, r package version 4.1.23

29. Breiman, L., 2001. Random forest. Mach. Learn. 45, 5–32.

30. Shah, A.D., Bartlett, J.W., Carpenter, J., Nicholas, O., Hemingway, H. Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study. American Journal of Epidemiology.2014;179(6):764-774

31. Pedersen A. B., Mikkelsen E. M., Cronin-Fenton D., Kristensen N.R., Pham T.M., Pedersen L., Petersen I. Missing data and multiple imputation in clinical epidemiological research. 2017;9:157-166.

32. M. Tan, H.B. Fang, G.L. Tian, and G. Wei, Testing multivariate normality in incomplete data of small sample size, J. Multivariate Anal. 93 (2005), pp. 164–179.

33. R Development Core Team, 2024. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL:http://www.R-project.org/.

34. Brian Ripley, Bill Venables, Douglas M. Bates, Kurt Hornik, Albrecht Gebhardt, David Firth, 2021, https://CRAN.R-project.org/package=MASS , r package version 7.3-58.2.

35. Stef van Buuren, Karin Groothuis-Oudshoorn, Gerko Vink, Rianne Schouten, Alexander Robitzsch, Patrick Rockenschaub, Lisa Doove, Shahab Jolani, Margarita Moreno-Betancur, Ian White, Philipp Gaffert, Florian Meinfelder, Bernie Gray, Vincent Arel-Bundock, Mingyang Cai, Thom Volker, Edoardo Costantini, Caspar van

Lissa, Hanne Oberman, https://CRAN.R project.org/package=mice ,2021, r package version 3.15.0.

36. Tobias Rockel, 2021, https://github.com/torockel/missMethods, r package version 7.3-58.2 0.4.0.

37. Alpar, R. Uygulamalı Çok Değişkenli İstatistiksel Yöntemler. 5th Ed. Ankara: Detay, 2017.

38. Yetkin, Z. Kronik periodontitisli hastalarda diştaşı temizliği ve kök düzeltmesi ile %1'lik kitosan jel, diştaşı temizliği ve kök düzeltmesi ile %15 metronidazol içeren %1'lik kitosan jel, ve yalnız diştaşı temizliği ve kök düzeltmesi işlemlerinin klinik parametreler üzerindeki etkisinin karşılaştırılmalı olarak incelenmesi, Doktora tezi, Hacettepe Üniversitesi, Türkiye, 2001.

39. Zhu, X., (2014), Comparison of four methods for handing missing data in longitudinal data analysis through a simulation study.Open Journal of Statistics,4, 933-944.

40. Dragset, I. G. Analysis of Longitudinal Data with Missing Values. M. S. Thesis, Norwegian University of Science and Technology, Norway, 2009.

41. Schwerter, J., Gurtsikaia, K., Romero, A., Zeyer-Gliozzo, B., Pauly, M. Evaluating tree-based imputation methods as an alternative to MICE PMM for drawing inference in empirical studies. 2024. https://doi.org/10.48550/arXiv.2401.09602

8. EKLER

EK 1: Tez Çalışması Orijinallik Raporu

turnitin

Dijital Makbuz

Bu makbuz ödevinizin Turnitin'e ulaştığını bildirmektedir. Gönderiminize dair bilgiler şöyledir:

Gönderinizin ilk sayfası aşağıda gönderilmektedir.

Gönderen:	Tuncay Yanarateş
Ödev başlığı:	MAKİNE ÖĞRENİMİ TABANLI ÇOKLU ATAMA YÖNTEMLERİNİN ...
Gönderi Başlığı:	MAKİNE ÖĞRENİMİ TABANLI ÇOKLU ATAMA YÖNTEMLERİNİN ...
Dosya adı:	Tuncay_Yanarates_Tez.pdf
Dosya boyutu:	1.16M
Sayfa sayısı:	48
Kelime sayısı:	9,667
Karakter sayısı:	56,533
Gönderim Tarihi:	23-Mar-2025 01:34ÖS (UTC+0300)
Gönderim Numarası:	2622371204

T.C.
HACETTEPE ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

Windows'u
Windows'u etkir

MAKİNE ÖĞRENİMİ TABANLI ÇOKLU ATAMA YÖNTEMLERİNİN PARAMETRİK OLMAYAN ÇOKLU KARŞILAŞTIRMALARA UYGULANMASI VE BENZETİM ÇALIŞMASI

ORİJİNALLİK RAPORU

%4	%4	%2	%1
BENZERLİK ENDEKSİ	İNTERNET KAYNAKLARI	YAYINLAR	ÖĞRENCİ ÖDEVLERİ

BİRİNCİL KAYNAKLAR

1	www.openaccess.hacettepe.edu.tr:8080 İnternet Kaynağı	%2
2	acikerisim.omu.edu.tr İnternet Kaynağı	<%1
3	export.arxiv.org İnternet Kaynağı	<%1
4	Özüdoğru, Anil. "Diz Osteoartritli Hastalarda Kapalı Kinetik Zincir Egzersizleri ve Açık Kinetik Zincir Egzersizlerinin Karşılaştırılması", Dokuz Eylul Universitesi (Turkey), 2024	<%1

9. ÖZGEÇMİŞ